A PRE-REGISTERED AUDIT OF HIERARCHICAL RETRIEVAL: COST, UTILITY, AND ASSIGNMENT CONFIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating hierarchical retrieval requires measuring what its structure costs, whether it improves answers, and what its stored confidence can support. We audit these questions inside one RAPTOR-style pipeline. A capacity-constrained (truncated) variational DPMM (Dirichlet-process mixture model) serves as a confidence-bearing audit instrument rather than a proposed method: its fit stores a posterior responsibility per node–cluster pair, and the index keeps a scalar derived from it. In pre-registered comparisons on NovelQA (60 novels, 1,526 questions), the hypothesis that this index outperforms both the GMM+BIC and the Leiden index is not supported under either writer–reader backbone, and the registered comparison with flat dense retrieval does not establish a semantic-tree advantage. We report associations, not causal claims beyond this pipeline. The DP index stores 6,205,729 summary tokens against 1,602,734 for fixed-block contiguous grouping, without an established paired accuracy advantage (0.5770 against 0.5775). Rebuilds under 4 further construction seeds yield median within-book accuracy standard deviations of 0.0368 to 0.0480 over 54 books, at the scale of the registered design target. On an eighteen-book development sample, outside the registration, maximum a posteriori reassignment under sampled variational parameters stays close to the fitted partition (mean adjusted Rand index 0.97), whereas a refit on row-permuted embeddings lands farther (0.72). A fixed confidence reranker lowers book-mean accuracy from 0.5770 to 0.5589, a descriptive result under the registered testing sequence. These results motivate reporting utility beside construction footprint and rebuild variation, and testing stored confidence against both the variation it is meant to track and the decision it is meant to inform.

1 INTRODUCTION

Answering a question about a book-length document means deciding what a context-limited reader gets to see. A hierarchical index makes that decision before any question is asked: passages are grouped, each group is summarized, and the process repeats over levels (Sarthi et al., 2024). The grouping rule therefore fixes which summaries exist, how much the index costs to build and store, and which context the reader receives, so accuracy alone cannot justify a choice between grouping rules.

Within one implementation, rival grouping rules differ by small reported margins with no clear winner across datasets (Ou & Lapata, 2026), and in the indexes we cite the fit's own uncertainty about its grouping is neither stored nor audited (§2; a selected lineage, not a survey). Comparing semantic indexes with one another leaves a prior question unasked: does semantic grouping justify the construction it adds, measured against an inexpensive index built from the same document? Answering it takes flat retrieval over the same chunks, contiguous and document-structure grouping, each index's construction footprint, and rebuilds under other construction seeds, examined together inside one pipeline.

We began from a registered primary question. With a truncated variational Dirichlet-process mixture (Blei & Jordan, 2006) substituted at the grouping step of a RAPTOR-style pipeline, the primary evaluation covers 60 full-length novels with 1,526 questions, under a pre-registered conjunctive hy-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pothesis that the DP arm beats both the Leiden and the GMM+BIC index. We retain that hypothesis and its outcome: the registered comparisons do not establish the advantage, both point estimates positive and both intervals including zero, and the backbone robustness check reaches the same outcome (§4). This is not evidence of equivalence; it sets the scope of what follows.

Our contributions are empirical audits within a shared retrieval pipeline. First, we evaluate grouping choices jointly by question-answering accuracy and construction footprint, using pre-registered comparisons and flat and non-semantic structural references (§4). Second, we measure how rebuilding an index changes its partition, book-level accuracy, and chosen answers, keeping these outcomes distinct from corpus-average differences between methods (§5). Third, we evaluate stored assignment confidence both for its stability under within-fit reassignment versus refitting and for its use by a fixed retrieval reranker (§6). The posterior diagnostic is exploratory, and the reranking result is descriptive under the registered testing sequence. Together, these measurements specify what evidence is needed to interpret a hierarchy's accuracy and its confidence within the evaluated setting.

Each of them asks for its own measurement: the utility that corresponds to a construction cost, the sensitivity of that utility to rebuilding, and the variation a stored confidence is meant to track and the decision it is meant to inform.

2 Related Work

Hierarchical indexes and their clustering step. RAPTOR (Sarthi et al., 2024) introduced the recursive cluster-and-summarize index; grouping rules have since been compared within a shared implementation at book scale (Ou & Lapata, 2026) and replaced outright by Leiden community detection on a k-NN graph (Traag et al., 2019; Liu et al., 2026b). Other successors redesign the index-construction pipeline itself; we hold the pipeline's interface fixed and vary only the statistical form of the assignment, so they enter as evidence about clustering choices, not as arms (SiReRAG (Zhang et al., 2025) among them; Appendix A, Table 5; what a swap changes besides that form: §3.2). The per-assignment posterior written beside the index at build time is our addition, a claim about what the build writes rather than about accuracy. Our contribution is the joint audit of construction footprint, structural controls, rebuild variation, and the consumption of stored confidence within one evaluation, with registered comparisons kept distinct from exploratory diagnostics. Reader-side and budget-allocation accounts ask what a construction advantage is worth under a stronger reader or a better budget split; with GPT-4o readers, document-order packing without a hierarchy has been reported to outperform RAPTOR on several long-document benchmarks (Laitenberger et al., 2025). We hold both fixed, carry document-order repacking as a registered descriptive control (Figure 2), and re-read it under a larger probe reader (Appendices W and A). Bayesian nonparametric methods for hierarchy predate retrieval augmentation (Blei et al., 2003; Teh et al., 2006; Heller & Ghahramani, 2005; Teh, 2006); Appendix A contrasts them with our use of a partition. Known pathologies of variational approximations (underdispersion (Blei et al., 2017), second moments mis-estimated where means are accurate (Giordano et al., 2018)) motivate the diagnostic of §6; we contribute no new inference method.

3 Audit Design and Evaluation

3.1 Audit structure

The audit compares grouping choices by question-answering accuracy and by construction footprint, beside flat and non-semantic grouping controls; it evaluates rebuilding under further construction seeds apart from the comparison between methods; and it examines stored confidence as reassignment within a fit, as agreement across refits, and through a consumer that ranks with it (Figure 1). This is the design of the present audit, not a procedure validated in other settings.

3.2 The shared pipeline and its clustering seam: a mechanistic account

RAPTOR (Sarthi et al., 2024) builds a leveled index (Figure 1) by chunking the text, embedding the current level's nodes, reducing with uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) and grouping with a Gaussian mixture whose component count the Bayesian

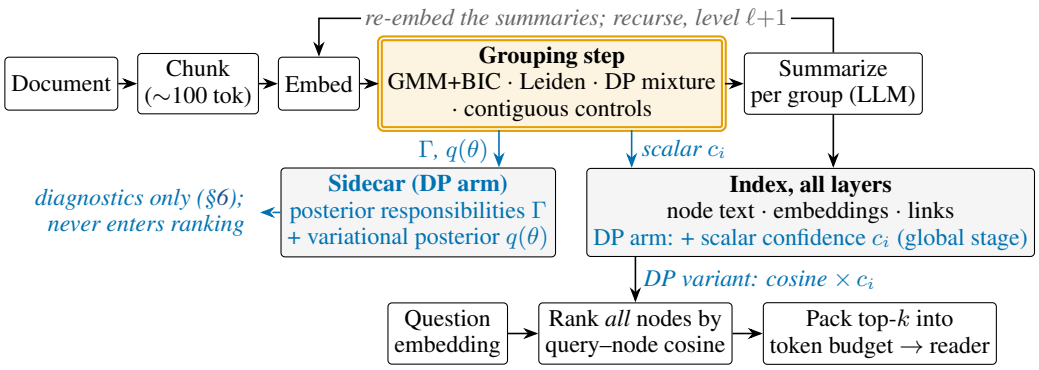

Figure 1: Offline construction (top) groups and summarizes nodes recursively; the grouping step is the only operator replaced. At query time (bottom) every layer is searched together by cosine and, of the top-k nodes, those of positive cosine are taken in rank order up to the context budget, so k and the budget are two limits and either can bind (Table 16). The primary DP arm ranks by cosine alone, the reranked arm by cosine $\times\, c_i$; $\Gamma = [r_{ik}]$ is the responsibility matrix written to the build-time sidecar, the index itself carrying the derived scalar c_i (§3.3, Appendix K).

information criterion (BIC) (Schwarz, 1978) selects (globally and then locally within each global cluster, letting a node join more than one group when its membership probability clears a fixed 0.1, the constant of RAPTOR's released implementation), then summarizing each group with an LLM and recursing. At query time collapsed-tree retrieval ranks every node of every layer by cosine similarity into a fixed token budget (Figure 1). The grouping step is the only place the method decides which nodes belong together, hence which summaries exist: the clustering seam. We share chunking, embedding, summarization, reading, and the default collapsed-tree retrieval rule across arms. The seam contrasts also carry their representation and fitting choices: the mixture arms reduce with UMAP, whereas the default Leiden arm groups raw embeddings, and two representation-matched Leiden controls bound that gap (Table 30). Among the first registration's arms that retrieve, the lexical floor (own scorer), the posterior consumer (own ranking), and the repacking control (own packing) each change one retrieval operator; the others differ only in the index retrieved from. **No-tree floors**: closed-book (no context), flat dense retrieval over leaf chunks, and flat BM25 (Robertson & Zaragoza, 2009) over the same chunks (lexical floor). **Tree arms at the seam**: the original GMM+BIC clustering (Sarthi et al., 2024); Leiden graph clustering on a k-NN graph (Traag et al., 2019; Liu et al., 2026b) on raw embeddings, plus the two representation-matched Leiden controls; and the variational DP mixture below. **Posterior consumer**: confidence-weighted reranking (§3.3) over the DP-mixture tree. **Structural controls**: contiguous grouping (no embedding or clustering input), document-order repacking, chapter-boundary grouping, and cluster-count-matched variants giving a competitor arm the DP arm's per-layer cluster counts. **Ceiling**: an oracle-evidence reader over gold evidence spans. Map by contrast: Figure 5; arm table and comparator disclosure: Appendix C.

At the seam we substitute the BIC-selected Gaussian mixture with a truncated variational Dirichlet-process mixture (Blei & Jordan, 2006), an audit instrument rather than a new inference method. The swap removes the BIC scan at every clustering step (one fit replaces each sweep) but at our settings does not make K a quantity the data selects: the truncation is an upper bound on the reported K that the fitted count frequently exhausts (rule and counts: Appendix K). What it does change is the object left behind: the baseline's membership probabilities are E-step quantities conditional on a point estimate of the parameters and on a count fixed by the separate BIC scan (ours by the truncation just stated), so the baseline's fit leaves no distribution over its own parameters; the variational fit returns, in the same family on the same coordinates, a posterior $q(\theta)$ over parameters and weights beside the responsibilities, and that posterior is what §6 audits. Stick-breaking construction, inference settings and the model class's cautions: Appendix K.

3.3 What assignment confidence is and how retrieval uses it

The variational fit yields, for every node i and every k in the fit's active set act, a responsibility $r_{ik} = q(z_i{=}k)$ under the fitted approximation (the mean-field factor for node i, which is conditional on the whole fit, not on $\mathbf{x}_i$ alone), and we use it twice. For **membership**, a node joins cluster k when

$r_{ik} \geq t_{\text{eff}} = \min(0.1,\, 0.5/K_{\text{eff}})$, where K_{eff} counts this fit's components holding some node's argmax responsibility; this reproduces RAPTOR's 0.1 cut at $K_{\text{eff}} \leq 5$ and relaxes proportionally beyond it. For **assignment confidence**, each clustered node stores

$$c_i \;=\; 1 - \frac{H(\tilde{\mathbf{r}}_i)}{\log K_{\text{eff}}}, \qquad H(\tilde{\mathbf{r}}_i) = -\sum_k \tilde{r}_{ik} \log \tilde{r}_{ik}, \qquad \tilde{r}_{ik} = \frac{r_{ik}}{\sum_{k' \in \text{act}} r_{ik'}}, \tag{1}$$

one minus the normalized entropy of the node's active-component responsibilities, a certainty in $[0, 1]$ written into the index at build time (c_i=1 when a single component is active). We name c_i a normalized allocation-certainty heuristic. It belongs to the global clustering stage and does not depend on the query. Component overlap, multi-topic membership, allocation ambiguity, and representation geometry mix into this entropy inseparably, so we interpret it only as allocation certainty and do not validate it as an epistemic-uncertainty, retrieval-correctness, or answer-confidence signal; the standards an answer-confidence signal would have to meet are those of Soudani et al. (2025). Its one registered consumer is a fixed multiplicative reranker over the same tree; as deployed, it ranks by cosine $\times\, c_i$, takes the top-k slice on that product, and then drops candidates whose raw cosine is non-positive (the registered order differs: §6). It changes the ranking, not the tree, and reads the index's own query-independent build-time confidence rather than the reader's answer-time confidence (Song et al., 2026b;a) (tie-breaking, orphan-rate logging, and the two-stage simplification: Appendix K).

3.4 Evaluation and inferential status

Our confirmatory benchmark is NovelQA (Wang et al., 2025), multiple-choice question answering over full-length English novels. The public snapshot we build on (Hugging Face revision `60f8de6a`, downloaded 2026-07-06) has 60 public-domain books and 1,526 usable questions by our harness's count (Appendix B); books range from 86K to 2.58M tokens, so no context window here sees a whole book. Collapsed-tree retrieval takes the top-20 nodes into a 2,000-token context for the main run; a descriptive budget curve revisits the cached trees at three further budgets (Appendix G). Top-k caps the candidate pool before token-budget packing, so a shared token cap is not shared context use: at the main budget every flat top-k candidate fits for every question, against 27.3% of questions for the DP arm (Table 16). `Llama-3-8B-Instruct` is both summarizer and reader, with greedy decoding, fixed across arms (completion parser and the registered parser-shape monitor: Appendix Z), and `Qwen2.5-7B-Instruct` is the second writer–reader backbone of the registered backbone robustness check, summarizing and reading on its own rebuilt trees; confirmatory results are conditional on the primary pair. Every other component setting is frozen across arms (Appendix C). The registered primary unit is the unweighted book mean, with paired comparisons at the book level. A transfer check on further datasets is reported apart from the primary analysis (Appendices M and N).

QuALITY (Pang et al., 2022) enters here as a development diagnostic and as the halted confirmatory attempt of Appendix J.1, which is why NovelQA is an adaptively selected benchmark. The primary hypothesis, arm set, run order, analysis code, spend cap (USD 850), and abort criteria were frozen, reopened once, and re-frozen into the hash-pinned specification. Only then was the answer key, hash-pinned at receipt, unsealed at the registered run-order gate; no freeze, re-freeze, build, or development decision taken before that gate saw it, an ordering that rests on our own version control and run ledgers, neither hash carrying a third-party timestamp (Appendix J). The primary predicate (H1) is conjunctive: the DP arm must beat both the Leiden arm and the GMM+BIC arm, evaluated by a cluster-robust paired comparison at the book level (60 books, 59 degrees of freedom), each interval lying above zero; the design target that sized the study is not that threshold. Because H1 was not established, the fixed-sequence reranking comparison is descriptive. The rest of the registration (interval semantics and error rates, power at the design target $\delta^* = 3.4$ pp, testing order, gates and read-outs) is in Appendix J. The claim register is Appendix Q. These analyses differ in sample and in inferential status:

Analysis	Sample	Status
Primary evaluation (§4)	NovelQA, all 60 books and their questions	first registration, a single construction seed
Rebuild study (§5)	accuracy: 54-book census; agreement: 12-book draw	second registration, fixed after the first evaluation; deviations disclosed
Posterior diagnostic (§6)	18 development books	outcome-informed exploratory follow-up

Table 1: Registered evaluation arms and structural controls on NovelQA (60 books, 1,526 questions; primary backbone). In the last two rows, two registered structural controls that take no embedding or clustering input; question-weighted means are omitted here (–; Appendix W pools questions for both in its post-hoc reader band). Registered primary unit: unweighted book mean; question-weighted mean co-reported. A zero is structural, not unmeasured. Chapter boundaries were used in 36 of the 60 books; the other 24 (40.0%) use fixed-block contiguous grouping.

Arm	Acc (book mean)	Acc (question mean)	Summary tokens
Closed book	0.3845	0.3893	0
Flat dense	0.5710	0.5616	0
GMM+BIC tree	0.5606	0.5531	4,822,747
Leiden tree	0.5674	0.5623	1,238,489
Leiden tree, UMAP-matched	0.5703	0.5564	1,402,350
DP-mixture tree	0.5770	0.5675	6,205,729
+ confidence reranking	0.5589	0.5511	(DP index reused)
Fixed-block contiguous (non-semantic)	0.5775	–	1,602,734
Chapter-boundary grouping	0.5837	–	1,849,571

Table 2: Registered comparisons on NovelQA (60 books). (A) The two conjuncts of H1 under each writer–reader backbone (primary `Llama-3-8B-Instruct`, robustness `Qwen2.5-7B-Instruct`): paired book-mean differences with cluster-robust two-sided 95% intervals; H1 requires both intervals of a backbone to lie above zero, and the backbones are never pooled. (B) Tree minus flat dense retrieval on the primary backbone: one-sided 95% bounds, not two-sided intervals; the registered floor criterion needs the lower bound above zero, and the upper bound is a registered co-report (none registered for the UMAP-matched arm, –).

(A) H1 conjuncts, two-sided intervals

Backbone	DP − Leiden	DP − GMM+BIC
Primary	$+0.0096$ $[-0.0114, +0.0306]$	$+0.0164$ $[-0.0050, +0.0378]$
Robustness	$+0.0045$ $[-0.0168, +0.0259]$	$+0.0036$ $[-0.0146, +0.0217]$

(B) Tree minus flat dense, primary backbone, one-sided bounds

Tree	Difference	Lower	Upper
GMM+BIC	-0.0104	-0.0270	$+0.0062$
Leiden	-0.0036	-0.0154	$+0.0083$
DP mixture	$+0.0060$	-0.0147	$+0.0266$
Leiden, UMAP-matched	-0.0007	-0.0175	–

4 Cost and Utility at the Registered Seed

Validity floors. All 60 books enter every arm, and the registered validity gate passes: flat dense retrieval beats closed book by $+0.1865$ (one-sided lower bound $+0.1436$), consistent with useful retrieval but not isolating its contribution. The closed-book reader reaches 0.3845, above the trivial floors (Appendix L) and below every retrieval arm (Table 1), with a wide per-book range (0.1579 to 0.6346, median 0.3684) that does not by itself identify memorization.

The registered hypothesis under both backbones. At the registered construction seed the DP index reaches book-mean accuracy 0.5770, against Leiden's 0.5674 and GMM+BIC's 0.5606 (Table 1). Neither conjunct of the registered predicate (§3.4) is established: each interval includes zero and neither point reaches the registered effect. The DP−Leiden difference is $+0.0096$ (95% CI $[-0.0114, +0.0306]$) and the DP−GMM difference is $+0.0164$ ($[-0.0050, +0.0378]$); both intervals are reported unconditionally. The swap flips correctness on 15.7% of the questions against Leiden and on 13.4% against the mixture (Appendix V): the small mean differences sit on substantial question-level churn. The backbone robustness check, scored at its own α and never pooled, reaches the same registered outcome, so H1 is not supported in either backbone (Table 2, panel A). The comparison was sized for an absolute difference of 0.034 (3.4 pp), and the two comparators end in different places against it: the Leiden interval excludes an effect that size, the GMM+BIC interval does not. The registered verdict over the conjunction is powered abandonment, so the DP-over-Leiden direction is discontinued: no evidence of an effect of 3.4 pp or more, and the null is not thereby proven; we lean on the intervals, not on that label: they establish no advantage, and with no equivalence margin registered they license no registered equivalence claim (the post-hoc equivalence reading and the dispersion threshold behind the label: Appendix S). The registered outlier-sensitivity and partial-pooling co-reports read the same way (Appendix L).

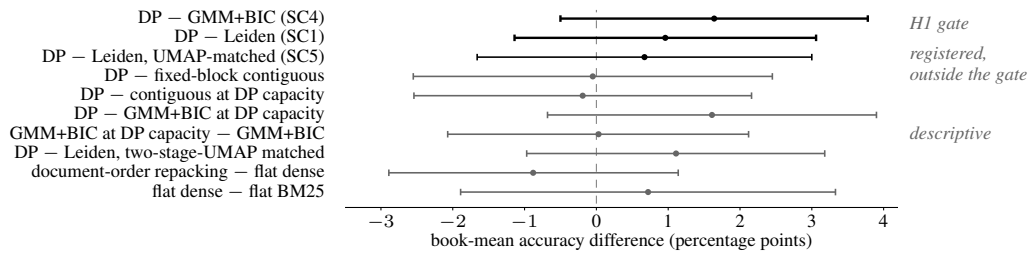

Figure 2: SCn throughout is the registration's own label for a contrast, not a rank: SC1 and SC4 are the two conjuncts of the primary predicate, SC5 the UMAP-matched Leiden contrast, SC3 the reranker comparison of §6, SC7 the UMAP-matched floor. The registered comparisons of Table 21 (Appendix L) on one axis, primary backbone: paired book-mean differences over the 60 books, cluster-robust 95% intervals, dashed line at no difference. The two H1 rows are evaluated jointly as the registered conjunctive predicate. SC0 (one-sided) has no finite upper bound to draw.

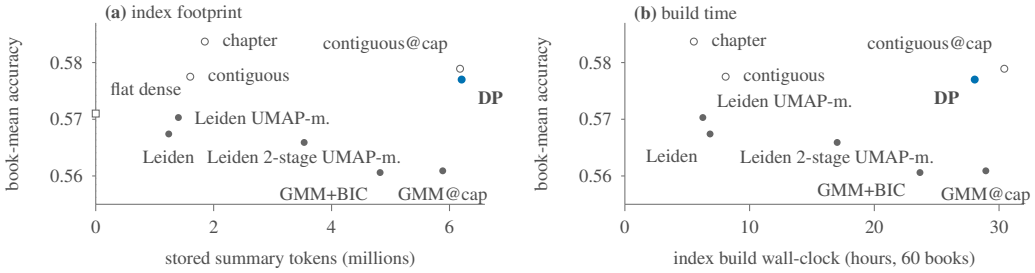

Figure 3: Construction cost beside book-mean accuracy for every tree-building arm (primary backbone, 60 books; filled, the semantic family with DP in colour; open, the non-semantic controls and the flat reference): (a) stored summary tokens; (b) recorded index build wall-clock (flat arm not timed). The accuracy axis does not start at zero and no per-arm interval is registered: separation is descriptive, chapter against DP is not a registered contrast, the chapter and contiguous arms build identical trees on 24 of them (40.0%), and ordering is read from Figure 2 (Appendix U).

Semantic hierarchy against flat retrieval. The registered per-arm floor criteria all fail (Table 2, panel B), so the registered headline cell applies: no tree variant, including ours, is shown to beat flat dense retrieval at the registered per-arm resolution among the arms carrying a registered floor comparison (arms outside it: Tables 21 and 13). The registration attaches two obligatory co-reports to this comparison. For the UMAP-matched Leiden arm the point estimate is -0.0007 and the one-sided lower bound is -0.0175, so the statement covers that variant. On one-sided 95% upper limits, the GMM+BIC and Leiden advantages over flat are below $+0.02$ ($+0.0062$ and $+0.0083$) while the DP arm's is $+0.0266$, so a small DP advantage over flat cannot be excluded and the frozen gate licenses no mechanism-level displacement claim in either direction. Both duties follow the backbone robustness check into its own cell, where the floor criteria fail as well (Appendix S). The second registration's stronger flat retrievers are dense–lexical rank fusion and a cross-encoder rerank at a base and, unregistered, at a large capacity. In point estimate they sit above the dense floor, the Leiden index and the DP index (3, 3 and 3 of 3). Of 9 paired intervals 1 excludes zero (rank fusion against the Leiden index). These comparisons do not replace the first registration's floor test (Appendix L.1). A registered length-stratified read-out (tercile cells plus a length–delta correlation, boundaries frozen before unsealing) establishes its registered direction in 0 of 12 cells (Appendix F).

Footprint and the structural controls. The most informative control is fixed-block contiguous grouping, which takes no embedding or clustering input: it reaches 0.5775, and the DP arm's paired difference against it spans zero (-0.0005, $[-0.0255, +0.0245]$), as does the difference for a contiguous tree at the DP arm's per-layer cluster counts (0.5789). The DP index stores 6,205,729 summary tokens against the contiguous index's 1,602,734, so this comparison documents a substantially larger summary footprint without establishing a corresponding accuracy advantage; stored summary tokens are not the whole cost. Against Leiden and GMM+BIC the DP index stores $5.0\times$ and $1.29\times$ their summary tokens, at a recorded build wall-clock of 28.05 h against 6.83 h and 23.66 h, an axis ordered rather than controlled: arm order is fixed and the cache state each build

Table 3: Rebuilding under 4 further construction seeds (second registration, primary backbone; different samples per row). Agreement: per book, the median ARI over seed pairs between the leaf layer and the first summary layer, averaged over the locked 12-book draw. Accuracy sd: each book's sample standard deviation across the 5 seeds, median over the 54-book census. Flips: changes of the chosen option over comparable seed pairs (both runs chose an option) of the census, which need not change correctness; the registration defines a flip as a change of correctness, and this reported definition departs from it. Full table: Table 8.

	GMM+BIC	Leiden	Leiden, UMAP-matched	DP mixture
Agreement (ARI, draw)	0.576	0.706	0.756	0.685
Accuracy sd (census)	0.0447	0.0372	0.0368	0.0480
Flips, % (census)	20.0	13.7	14.3	19.8

meets is part of the cell it was measured in (Figure 3, Appendix U). Giving GMM+BIC the DP arm's per-layer cluster counts moves it slightly. The count-matched arm also swaps the two-stage fit for a flat fit per layer, so cluster count is neither identified nor ruled out. The contrasts involving the two representation-matched Leiden controls remain inconclusive: DP minus UMAP-matched Leiden is $+0.0067$ ($[-0.0166, +0.0300]$), and UMAP-matched minus raw Leiden is $+0.0028$ ($[-0.0153, +0.0210]$); the point estimates are small, but the intervals admit changes large enough to reorder the arms. No control here isolates an accuracy benefit attributable to semantic grouping (registered controls with their intervals: Figure 2 and, in Appendix L, Table 21).

Descriptive read-outs and directional cells. Descriptive read-outs trace where the arms differ, without a mechanism claim. Gold-evidence recall is verbatim by construction, so it bounds evidence delivery rather than clustering quality: summary share rises across Leiden, GMM+BIC, and DP while conditional recall of gold spans falls, and pruning the DP arm's summary layer lowered that recall on no book (Appendices L and D). The chapter-boundary control (author markup where present; Figure 3) posts the highest point estimate among the first registration's primary-backbone non-oracle arms (0.5837 over all 60 books). Its value on the 36 chapter-parsed books appears with its matched comparisons in Table 13. These chapter-boundary results are descriptive and do not establish an accuracy advantage. The post-hoc decomposition of Appendix P likewise supports comparison between arms, not absolute shares (its subset is measurably easier than the remainder; Appendix X). Of the 21 registered directional cells, 0 met the criterion of the displacement account that motivated the design (Appendices F and I); no directional conclusion follows.

5 Reconstruction Sensitivity

A second registration, fixed after the first evaluation, rebuilds the arms of Table 3 under 4 further construction seeds (Appendix C.1). Partition agreement is measured on a locked 12-book draw and accuracy on a 54-book census, the confirmatory corpus with the 6 development-track books held out; the samples differ and are not pooled. Rebuilding changes both the partition and the answers. No rule reproduces its own partition: mean leaf-to-first-layer agreement on leaves singly assigned under every arm and seed runs 0.576 to 0.756 across arms and no book reaches full agreement (intersection-union $p = 0.00024$ against unchanged partitions). The per-book accuracy spread across seeds (sample standard deviation, $n - 1$ in the denominator), median 0.0368 to 0.0480, is at the scale of the design target $\delta^* = 0.034$, and the chosen answer changes on 13.7 to 20.0% of comparable seed pairs (Table 3). These are within-book reconstruction measurements, not uncertainty intervals for corpus-level differences between methods, which are a different statistic (variance components and an exploratory multi-seed mean: Appendix C.1); the one-seed H1 verdict is unchanged.

The agreement analysis reports a mask chosen with the draw partly complete and some values visible, a recorded deviation; the draw's population was narrowed after an initial draw had been seen; and of the three code revisions behind the rebuilds, two carry an unauditable uncommitted delta (Appendix C.1). We therefore report reconstruction sensitivity without attributing all of the observed variation to the construction seed alone.

Table 4: The variation sources of §6; row P is its row-permuted refit. Every ARI is computed against the reference fit's MAP labeling; q is that fit's variational posterior; B and B′ use two parameter draws per book on the reference fit's active components; means are over 18 of the 19 development books (one excluded by a replay gate, Appendix R; per-book range in parentheses).

Axis	What varies	Mean ARI (book range)
A (refit)	refit after a 10^{-7} perturbation of the input embeddings	0.6672 (0.5837–0.7485)
B′ (within fit)	parameters drawn from q, then MAP relabeling	0.9668 (0.9550–0.9772)
B (within fit)	parameters drawn from q, then categorical assignment sampling	0.9256 (0.9016–0.9495)
P (row-permuted refit)	refit on row-permuted input with the reduced embeddings held fixed	0.7171 (0.6397–0.7910)

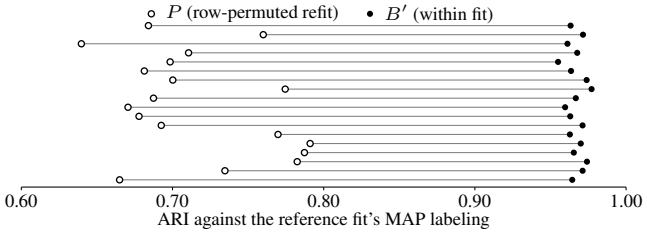

Figure 4: Within-fit reassignment (row B′ of Table 4; filled: MAP relabeling under parameters drawn from the fit's variational posterior q) and row-permuted refit (row P; open: reduced embeddings held fixed), book by book, on the 18 development books. P minus B': mean -0.2497, book-bootstrap 95% CI $[-0.2700, -0.2288]$; P above B' on 0 of 18 books (Holm over the four-contrast family, $p = 3.1 \times 10^{-5}$). Books appear in the driver's order, not sorted by either axis. Post hoc, exploratory, development sample only (Appendix O).

6 Assignment Confidence: Stability and Retrieval Use

This diagnostic is development-stage: it was specified after the observation that prompted it and before the measurements reported here, separately from the primary registration. On the global clustering stage's variational fit (§3.3), whose marginals r_{ik} enter Equation 1, it compares two sources of assignment variation, both scored against the reference fit's maximum a posteriori (MAP) labeling by ARI: *within-fit reassignment*, parameters drawn from the fitted variational posterior and then MAP relabeling (row B′ of Table 4), and the *row-permuted refit*, the same reduced embeddings refitted after their row order is permuted (row P). The sample is 18 development books drawn from InfiniteBench En.QA (Zhang et al., 2024) (a further book excluded by a replay gate), five of which are also NovelQA titles.

Within-fit reassignment barely moves the partition (mean ARI 0.9668), whereas row-permuted refits land farther (0.7171); the paired difference is -0.2497 (book-bootstrap 95% CI $[-0.2700, -0.2288]$; Figure 4). The separation survives 32 draws per book and dropping the 5 books shared with NovelQA. Certainty within a fit therefore does not imply agreement across refits; the diagnostic does not establish miscalibration of the stored responsibilities, nor does it estimate the width of the full partition posterior (further axes and checks on the source of the gap: Appendix O). A third axis, a refit after a 10^{-7} perturbation of the input embeddings (row *A*), is tabulated but is not the yardstick we argue from (§6; the sensitivities, the A-versus-B′ gap and row *B*: Appendix O).

Checks on the source of the stability gap. *Is the perturbation axis itself an artifact?* A zero-information control (row-order permutation: the point cloud is unchanged, the reduction refitted) is not separable from the 10^{-7} perturbation by a sign test, while a bootstrap interval on the mean difference excludes zero. That is why §6 argues from row-permuted refits, which need no floating-point-scale perturbation. Both readings, a plug-in log-likelihood comparison in which permuted refits show no clear mean difference from the untreated fit, the reduction's share of that sensitivity and a tempering diagnostic are in Appendix O. *Approximation or model?* On the development sample, under a diagnostic model with exactly computable mass, the fit's own labeling is the highest-mass candidate in none of the 18 books, on point masses that do not separate approximation from model (Appendix O).

In the separate NovelQA evaluation, the registered consumer of the stored responsibilities (§3.3) has lower accuracy than the index it reranks: 0.5589 against 0.5770, a reranked-minus-DP difference

of -0.0181 ($[-0.0338, -0.0023]$). Under the registered fixed-sequence rule this comparison is descriptive (the primary gate did not resolve, so no alpha reaches it), and the interval lies entirely below zero. What that adjudicates is the consumption rule, not the quantity: the registered consumer multiplies by a query-independent scalar whose median over the 195,837 layer-0 nodes is 0.998 (Appendix V), and post hoc it changes the retrieved node set on 79.0% of queries and the final answer on 11.7%. The registration excludes negative-cosine candidates before reranking, whereas the deployed code drops non-positive-cosine candidates after the top-k slice; the two orders return different sets only through a tie at score zero, and how often that arose in the main run was not measured (Appendix J). The two studies concern different settings, and neither shows that refit instability caused the reranker's lower accuracy.

7 Discussion, Limitations, and Conclusion

What to measure. The audit suggests three practices (Appendix O.1). First, report a grouping choice's utility beside its construction footprint, stored summary tokens and build time, and compare it with flat and contiguous controls built from the same chunks, not only with other semantic indexes: here a contiguous control was not separated from the confidence-bearing index while storing far fewer summary tokens. Second, rebuild the index under several construction seeds and report the per-book spread and the answer flips apart from the corpus-mean intervals, since a small mean difference can sit on large book-level movement. Third, name the variation a stored confidence is meant to track, test it against that variation, here a refit rather than a reassignment within a fit, and test the decision that consumes it directly, since stability and utility were separate findings in this audit.

Limitations. *Scope.* One pipeline, a primary benchmark of fiction, and small open-weights models carry the results. A single 8B reader carries the confirmatory run; a post-hoc, exploratory re-read by a larger same-generation reader compares the DP arm with Leiden, flat dense retrieval and the closed-book floor, only the floor distinguished at this n (Appendix W). Residual exposure is diagnosed by the closed-book floor, not excluded by dates, and in the registered comparisons one model both writes the summaries and answers. The transfer check is reported cell by cell in Appendices M and N, unpooled and without a label; its seam contrasts split in sign, and across both registrations the intervals excluding zero are 1 against the DP arm and 1 for it, in different slices. Most of the DetectiveQA questions in that check are AI-supervised (Appendix B), a correlated-error risk the analysis does not stratify away. *Registration and rebuilds.* NovelQA was adaptively selected, and its chronology rests on our own ledgers (§3.4). Every confirmatory tree was built under one seed (224); construction is not fully deterministic given the seed (exceptions are disclosed), and the rebuilds carry the deviations and the unauditable code delta of §5. The covariance-shape axis, measured on the development track only, is not propagated (Appendix S). *Confidence.* The stored confidence reflects the global clustering stage only (its sign interaction with similarity: Appendix S). The stability diagnostic is exploratory and scoped to the development sample with its source overlap, and licenses only the reassignment-versus-refit comparison of §6, not a claim that the posterior fails to capture real-world uncertainty. *Retrieval stack and release.* A shared token cap is not shared context use (§3.4). The index contrast bounds its lexical axis by a BM25 arm, not by sparse–dense fusion, for the reason Appendix S gives. Most reported statistics can be re-aggregated from released per-record correctness, with the exceptions Appendix J.2 names, but rescoring an answer independently needs its chosen letter, which the released records withhold.

Conclusion. In this RAPTOR-style pipeline, the registered comparisons did not establish the DP index's advantage over both semantic comparators under either backbone, and none of this is evidence for the null. The audit additionally measures the summary footprint of grouping choices, the variation across index rebuilds, and the distinction between confidence within a fit and its use in retrieval. These findings support evaluating a hierarchy through its construction resources and reconstruction behavior alongside accuracy, and testing a stored confidence signal against the variation and the decision for which it is used.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the conceptual framework and theoretical models, to propose and refine hypotheses, to design and critique the methodology and experiments, to formulate mathematical claims and assist with their proofs, to implement methods, to clean and reformat dataset loaders and manifests, to support qualitative and thematic analysis, to interpret results, and to assist with translation. We did not use generative AI tools to generate synthetic data sets. Additionally, we used generative AI tools to create and edit code (including the scripts that generate figures and tables), to draft and edit parts of the paper, to summarize and identify related literature, and to format references. We also used generative AI tools as external manuscript reviewers; their critiques were checked against the ledgers, the code, and the cited sources before adoption. We have reviewed all AI-assisted work. Accuracies and their intervals are recomputed from the append-only per-query ledgers, apart from the copied and quoted figures Appendix J.2 names, and the paper's build gate fails on any result number written in digits in its text that neither enters through a generated macro file nor is registered with its source in a checked allowlist, and checks number words only in part. For every citation we kept an internal record of the passages consulted and of where only the abstract was read; these records are not part of the supplementary material. Novelty and significance judgments, the scientific claims, the pre-registered design, the choice of when to freeze the registration and when to unseal the answer key, the decision to halt the earlier QuALITY attempt at its registered sanity gate and continue on NovelQA, and all final wording were decided and approved by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI. No LLM is an author.

REPRODUCIBILITY STATEMENT

The first registration's full protocol (hypotheses, arm set, run order, analysis code, spend cap, and abort rules) was frozen and hash-pinned before the gold answer key was unsealed, and the analysis gate consumes the frozen commit; the second registration (Appendix C.1) was frozen after the confirmatory run. Tree builds are not fully deterministic given the fixed seed (224; measured exceptions are disclosed in §7 and listed in Appendix S); answer-time model usage is recorded in an append-only per-query ledger with tokens, cost, latency, and cache state, build-time summarization in per-book build records, and the accuracies and their intervals are recomputed offline from the per-query ledgers (apart from the copied and quoted figures Appendix J.2 names), the index budgets from the build sidecars, and the costs from the spend ledger. Code is released in anonymized form at `https://anonymous.4open.science/r/dptreerelease` and, together with the frozen protocol (during review, its English rendering and an anonymized copy of the original-language document; Appendix J), the per-query outcome ledgers, a question-to-book manifest (no question or answer text), and a per-question answer-change flag for each budget-curve condition, as supplementary material. From these, the accuracy statistics of the evaluation arms of Table 1 and of the registered contrasts over them (book means, paired differences, and their intervals) are reaggregated by one released command (`python -m scripts.reaggregate_from_release`). Appendix J.2 states what else the release carries and what it withholds, which quantities rederive from it alone, and which need more than it, such as the build cache or the benchmark's own answer key. The component settings frozen across every arm (chunk, summary caps, prompt, encoder, decoding, hardware) are listed in Appendix C. The primary backbone is `meta-llama/Meta-Llama-3-8B-Instruct` (Grattafiori et al., 2024) (Meta Llama 3 Community License Agreement, gated release), used as both summarizer and reader. The backbone robustness check swaps both roles to `Qwen/Qwen2.5-7B-Instruct` (Yang et al., 2024) (Apache-2.0), so every tree is rebuilt with that model's summaries. Both licenses were read from the providers' Hugging Face model cards; neither model is redistributed.

ETHICS STATEMENT

This work uses publicly released benchmarks of published fiction and non-fiction documents; no private or personally identifying data is processed. NovelQA is distributed under Apache-2.0 with gated access; its answer key is provided separately on request, and its terms state that the input data are only for internal evaluation use and ask that they not be publicly spread online. DetectiveQA and LongBench v2 are distributed under Apache-2.0. The ∞Bench dataset README declares Apache-

2.0 (its machine-readable card field is unset; the companion code repository is MIT), and QuALITY ships per-article license metadata. ∞Bench is used for development-stage diagnostics; QuALITY was used for development-stage diagnostics and for an earlier confirmatory attempt halted at its registered sanity gate, whose history is in Appendix J.1.

REFERENCES

Alexander Adkins and Teimuraz Trapaidze. The recall trap: A recall-maximizing retriever configuration reduces issue resolution in fixed-budget code context. *arXiv preprint arXiv:2608.14838*, 2026.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proc. ACL*, pp. 3639–3664, 2025. doi: 10.18653/v1/2025.acl-long.183.

Ananto Nayan Bala. Recall is not enough: A reader-context diagnostic for budget-constrained retrieval-augmented generation. arXiv:2607.00725, 2026.

Cecilia Balocchi and Sara Wade. Understanding uncertainty in Bayesian cluster analysis. arXiv:2506.16295v2, 2026.

Nicola Bariletto, Nhat Ho, and Alessandro Rinaldo. Conformalized Bayesian inference, with applications to random partition models. arXiv:2511.05746, 2025.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system. In *Proc. CAIN*, pp. 194–199, 2024. doi: 10.1145/3644815.3644945.

Sisong Bei, Mikhail L. Arbuzov, Ziwei Dong, Dmitri Kalaev, and Alexey Shvets. Context compression is not one thing: Readable symbolic re-expression vs. coherent summary at matched budget. arXiv:2606.14875, 2026.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24 (4):295–300, 1982. doi: 10.1080/00401706.1982.10487790.

Somnath Bhadra and Michael J. Daniels. Variational Bayes and truncation approximations for enriched Dirichlet process mixtures. arXiv:2603.12427, 2026.

David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Proc. NeurIPS*, 2003.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Henrique Bolfarine, Hedibert F. Lopes, and Carlos M. Carvalho. Lower-dimensional posterior density and cluster summaries for overparameterized Bayesian models. *Statistics and Computing*, 36 (3):107, 2026. doi: 10.1007/s11222-026-10859-z.

Laura Caspari, Kanishka Ghosh Dastidar, Michael Dinzinger, Jelena Mitrović, and Michael Granitzer. When is complex chunking worth it? A multi-objective evaluation of chunking methods at scale. arXiv:2608.16586, 2026.

Camilo Chacón Sartori and José H. García. A fixed-budget, cluster-aware standard for LLM-as-a-judge evaluation: A multi-hop RAG stress test. arXiv:2605.27789, 2026.

Xinyue Chen, Pengyu Gao, Jiangjiang Song, Xinjian Chen, and Xiaoyang Tan. HiQA: A hierarchical contextual augmentation RAG for multi-document QA. arXiv:2402.01767, 2024.

Charbel Chucri, Rami Azouz, and Joachim Ott. Recursive abstractive processing for retrieval in dynamic datasets. arXiv:2410.01736, 2024.

Theodore O. Cochran. Progressive disclosure for LLM-maintained wiki knowledge bases: a preregistered ablation. arXiv:2607.04576, 2026a.

Theodore O. Cochran. Single-round vector RAG vs an LLM-compiled wiki: A preregistered comparison on a small multi-domain research corpus. arXiv:2605.18490, 2026b.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proc. NAACL*, 2021.

Pierpaolo De Blasi, Stefano Favaro, Antonio Lijoi, Ramsés H. Mena, Igor Prünster, and Matteo Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229, 2015.

Xinle Deng, Ruobin Zhong, Hujin Peng, Xiaoben Lu, Yanzhe Wu, Guang Li, Buqiang Xu, Yunzhi Yao, Jizhan Fang, Haoliang Cao, Junjie Guo, Yuan Yuan, Ziqing Ma, Yuanqiang Yu, Rui Hu, Baohua Dong, Hangcheng Zhu, and Ningyu Zhang. MemTrace: Tracing and attributing errors in large language model memory systems. arXiv:2605.28732, 2026.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. arXiv:2404.16130, 2024.

Dongzhe Fan, Zheyi Xue, Siyuan Liu, and Qiaoyu Tan. Do we still need GraphRAG? Benchmarking RAG and GraphRAG for agentic search systems. arXiv:2604.09666, 2026.

Junaid Farooq. Semantic compression trees: Multi-resolution knowledge retrieval via hierarchical semantic residuals. arXiv:2608.21610, 2026.

Naihe Feng, Yi Sui, Shiyi Hou, Ga Wu, and Jesse C. Cresswell. Conformal agent error attribution. arXiv:2605.06788, 2026.

Dennis Frauen, Marie Brockschmidt, Konstantin Hess, Haorui Ma, Yuchen Ma, Abdurahman Maarouf, Maresa Schröder, Jonas Schweisthal, Yuxin Wang, Athiya Deviyani, Sonali Parbhoo, Rahul G. Krishnan, and Stefan Feuerriegel. Causal methods for LLM development and evaluation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (V.2)*, pp. 13170–13177. ACM, 2026. doi: 10.1145/3770855.3818647.

Anupama Garani. A systematic taxonomy of failure modes in retrieval-augmented generation systems. In *Proceedings of the 6th Workshop on Trustworthy Natural Language Processing (TrustNLP 2026)*, pp. 413–424, 2026. doi: 10.18653/v1/2026.trustnlp-main.27.

Ryan Giordano, Runjing Liu, Nelle Varoquaux, Michael I. Jordan, and Tamara Broderick. Measuring cluster stability for Bayesian nonparametrics using the linear bootstrap. arXiv:1712.01435, 2017.

Ryan Giordano, Tamara Broderick, and Michael I. Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19(51):1–49, 2018. URL http://jmlr.org/papers/v19/17-670.html.

Krish Goel and Mahek Chandak. HIRO: Hierarchical information retrieval optimization. arXiv:2406.09979, 2024.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Interpolating between types and tokens by estimating power-law generators. In *Proc. NeurIPS*, pp. 459–466, 2005.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

Kai Guo, Xinnan Dai, Zhibo Zhang, Nuohan Lin, Shenglai Zeng, Jie Ren, Haoyu Han, and Jiliang Tang. Why retrieval-augmented generation fails: A graph perspective. arXiv:2605.14192, 2026.

Shubham Gupta, Zichao Li, Tianyi Chen, Cem Subakan, Siva Reddy, Perouz Taslakian, and Valentina Zantedeschi. Hierarchical retrieval at scale: Bridging transparency and efficiency. arXiv:2502.07971, 2025.

Katherine A. Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *Proc. ICML*, 2005.

Tom Hosking, Hao Tang, and Mirella Lapata. Hierarchical indexing for retrieval-augmented opinion summarization. *Transactions of the Association for Computational Linguistics*, 12:1533–1555, 2024. doi: 10.1162/tacl_a_00703.

Jakir Hossain and Ahmet Erdem Sarıyüce. Core-based hierarchies for efficient GraphRAG. In *Proc. KDD*, pp. 1722–1733, 2026. doi: 10.1145/3770855.3818007.

Amine Kobeissi and Philippe Langlais. Decomposing retrieval failures in RAG for long-document financial question answering. arXiv:2602.17981, 2026.

Alex Laitenberger, Christopher D. Manning, and Nelson F. Liu. Stronger baselines for retrieval-augmented generation with long-context language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proc. EMNLP*, pp. 32559–32569, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1656. URL https://aclanthology.org/2025.emnlp-main.1656/.

Kin Kwan Leung, Mouloud Belbahri, Yi Sui, Alex Labach, Xueying Zhang, Stephen Anthony Rose, and Jesse C. Cresswell. Classifying and addressing the diversity of errors in retrieval-augmented generation systems. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3185–3207, March 2026. doi: 10.18653/v1/2026.eacl-long.147. URL https://aclanthology.org/2026.eacl-long.147/.

Zezheng Lin and Fengming Liu. Position: Mechanistic interpretability must disclose identification assumptions for causal claims. arXiv:2605.08012, 2026.

Peiyang Liu, Xi Wang, Di Liang, and Wei Ye. The laws of context allocation: Causal measurement and closed-loop orchestration in generative search. arXiv:2608.23252, 2026a.

Yan Liu, Xiaodong Xie, Xin Wan, Yi Pan, and Cheng Wang. Enhancing RAPTOR with semantic chunking and adaptive graph clustering. *Frontiers in Computer Science*, 7, 2026b. doi: 10.3389/fcomp.2025.1710121.

Guanran Luo, Zhongquan Jian, Wentao Qiu, Meihong Wang, and Qingqiang Wu. DTCRS: Dynamic tree construction for recursive summarization. In *Proc. ACL*, pp. 10948–10963, 2025. doi: 10.18653/v1/2025.acl-long.536.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426, 2018.

Jeffrey W. Miller and Matthew T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Proc. NeurIPS*, 2013.

Saeedeh Momtazi and Dietrich Klakow. Hierarchical Pitman–Yor language model for information retrieval. In *Proc. SIGIR*, 2010.

Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, 39(6):765–774, 2021. doi: 10.1038/s41587-020-00801-7.

Khai Nguyen and Peter Mueller. Summarizing Bayesian nonparametric mixture posterior — sliced optimal transport metrics for Gaussian mixtures. arXiv:2411.14674v5, 2025.

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

Jingjie Ning and Xueqi Li. Same agent, different answers: A repeat-aware audit of corpus-induced answer churn in retrieval-augmented QA. arXiv:2608.22856, 2026.

Litu Ou and Mirella Lapata. Storyline trees: Hierarchical representations for long-form narratives. arXiv:2606.20900, 2026.

Garritt L. Page, Andrés F. Barrientos, David B. Dahl, and David B. Dunson. Uncertainty quantification in Bayesian clustering. arXiv:2511.16040, 2025.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. QuALITY: Question answering with long input texts, yes! In *Proc. NAACL*, pp. 5336–5358, 2022.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Donna Pham. Indexing long documents for LLM-based analysis. arXiv:2608.21237, 2026.

Saloni Potdar, Daniel Lee, Omar Attia, Varun Embar, De Meng, Ramesh Balaji, Chloe Seivwright, Eric Choi, Mina H. Farid, Yiwen Sun, and Yunyao Li. Comprehensive evaluation for a large scale knowledge graph question answering service. arXiv:2501.17270, 2025.

Renyi Qu, Ruixuan Tu, and Forrest Bao. Is semantic chunking worth the computational cost? arXiv:2410.13070, 2024.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Daeyoung Roh and Donghee Han. Before reasoning can fail: Pre-evidence procedural failures in agentic RAG. arXiv:2608.02011, 2026.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *Proc. ICLR*, 2024.

Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

Sadanand Singh, Allam Reddy, and Manan Chopra. Beyond the reranker: Do RAG retrieval enhancements help once a strong reranker is present? arXiv:2606.28367, 2026.

Anders Søgaard, Daniel Hershcovich, and Miryam de Lhoneux. A two-sided discussion of preregistration of NLP research. In *Proc. EACL*, pp. 83–93, 2023. doi: 10.18653/v1/2023.eacl-main.6.

Zhipeng Song, Xiangyu Kong, Xinrui Bao, Yizhi Zhou, Jiulong Jiao, Sitong Liu, Yuhang Zhou, and Heng Qi. LLM-confidence reranker: A training-free approach for enhancing retrieval-augmented generation systems. *Expert Systems with Applications*, 314:131627, 2026a. doi: 10.1016/j.eswa.2026.131627.

Zhipeng Song, Yizhi Zhou, Xiangyu Kong, Jiulong Jiao, Xuezhou Ye, Chunqi Gao, Xueqing Shi, Yu Wang, Yuhang Zhou, and Heng Qi. CAR: Query-guided confidence-aware reranking for retrieval-augmented generation. arXiv:2605.04495, 2026b.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In *Findings of ACL*, pp. 16596–16616, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.852. URL https://aclanthology.org/2025.findings-acl.852/.

Lin Sun, Linglin Zhang, Jingang Huang, Change Jia, Zhengwei Cheng, and Xiangzheng Zhang. BEAR: Budgeted evidence allocation for multi-document reasoning. arXiv:2601.18116, 2026.

Zhihui Sun. SVD-RAG: Efficient tree-organized retrieval-augmented generation via singular value decomposition. arXiv:2607.10316, 2026.

Yee Whye Teh. A hierarchical Bayesian language model based on Pitman–Yor processes. In *Proc. COLING-ACL*, 2006.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

Ryszard Tuora, Mateusz Galiński, Michał Godziszewski, Michał Karpowicz, Mateusz Czyżnikiewicz, Adam Kozakiewicz, and Tomasz Ziętkiewicz. UnWeaving the knots of GraphRAG — turns out VectorRAG is almost enough. arXiv:2603.29875, 2026.

Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. NovelQA: Benchmarking question answering on documents exceeding 200k tokens. In *Proc. ICLR*, 2025.

Pengyu Wang, Benfeng Xu, Shaohan Wang, Mingxuan Du, Xin Zeng, Huarui Wu, Lei Zhang, and Licheng Zhang. BM25 wins at scale: A scaling study of retrieval-augmented generation paradigms. arXiv:2607.26497, 2026a.

Shu Wang, Yingli Zhou, and Yixiang Fang. BookRAG: A hierarchical structure-aware index-based approach for retrieval-augmented generation on complex documents. *Proceedings of the VLDB Endowment*, 19(9):2358–2371, 2026b. doi: 10.14778/3819518.3819556.

Chunyu Wei, Huaiyu Qin, Siyuan He, Yunhai Wang, and Yueguo Chen. T-Retriever: Tree-based hierarchical retrieval augmented generation for textual graphs. arXiv:2601.04945, 2026.

Xing Wei and W. Bruce Croft. LDA-based document models for ad-hoc retrieval. In *Proc. SIGIR*, 2006.

Xiaoqing Wu, Feifei Li, Haoliang Ming, and Wenhui Que. AuthTrace: Diagnosing evidence construction in thematically dense single-author corpora. arXiv:2605.25382, 2026.

Zhe Xu, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, and Xipeng Qiu. DetectiveQA: Evaluating long-context reasoning on detective novels. In *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=9ExIs5ELlk.

Zhen Xu, Shang Zhu, Jue Wang, Junlin Wang, Ben Athiwaratkun, Chi Wang, James Zou, and Ce Zhang. When does divide and conquer work for long context LLM? A noise decomposition framework. In *Proc. ICLR*, 2026.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. arXiv:2412.15115, 2024.

Yang Yang. Structure-aware semantic chunking with title-chain prefixes: A 1600-query evaluation and the measurement trap in text-transform ablations. arXiv:2608.00824, 2026.

Jiangnan Yu, Kisson Songqi Lin, and Jilong Wu. WhenLoss: Diagnosing write and retrieval bottlenecks in long-context memory systems. arXiv:2605.24579, 2026.

Nan Zhang, Prafulla Kumar Choubey, Alexander Fabbri, Gabriel Bernadett-Shapiro, Rui Zhang, Prasenjit Mitra, Caiming Xiong, and Chien-Sheng Wu. SiReRAG: Indexing similar and related information for multihop reasoning. In *Proc. ICLR*, 2025.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proc. ACL*, pp. 15262–15277, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.814.

Ziwen Zhao and Menglin Yang. Hierarchical abstract tree for cross-document retrieval-augmented generation. arXiv:2605.00529, 2026.

MAP OF THE APPENDICES

Twenty-six appendices follow, in six groups. Each entry names a section and what it holds; nothing is claimed here that is not claimed there.

What was registered. Appendix J summarizes the registration and carries the re-freeze edit and its record; Appendix Q is the claim register, one row per entered claim and what licenses it; Appendix T the registered option-position probe; Appendix Z the registered parser-shape monitor, with every row it flags listed by shape and no completion printed.

The corpora and the indexes. Appendix B records which NovelQA snapshot was run and what the transfer set of Appendix M holds; Appendix K the implementation of the clustering seam itself; Appendix C the per-arm index diagnostics; Appendix U what each arm cost to build and to store; Appendix D the pruned summary layer of the DP arm; and Appendix E the chapter-boundary control, including where its markup was unusable.

The registered read-outs. Appendix L holds the registered secondary read-outs and the control table; Appendix F the length-stratified deltas; Appendix I the complexity subsets; Appendix G the QA budget curve and Appendix H the budget–recall curve; Appendix M the transfer check, cell by cell, on both datasets and both backbones; and Appendix N the two further slices the second registration registered separately.

The stored confidence and how it was diagnosed. Appendix O states the diagnostic procedure and the checks fixed before it ran; Appendix P states the fixed frame for reading the comparisons and its bookkeeping identity and sketches its proof; and Appendix R sets the stored responsibilities against cosine geometry in full.

Read-outs carrying no registered status. Appendix V declares its post-hoc exploratory read-outs, among them §V.1, which runs the diagnostic procedure over the confirmatory corpus after the development result was in hand, computed after the confirmatory verdict was known, registered as no analysis (the second registration's deviation record cites it as prior information for its test H7a), and with nothing outside that appendix depending on it; §V.2, which closes that appendix outside that declaration, rebuilds the DP arm's trees from collapsed-Gibbs partitions and names the designed sets that did not run; Appendix W reports a larger reader's post-hoc read of the frozen indexes, with the second registration's reader run beside it, and Appendix Y what the larger reader's unparsed completions contained; and Appendix X the executed failure audit.

Context and boundaries. Appendix A extends the lineage of §2 and states how that lineage was selected; Appendix S gives the boundaries of §7 in the form their measurements require.

A RELATED WORK: EXTENDED LINEAGE

How this lineage was selected. The lineage below is a selection, not a systematic survey, and no priority claim rests on it. A work enters when it bears on the clustering seam we replace, on the reader-side or budget account we hold fixed, when it names something adjacent that we do not do and so bounds a claim of ours, or when we cite it for a specific fact we use. The project's reference pool is wider than the set cited here; the difference is what we chose to discuss, not what a search returned, and the claims in §1 are phrased so that none of them depends on the set being complete.

The clustering seam. The nearest neighbor to our intervention is query-conditional: DTCRS (Luo et al., 2025) rebuilds the tree per question, seeding cluster centers with sub-question embeddings and gating tree construction on a question-type classifier. It reports large gains over RAPTOR on QASPER and QuALITY but not on NarrativeQA, and its own ablation locates them in the type gate and the question decomposer: removing the gate costs 7.3 and 4.6 F1 on QASPER (Dasigi et al., 2021) across its two readers and removing the decomposer 7.0 and 5.4, while swapping the clustering scheme moves total F1 by at most 0.7 in either direction, and it finds that for extractive and boolean questions recursive summarization "provides no advantage and may even have a negative impact." A dynamic-corpus variant (Chucri et al., 2024) reports parents-per-leaf of 1.0–1.03 across four benchmarks, RAPTOR-style overlapping assignment degenerating to a near-hard partition in practice. The complementary swap also appears: SVD-RAG (Sun, 2026) holds the grouping fixed and replaces the abstractive summarizer with an SVD-extractive one, the mirror image of our inter-

vention, with the clustering never varied. A third direction adds an index rather than replacing one: SiReRAG (Zhang et al., 2025) keeps a recursive-summarization similarity tree and builds a second, relatedness tree from propositions and entities extracted from the text, grouping propositions by shared entities, then flattens both trees into one retrieval pool. Its comparison is therefore between index designs, not between assignment rules at a fixed seam, and its evaluation is on short-passage multi-hop sets rather than at book length; we take the design description from it and none of its reported numbers.

Within a single implementation the spread between grouping rules has been measured directly. A third-party seam-only ablation reports, under a single reader, three rules on three book-length benchmarks separated by at most 0.34 points on DetectiveQA, 1.65 on NovelQA and 2.44 on LiteraryQA, with the ordering changing between datasets: its order-aware variant, which merges only adjacent units but decides each merge by semantic compatibility, leads on NovelQA and trails on the other two (Ou & Lapata, 2026). Those are the margins §1 calls small; that paper's own reading is that no single alternative dominates, and Appendix J records how the same finding bore on our choice of comparator.

Outside the RAPTOR line, hierarchical retrieval at scale reports k-means-based trees beating GMM-based ones on all four text datasets tested (Gupta et al., 2025); a GraphRAG (Edge et al., 2024) variant swaps Leiden communities for k-core hierarchies, citing the exponential multiplicity of near-optimal modularity partitions (proved there for sparse knowledge graphs of constant average degree, most nodes of low degree) as grounds that the communities are inherently non-reproducible (Hossain & Sarıyüce, 2026); an LLM-proposed per-document index schema caps each page at a fixed size, with principled size-adaptation explicitly left to future work (Pham, 2026) (our truncated DP mixture compares grouping rules under a fixed component cap and does not claim that axis); a whole-index swap (grouping rule, rebalancing and abstract cap together) from a Gaussian-mixture tree to an iteratively merged one reports $+1.6$ and $+4.9$ F1 on two multi-hop sets with the retrieval agent and reranker held fixed, and $+1.45$ and $+0.49$ with those components absent (Zhao & Yang, 2026); and at small scale, a heading-derived compression tree reports the ordering our floor comparisons later measure: on 50 QASPER papers, when the relevant document is supplied, its flat dense baseline (F1 0.277) sits above both its own tree (0.274) and RAPTOR (0.272) with neither hierarchy separating from flat under a paired bootstrap; when the document must instead be found, both fall below flat and do separate (Farooq, 2026). Hierarchies also arise outside the recursive-summary construction altogether: levels can be derived by minimizing an information-theoretic objective over an attributed graph, on the argument that fixed per-layer compression quotas "fail to adapt to the data's intrinsic organization" (Wei et al., 2026), or by mapping sentences to paths in a learned discrete index and retrieving subpaths rather than summarizing recursively (Hosking et al., 2024). The nearest case reads the levels off the document itself (layout blocks whose hierarchical level an LLM judges, assembled into a tree that "serves as the role of its table of contents") and takes RAPTOR as a baseline without varying its grouping rule (Wang et al., 2026b); that construction is the one our chapter-boundary control approximates (Appendix E), and it involves no clustering of embeddings at all. Where a hierarchy comes from, and how far its grouping and assignment confidence can be trusted, are separable questions; this paper's grouping audit and confidence diagnostic address the latter. A third axis is neither of those: how the tree is read back. Some work leaves construction to RAPTOR untouched and varies only the query-time traversal, fixing how much context a question returns by a pair of scalar thresholds (an absolute similarity floor for breadth and a parent-to-child similarity gain for depth) against the complaint that a "static definition of 'k' in conventional querying does not account for the variability in information needs" (Goel & Chandak, 2024). What we hold fixed is the traversal *mode*: every tree arm here is read by the same collapsed ranking over a flattened node set, and no arm descends the tree with a per-query stopping rule. Even the confidence-reranked consumer reweights scores inside that same ranking (§3.3) rather than changing how the tree is walked. How a grouping rule fares under a traversal that decides its own depth is therefore untested here.

The structure-versus-flat line. Whether structure earns its cost is contested in the surrounding literature, and the contest runs in both directions. A scaling study that holds the questions and a bedrock of relevant and adversarial documents fixed while varying corpus size across nested tiers reports a lexical baseline ahead of dense retrieval throughout, with the scalable graph variants remaining below it at the shared tiers they complete (Wang et al., 2026a). A paper proposing its own decomposed index reports along the way that plain vector retrieval beats standard graph retrieval on end-to-end question answering (Tuora et al., 2026). A benchmark that standardizes backbone,

retrieval budget and inference protocol before comparing paradigms reports the reverse where the questions are multi-hop (graph structure "remains advantageous for complex multi-hop reasoning" once its offline cost is amortized) and prints that offline cost beside accuracy (Fan et al., 2026). We draw no support from either side: none of these varies our axis, since they swap retrieval paradigms wholesale, or move corpus size, or change the search protocol, where, for the arms at the seam, we hold corpus, reader and packer fixed and change only the rule that groups chunks. Our own comparison is abandoned at its registered design target and licenses no statement in either direction (§4); what this line supplies is a reason to standardize budget and backbone before comparing indexes at all, which is the control we adopt. A neighbouring genre asks our question one operator lower down, pricing the chunking step rather than the index built on top of it: a three-task comparison of semantic against fixed-size chunking finds "the computational costs associated with semantic chunking are not justified by consistent performance gains" (Qu et al., 2024), and a multi-objective evaluation measuring retrieval effectiveness and system cost together reports that "computationally expensive methods rarely provide consistent gains over simpler chunking" and that "chunking should be seen as a multi-objective design decision" (Caspari et al., 2026). Their index is flat and their operator is the chunk boundary, so neither bears on our seam; what they establish is the form of the question, which is why we report construction cost beside accuracy rather than after it.

The size of the published seam effect. The comparator reports $+10.33$ pp over the GMM seam on QuALITY at its peak threshold, with no interval (Liu et al., 2026b); that figure is the clustering step's own contribution, measured on leaves its semantic chunker produced at a fixed threshold, not a re-implementation of ours. Under our frozen harness the same-seam analogue (Leiden minus GMM+BIC) is $+0.0068$ on point estimate (no interval is registered for this pair), and no pairwise difference among the GMM+BIC, Leiden (plain and UMAP-matched) and DP-mixture trees exceeds $+0.0164$, whose interval spans zero (§4). Reader, encoder, dataset, and protocol all differ; we do not adjudicate which factor absorbs the gap.

The non-semantic control has a precedent in RAPTOR's own appendix. RAPTOR's Appendix B (Sarthi et al., 2024) already compares its clustering against the non-semantic alternative at the same seam: it builds "a balanced tree by recursively encoding and summarizing contiguous text chunks," sets the window from "the average cluster size observed in RAPTOR, which is approximately 6.7 nodes" (hence seven), and holds the embedding model, the reader and the collapsed-tree traversal fixed across both arms. Our fixed-block contiguous grouping control shares that recursive form and the order constraint. It differs in how the block size is set (ours is derived from the frozen token budget rather than matched to a realized cluster size, and the count-matched twin is the separate *contiguous at DP capacity* row) and in that no embedding enters the grouping at all. What separates the two accounts is the resolution, not the design: that appendix reports 56.6 against 55.8 on QuALITY under a single reader with no interval and reads the gap as an improvement, while the same control, read at the registered interval convention on the primary backbone (§4), is not separated from the DP arm, with the caveat that a difference of the size their table reports would not be resolvable at our sample either (Appendix S), so the two readings disagree about what counts as resolved rather than about a measured quantity. The precedent is for the control, not for the finding.

Bayesian nonparametric precedents before retrieval augmentation. Bayesian nonparametric methods for hierarchical structure long predate retrieval augmentation. The nested Chinese restaurant process learns topic hierarchies (Blei et al., 2003), the hierarchical Dirichlet process shares mixture components across groups of data (Teh et al., 2006), and Bayesian hierarchical clustering decides which clusters to merge by marginal likelihood rather than by an ad-hoc distance (Heller & Ghahramani, 2005). For language, Pitman–Yor priors were adopted because they generate the power-law word frequencies that text exhibits (Goldwater et al., 2005; Teh, 2006). In each of these the learned structure or model is the object of interest: a topic hierarchy, a clustering, a language model. What separates Pitman–Yor priors from the Dirichlet process, including the growth in the number of clusters, is characterized within the Gibbs-type family (De Blasi et al., 2015). Probabilistic text models have also served retrieval directly, as the document model that scores a query, from LDA-based document models for ad-hoc retrieval (Wei & Croft, 2006) to Pitman–Yor smoothing of document language models (Momtazi & Klakow, 2010). Here the partition is an instrument, one construction step of a summary index judged by what a fixed retrieval budget can hold, inside a

recursion that stays RAPTOR's (§3.2); none of these models enters as an arm, and the prior we fit is part of that lineage rather than a contribution of this paper.

What we store is not a partition posterior to summarize. Summarizing a posterior over clusterings is a developed problem in Bayesian cluster analysis: a point estimate carrying a credible ball (Wade & Ghahramani, 2018), or a set of representative partitions when the posterior is multimodal (Balocchi & Wade, 2026). Both summarize a partition posterior explored by sampling, where dependence between assignments is what the summary is built to describe. The index stores neither: a mean-field variational fit factorizes into independent per-node marginals, so it carries none of that dependence, and we never summarize it under a loss function. That second clause is what separates the index from the wider line as well: post-processing that turns any sampled partition posterior into a credible set (Page et al., 2025), a conformal construction attaching what its authors call assumption-free posterior coverage to a region of partition space (Bariletto et al., 2025), a shift of the summary target from the induced partition to the mixing measure (Nguyen & Mueller, 2025), and a projection of an overparameterized fit onto a lower-dimensional surrogate carrying its own uncertainty (Bolfarine et al., 2026) all take a posterior meant to be summarized as their input, and we compute none of them. Even the variational–sampling pair we use as a diagnostic axis is an object of study in that literature rather than an instrument, with a truncated variational estimator derived and offered as initial values for a blocked Gibbs sampler (Bhadra & Daniels, 2026); we run the two to compare partitions, not to improve either. The diagnostic of §6 accordingly compares two spreads rather than summarizing one posterior, namely movement across row-permuted refits (§6) against movement under parameters drawn from the fitted posterior, and exact candidate-wise masses concentrate as well (§6; Appendix O), without separating model from approximation. It is a narrower question than summarizing a partition posterior, and does not stand in for one. A separate line measures the stability of a Bayesian nonparametric clustering directly, approximating the bootstrap by local data perturbation so that stability can be assessed without rerunning the fit on every resample (Giordano et al., 2017); its varying quantity is the data, where the two spreads we compare hold the data fixed and vary the row order of the refit or the parameters drawn from the fitted posterior, so it bounds a claim we do not make rather than supplying one we do.

The reader-side line. With GPT-4o readers, ordering retrieved chunks by document position and dropping the hierarchy outperforms RAPTOR on several long-document benchmarks (Laitenberger et al., 2025), and a reader-context diagnostic finds that an evidence-packing policy's edge over naive packing stays significant at every reader scale it tests, while its edge over a hand-tuned heuristic is absorbed once the reader passes the smallest rung, a smallest-reader win its own authors decline to call established (Bala, 2026). We differ from the latter in object of study: that work measures answer survival under a packing *policy* on flat multi-hop corpora with tight budgets, whereas we hold a frozen collapsed packer fixed and vary the *index topology* it packs from; redundancy between ancestors and descendants, and summary-versus-leaf displacement, are only definable once the index has levels. A budget-allocation line accepts the budget as the binding constraint and optimizes its split: deciding, over an offline hierarchical index, which documents and which levels of granularity receive the limited query-time evidence slots (Sun et al., 2026). A concurrent line on the same axis replaces relevance proxies with a leave-one-out causal probe that isolates generative reliance and calibrates the structural dilution of attention, then allocates budget on a deconfounded factorial grid (Liu et al., 2026a); their quantity is dilution of attention across a context, ours is a summary node occupying a slot a leaf would otherwise take, with the construction operator held fixed; the probe is theirs. Our own recall read-outs stay descriptive, since maximizing recall under a fixed budget can reduce end-task success (Adkins & Trapaidze, 2026). A terminological note travels with this line. The word displacement appears in both settings with different referents: there it is positional shift within a context window, ruled out as an artifact of the measurement; here it is one candidate occupying a retrieval slot instead of another (a channel the reranker-ablation literature has already described, observing that an off-topic summary node can displace the answer-bearing chunk and offering that as a plausible source of the dev–test sign change in their RAPTOR ablation (Singh et al., 2026)) and it is a quantity we set out to measure.

Failure-decomposition catalogues. A qualitative catalogue derives seven RAG failure points from field case studies (Barnett et al., 2024); a diagnostic benchmark separates evidence-construction failure from synthesis failure across retrieval paradigms (flat, memory, graph, wiki)

Table 5: Seam-level comparison of hierarchical-index lineages. Dashes mark axes a system does not vary or report; the last column names the evaluation each paper itself reports (HiQA also evaluates on its self-built MasQA). Where the token scales matter, we measure QuALITY passages at 2.4–8K tokens and NovelQA books at 86K–2.58M, two regimes separated by a gap, not a continuum.

	Grouping rule	Cluster count	Assignment posterior (and where it is kept)	Long-document evaluation
RAPTOR (Sarthi et al., 2024)	GMM on UMAP-reduced embeddings	BIC sweep	—	NarrativeQA, QASPER, QuALITY
+SC+AGC (Liu et al., 2026b)	Leiden on a k-NN graph	quality-function optimum	none (hard partition)	QuALITY, QASPER
N-RAPTOR, an ablation variant in (Ou & Lapata, 2026)	order-aware adjacent merging (semantic)	—	none	NovelQA, DetectiveQA, LiteraryQA
HiQA (Chen et al., 2024)	hierarchical contextual augmentation of chunks; its grouping is called soft from PCA/t-SNE plots alone	—	no quantitative object	MasQA (self-built), NarrativeQA, QASPER
DTCRS (Luo et al., 2025)	query-conditional GMM, sub-question seeds	sub-question count at the first layer; BIC above it	transient (tree rebuilt per query)	QASPER, QuALITY, NarrativeQA
SCT (Farooq, 2026)	document heading levels (font size in PDFs); no embedding clustering	—	n/a	QASPER (50 papers)
Ours	truncated VB-DPMM on the same two-stage UMAP	effective components under a fixed cap	responsibilities Γ in the build sidecar; the index keeps, per node, only the derived scalar c_i, consumed by the reranking variant	NovelQA (confirmatory); DetectiveQA, LongBench v2 (transfer)

swapped wholesale rather than varied one factor at a time (Wu et al., 2026); an oracle-based analysis decomposes retrieval failure into document-, page-, and chunk-level discovery on financial QA, with no reader-side term in the decomposition (Kobeissi & Langlais, 2026); a telescoping identity factors divide-and-conquer fidelity into task, model, and aggregator noise, a decomposition holding by construction rather than by causal attribution (Xu et al., 2026); and a fixed-budget diagnostic splits write-side from retrieval-side bottlenecks in long-context memory systems; we take the split as a naming of channels, not a causal attribution (Yu et al., 2026). Among catalogues that do not claim exclusivity, unlike Roh & Han (2026): a sixteen-type taxonomy states that its types are not mutually exclusive and annotates the earliest erroneous stage with one or more labels inside it (Leung et al., 2026), a thirty-three mode review organizes by distinct causal mechanism without claiming a partition or stating an assignment rule (Garani, 2026), and a deployed question-answering service routes failures to seven query-understanding and knowledge-graph buckets without stating whether they are mutually exclusive (Potdar et al., 2025). A further line locates failure inside the model rather than in the pipeline, tracing attribution graphs over transformer layers and contrasting the information flow of correct and failed predictions (Guo et al., 2026).

Executed per-record attribution. The audit that Appendix P defines is executed post hoc in Appendix X, on the subsample of questions whose gold spans can be located; a contemporaneous line builds a fuller instrument: Deng et al. (2026) transform memory pipelines into executable evolution graphs and add an attribution method that traces operation subgraphs to locate the cause of a failed case. Their pipelines are memory systems for long-horizon reasoning, not a clustered summary tree built once and read many times, so the instrument does not port directly; the methodological point does. Conformal prediction gives finite-sample, distribution-free coverage over contiguous

step intervals in multi-agent trajectories, localizing which step failed rather than recording which components a contrast changes together (Feng et al., 2026). Three of the lines already placed above execute attribution, or its causal analogue, as well: the leave-one-out budget probe (Liu et al., 2026a), circuit tracing over transformer layers (Guo et al., 2026), and, nearest of all in mechanism, the fixed-priority bucketing of 3,807 wrong cases among twelve thousand paired trajectories (Roh & Han, 2026), whose device is the one Appendix P defines and whose disjointness its authors scope as operational rather than semantic. Theirs are executed as the study's object; ours is executed after the fact, on 49.2% of the eligible questions, and carries no registered status. What is specific to a leveled index is the displacement class, which has no counterpart in a flat retrieval trace, and what we add is a reading fixed in advance of which components each contrast changes together. The shares in Appendix X are read under that fixed reading, not used to establish it.

The pre-registration lineage. Pre-registration is the subject of a standing methodological discussion in NLP (Søgaard et al., 2023), with recent LLM-era instantiations (Cochran, 2026a; Chacón Sartori & García, 2026), the latter pairing pre-registration with a fixed evaluation budget and cluster-aware (clustered-error) inference for judge-based RAG scoring. Two further instances are close in form rather than in subject. One reports its registered predictions in registration order and states plainly that one of them was refuted (Cochran, 2026b), the reporting shape a registered prediction takes when the data do not support it. The other registers a measurement rather than a result, asking whether aggregate accuracy can hide answer-level change when gains and losses cancel (Ning & Li, 2026); we did not measure repeat-adjusted excess answer churn, though Appendix V reports correctness flips between arms; we read the work as an argument for registering what a metric can conceal, not as an account of our own outcome. A third states its registered depth-interaction hypothesis and reports it null, having "filed a pre-registration prior to data collection" (Bei et al., 2026); its budget is matched by token count, ours by cluster capacity, so the shared phrase names different controls. A related caution arrives from outside this literature: an ablation that strips the context its own relevance judgements depend on degrades the measurement rather than the system (Yang, 2026). We came upon that work after our own protocol was frozen, so it is not a reason for our design; we record it because it names the hazard our fixed-in-advance controls and separately reported post-hoc audit (Appendix X) are exposed to. What we add is not the practice but its object: a cluster-count-matched anatomy of a single clustering seam.

B Dataset snapshot provenance

NovelQA. The confirmatory benchmark (Wang et al., 2025) is pinned to Hugging Face dataset revision `60f8de6a`, downloaded 2026-07-06; all counts refer to this snapshot and we do not track upstream updates. Upstream sources disagree about the public release's size: the benchmark paper reports 65 public books (1,640 public questions) of 89 total, the dataset card implies 61, and the pinned snapshot's public-domain folder contains 60 books / 1,526 questions; its book metadata marks 61 books public domain, our 60 plus the demonstration book. Of the 63 IDs B00–B62, three are outside that folder: B06, which has no entry in the release's book metadata; B45, which the metadata lists as copyright-protected and so outside the public set; and B30, which exists only as the demonstration book, which our protocol excludes by rule. We do not speculate about the question-count difference.

Transfer set. The DetectiveQA English subset (Xu et al., 2025) contributes 600 questions over 62 detective novels, of which 446 are AI-supervised and 154 human-annotated. That split is the snapshot's own annotator label, pinned by our loader; the DetectiveQA abstract instead reports 1,200 human-annotated questions across its Chinese and English subsets. As with NovelQA above, we report the snapshot we ran on and do not reconcile the two counts. The LongBench v2 (Bai et al., 2025) slice contributes 37 questions selected by our own filter (single-document, non-fiction domains, medium and long lengths); the label "non-narrative" is ours, not the benchmark's.

C Per-arm index diagnostics

Component settings, frozen across every arm. Leaf chunks are ∼100 tokens; summaries are capped at 256 tokens with a 3,500-token summarization context; the sum-

marization prompt is RAPTOR's verbatim; the encoder is a fixed sentence-transformer (`multi-qa-mpnet-base-cos-v1`, 768-d); where an arm reduces (§3.2), UMAP reduces to 10 dimensions with the frozen seed, single-threaded for determinism; and generation is served with greedy decoding on a single A100 80GB (the 70B readers of Appendix W on four).

Mixture prior hyperparameters, two of which are read from each book. "Frozen across every arm" is a statement about the arms; it is not a claim that the mixture's prior is independent of the data, and for two hyperparameters it is not. The variational fitter is called with seven explicit arguments and inherits every other setting from the pinned library version. The weight prior is a Dirichlet process at $\alpha_0 = 1$. Its normal–inverse-Wishart base measure is inherited whole: $\kappa_0 = 1$ and $\nu_0 = 10$ (the reduced dimension) are the same for every book, while μ_0 and Ψ_0 are set by the library to the mean and the empirical covariance *of the embeddings being clustered*, so both differ from book to book. The registration enumerates all four and marks the last two data-dependent, with the library version at which they were read; nothing here departs from it, and the collapsed-mass diagnostic of Appendix O is scored at the same κ_0, ν_0 corner. What this costs is comparability of the prior across books, and it does not reach across arms because only this arm has a prior at all: the mixture baseline is a maximum-likelihood `GaussianMixture` fit of the same library on the same coordinates, with no prior hyperparameters to read.

What the arm descriptions spell out. The first matched Leiden control is a global-UMAP-stage-matched control (the same global-UMAP stage the mixture arms use); the second is matched on the full two-stage UMAP skeleton (global reduction, then per-community local reduction). Contiguous grouping takes adjacent chunks with no embedding or clustering input to the grouping. The document-order control retrieves the identical node set and repacks it in document order instead of retrieval-rank order. How far that repacking moves the context was measured once before the main run, in a smoke test outside the confirmatory run: on the flat index of the 12 NovelQA books with the fewest characters (chosen for build time, not at random), all of them confirmatory books, and their 326 questions, all of them confirmatory questions, with a scripted reader, the rank correlation τ_a between the retrieval-rank order and the document order of the packed nodes had median $+0.0158$ (range -0.4211 to $+0.6211$), and $\tau_a > 0.5$ on 3 of 326 questions, so the two orders were close to unrelated. Renumbering the packed nodes can push the context past its token budget, in which case the lowest-ranked node is dropped and the control reads fewer nodes than flat dense; in the smoke test that overflow was under 1% (0/326 measured, rule-of-three upper bound 0.92%). The oracle-evidence reference is conditional: questions without gold evidence spans are excluded from its delta.

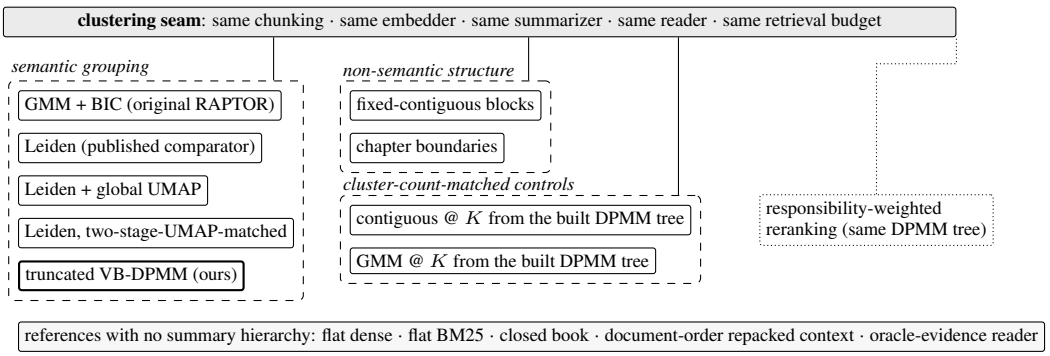

Figure 5: The design space, organized by what each contrast compares. Within the semantic family the arms replace the grouping operator (DPMM vs. GMM/Leiden); Table 30 lists the representation and fitting differences that travel with each swap. The semantic family vs. the contiguous and chapter controls compares meaning-based grouping with structural grouping; the cluster-count-matched pair imports the DP arm's realized per-layer cluster counts, intended to separate *how many* groups from *which* groups (what it does and does not identify: §4); the Leiden trio varies the reduction pipeline; tree arms vs. the flat references compare a hierarchical index with one that stores no summaries; the oracle bounds the retrieval side. The dotted variant changes ranking only, on the identical tree. Every other arm above the reference band shares the frozen non-seam pipeline.

Table 6(a) reports the build-side footprint of every index-building arm. All arms share the identical leaf layer by construction (the pipeline is frozen below the clustering seam), so the arms differ only in the summary levels built on top of it.

Summarizer-input truncation. Building a summary node truncates its input when the concatenated child text exceeds the summarizer's context window. The registered co-report (count basis and token basis, always together, aggregated over (book, layer) cells where both bases are recorded) appears in Table 6(b) for every index-building arm: the share of summary nodes whose input was truncated (count) and the share of input tokens cut (token), over all layers, at the first summary layer, and at the top layer. Truncation tracks cluster size, not tree depth: the raw and the single-stage-UMAP Leiden arms compress the corpus into few, large clusters whose top-layer nodes read the most text and lose the most, while the cluster-count-matched and contiguous constructions truncate little or not at all. These shares are pooled over books, which hides how unevenly the loss falls: over all layers the raw and the UMAP-matched Leiden arms exceed 1% of member tokens on 60 and 60 of the 60 books, the mixture arm on 26 and ours on 2, with per-book median losses of 0.204, 0.146, 0.005 and 0.002. The mixture arm's pooled share and ours therefore rest on a few books, including the outlier the registered sensitivity co-report excludes.

Table 6: Index-building arms; both panels cover the same arms in the same order. **(a) Footprint**: summary nodes, summary tokens (the stored summary-token footprint: the token count of the summary nodes, not the LLM cost of building them; Appendix U reports build wall-clock separately), and tree depth (min–max layers across the 60 books). The cluster-count-matched variants reproduce the DP arm's cluster-count budget exactly; the two-stage-matched Leiden variant matches representation, not capacity (its summary-node count is 56.0% of the DP arm's); the contiguous control and chapter arms build shallower trees. **(b) Summarizer-input truncation**, on the count basis (share of summary nodes truncated) and token basis (share of input tokens cut), reported together, as registered. "Top" is each arm's highest layer, named per arm, over the books that reach it.

(a) Index footprint

Arm	Summary nodes	Summary tokens	Layers
GMM+BIC	19,627	4,822,747	4–5
GMM+BIC at DP capacity	24,882	5,884,945	4–5
Leiden	4,855	1,238,489	3–3
Leiden, UMAP-matched	5,495	1,402,350	2–3
Leiden, two-stage-UMAP matched	13,934	3,534,905	3–4
DP mixture	24,882	6,205,729	4–5
Fixed-block contiguous	6,265	1,602,734	3–4
Contiguous at DP capacity	24,882	6,178,512	4–5
Chapter-boundary	7,243	1,849,571	3–4

(b) Summarizer-input truncation

	All layers		Layer 1		Top layer			
Arm	count	token	count	token	count	token	layer	books
GMM+BIC	0.043	0.068	0.053	0.086	0.000	0.000	L4	8
GMM+BIC at DP capacity	0.011	0.013	0.004	0.009	0.000	0.000	L4	4
Leiden	0.392	0.207	0.374	0.190	0.798	0.420	L2	60
Leiden, UMAP-matched	0.283	0.150	0.257	0.122	0.937	0.467	L2	57
Leiden, two-stage-UMAP matched	0.044	0.047	0.000	0.000	0.500	0.259	L3	5
DP mixture	0.027	0.044	0.022	0.052	0.000	0.000	L4	4
Fixed-block contiguous	0.000	0.000	0.000	0.000	0.000	0.000	L3	4
Contiguous at DP capacity	0.000	0.000	0.000	0.000	0.000	0.000	L4	4
Chapter-boundary	0.000	0.000	0.000	0.000	0.000	0.000	L3	5

Cluster-count-matched import. The two cluster-count-matched arms import the DP arm's per-layer cluster-count budget rather than choosing their own: at each layer the imported target equals the DP arm's realized count for that book. The per-layer evidence that the import held exactly is identical for both arms and reads, over the 60 books: layer 0 mean $k = 363.27$ (range 99–2182), layer 1 mean $k = 45.33$ (13–256), layer 2 mean $k = 5.92$ (2–33), and layer 3 (reached by 4

books) mean $k = 2.75$ (2–4). The target equalled the effective count in every cell, and the clamped, fallback, and degenerate-fit counters are all 0: no book required clamping or fell back to a non-imported count.

Retrieval composition, recall, and oracle gap, all evaluation arms. Table 7 completes the per-condition read-outs that §4 quotes selectively: the summary share of retrieved slots, conditional verbatim recall of gold-evidence spans, and the oracle-context gap (condition minus oracle-evidence reader, negative means below the oracle), for every evaluation arm of Table 1.

Table 7: Per-condition retrieval composition (summary share of retrieved slots, pooled over every slot the condition retrieved, unlike the book-mean share of Table 12), conditional verbatim gold-evidence recall (snippet-pooled: every gold snippet of an evidence-bearing question with a reconstructable reader context weighted equally, unlike the book-mean aggregation of Table 12), and oracle-context gap with 95% CI. The closed-book row retrieves nothing and the flat row retrieves only leaves, so their composition cells are fixed by construction; recall conditions on evidence-bearing questions whose reader context could be reconstructed, not on snippets known to be recoverable. Every row that carries a recall value scores the same 2,573 snippets, drawn from the same 1,519 questions: the evidence map is shared across conditions and no condition lost a reconstructable context, so the rows are comparable cell-for-cell down the column. The gap column is computed on common support: the 7 questions with no gold evidence spans are dropped from the condition and from the oracle alike, because the oracle retrieves nothing for them.

Condition	Summary share	Verbatim recall (pooled)	Oracle gap [95% CI]
Closed book	—	—	-0.3723 $[-0.4203, -0.3243]$
Flat dense	0.0000	0.0731	-0.1845 $[-0.2221, -0.1468]$
GMM+BIC	0.2523	0.0661	-0.1952 $[-0.2308, -0.1596]$
Leiden	0.1028	0.0700	-0.1879 $[-0.2250, -0.1508]$
Leiden, UMAP-matched	0.1130	0.0692	-0.1851 $[-0.2182, -0.1520]$
DP mixture	0.2745	0.0618	-0.1803 $[-0.2145, -0.1460]$
Confidence-reranked	0.2913	0.0610	-0.1978 $[-0.2315, -0.1641]$

What the Leiden arms adopt, and where they depart. We adopt the published comparator at the algorithm level and re-implement it inside our frozen harness, following their released code: a cosine k-NN graph whose k starts at 15 and grows by 5 per layer above the leaves, the RB-configuration quality function at resolution parameter 1.0, and recursive re-clustering of a community above 100 nodes under the same k unless that community is the whole input, in which case their deadlock fallback keeps it whole. Their code holds the resolution parameter fixed, so the decay schedule their text describes is never executed; we follow the code and disclose the gap. Two deviations are ours: the partition is seeded (theirs is not) and edges of non-positive weight are dropped, which is inert in their regime, where the graph is built on raw embeddings and the weights stay positive; on the UMAP coordinates the variants below cluster, non-positive similarities are not excluded and the drop is unmeasured. Neither the resolution parameter nor the k schedule alone fixes the comparator's cluster count: the oversized-community threshold splits what they leave, so the count is a joint product of the three. The original stack pairs a different reader and encoder, so published absolute numbers are not comparable to ours and are not reused. Because the mixture arms cluster after a UMAP reduction while the published form clusters raw embeddings, the UMAP-matched variant exists to bound how much of any gap the preprocessing explains; the registered conjunctive predicate targets the published raw-embedding form. The two-stage-matched variant aligns the *representation* pipeline of the mixture arms (both UMAP stages) so that the remaining factor is each algorithm's clustering rule together with its own cluster-count convention; Leiden carries no per-stage k convention and we introduce none, so we do not describe the residual factor as the algorithm alone. Its second stage measurably runs: 5,112 of 5,708 stage-1 communities (89.6%) are re-partitioned locally, so the variant does not degenerate to its single-stage form.

C.1 Rebuilding the index under other construction seeds

The registered comparison of §4 speaks to the index set realized at a single construction seed. A second registration, frozen after the confirmatory run and adjudicated by its own registered rules

(book-resampling bootstrap 95% intervals and per-book sign tests, with the book as the unit), rebuilds the GMM+BIC, Leiden, UMAP-matched Leiden and DP-mixture trees (Table 8) under 4 further seeds. The read-out below asks whether the same rule rebuilds the same index, whether that reaches accuracy, and which answers move. Each is scored on the primary backbone and on those trees; the no-tree floors, the repacking control, the second backbone and the remaining registered arms, including the contiguous and chapter controls, were not rebuilt. For those two controls the registration records that the trees do not depend on the construction seed, a structural fact that we state here rather than as a row of Table 8. The sixteen rebuilt ledgers carry three code revision strings, two of them a commit plus an uncommitted delta; Table 20 lists them in its rebuild row, with what the committed span between the bases reaches and why the rest is not auditable.

Does the same rule rebuild the same index? The partition-agreement sample is locked to a 12-book draw, stratified by token length, whose selection seed the authors did not choose and which is fixed before any count is visible; that seed is derived from the draw's population, which, as the registration records, was narrowed to exclude the development-track books after an initial draw over the whole corpus had been seen. Each book contributes the median over its 10 seed-pair agreements between the leaf layer and the first summary layer, and the mask (leaves singly assigned under every arm and seed) is the one the adjudicated artefact reports as its headline, a choice its deviation record lists as not registered; the same record notes that the choice was made with the draw partly complete and some books' values already seen, and gives its direction as unfavourable to us. Every arm falls below perfect agreement on every book of the draw (Table 8), which is the registered sign test at 12 of 12 books for each arm, with per-arm means from 0.576 to 0.756. The registered 95% bootstrap intervals over 20,000 book resamples are $[0.551, 0.602]$ for GMM+BIC, $[0.678, 0.735]$ for Leiden, $[0.734, 0.779]$ for its UMAP-matched variant and $[0.674, 0.697]$ for the DP arm, whose mean is 0.685. The registered intersection-union test over arms, bound by the DP arm, gives $p = 0.00024$. The null behind this p is that some arm rebuilds its partition unchanged: each arm's sign test takes perfect median seed-pair agreement on a book as its null, and the union of these nulls is rejected only when every arm's null is. The registered family carrying this test has a second member not yet read out, so no family-level adjustment is reported here. Seam instability is a property of every semantic grouping rule we rebuilt, not of the mixture alone; which branch of the registration's interpretation rule applies is not decided here, since that rule compares the arms after a cluster-count correction this paper does not report for the draw. By rule the draw's population excludes the corpus's longest book, which sits in the development track, and the registration cites an earlier draft of this paper as stating that seed instability was worst on that book. Instability found on this draw is therefore found on the easier sample, and its absence here would not have spoken for the longest books.

Does it reach accuracy? Per book, the sample standard deviation ($n - 1$ denominator) of accuracy across the 5 seeds has a median over the 54-book census, the confirmatory corpus with the development books held out, of 0.0368 to 0.0480 depending on the arm (Table 8); the per-book values behind these medians, standard deviation and range for every census and development book, are printed unsummarized in Table 9, as the registration requires. The registration names both yardsticks in advance, and its deviation record, while the further seeds were still being rebuilt and part of their accuracy data was on disk, fixed that both observed gaps are reported rather than one chosen once the medians are known: the design target $\delta^* = 0.034$ the study was sized for, and both observed arm gaps, 0.0096 against the primary comparator and 0.0164 against the mixture. The spread across seeds is at the scale of the registered design target; both arm gaps, a different estimand from the spread, are small. The registration gives this descriptive path a second item, a variance-component decomposition; beside the path it asks that two other paths be printed, a random-intercept model over books and seeds and a paired bootstrap over seed means; and it fixes the descriptive path in advance as the headline. The decomposition and the model are below, in Table 10 and Table 11; the bootstrap is not reported, for the reason given with them.

Which answers move? Between seeds, the reader's chosen option changes, not necessarily its correctness, on 17.0 to 25.5% of comparable seed pairs (pairs in which both runs chose an option) over the 301 questions of the locked draw, and on 13.7 to 20.0% over the 1,298 questions of the widened census; the per-arm cells are in Table 8. The retrieved leaf sets themselves overlap, over the widened census, at a median Jaccard of 0.93 to 0.95, computed over leaf identifiers only, so for the typical book wholesale changes in which leaves were retrieved directly do not account for the

Table 8: Rebuilding under 4 further construction seeds, primary backbone. Agreement is the per-arm mean of per-book median seed-pair ARI on the locked 12-book draw. Accuracy spread is the per-book sample standard deviation ($n-1$ denominator) across the 5 seeds, median over the 54-book census. Flip rates count changes of the chosen option (correctness need not change) and are reported for both the locked draw and the widened census, which are different samples. The development rows are the registration's 6-book development track, a different sample by a different rule, never pooled with the census; its agreement column is empty because partition agreement was measured on the locked draw only; the descriptive extension of agreement to the census that the registration provides for is not reported, a departure from it. The yardstick row places the registered target and both observed arm gaps in the spread column, beside the spreads they are compared with. In the development rows the two flip columns are the track's rate and the longest book's, as their header says.

Arm	Agreement	Accuracy sd	Flip % (draw)	Flip % (census)
GMM+BIC tree	0.576	0.0447	24.5	20.0
Leiden tree	0.706	0.0372	17.0	13.7
Leiden tree, UMAP-matched	0.756	0.0368	18.8	14.3
DP-mixture tree	0.685	0.0480	25.5	19.8
Yardsticks	—	$\delta^* = 0.034$ 0.0096, 0.0164	—	—
Development track, 6 books by length rank; not pooled with the census				
	Agreement	*sd, 6 books*	*Flip %, 6 books*	*Flip %, longest book*
GMM+BIC tree	—	0.0361	19.6	12.4
Leiden tree	—	0.0205	14.6	18.5
Leiden tree, UMAP-matched	—	0.0342	14.4	12.6
DP-mixture tree	—	0.0456	19.6	17.6

flips; the summary nodes among the retrieved slots (Table 7) are outside this statistic (Appendix V). This re-scopes the registered comparison rather than overturning it.

The development track, beside the census. The registration's development track holds 6 books selected at equal rank spacing on token length rather than by draw, the corpus's longest book among them, and it places both tracks in the same table without pooling them. Read with the same instrument, the per-book accuracy spread across the 5 seeds has a median of 0.0205 to 0.0456 across arms over the track, and of 0.0208 to 0.0720 on the longest book alone. The chosen option changes on 14.4 to 19.6% of comparable seed pairs over the track's 228 questions, and on 12.4 to 18.5% over the longest book's 34. Leaf-set Jaccard medians run from 0.93 to 0.95. Partition agreement was measured on the locked draw only, so the development rows of Table 8 carry none.

What else the registration asks this section to print. The registration also asks for all three leaf masks with the unmasked ranking beside any corrected one, for the draw it replaced, and for a single-book result seen before its rules were written, printed beside these results. Under unmasked labels the agreement column of Table 8 would read 0.411 (GMM+BIC), 0.579 (Leiden), 0.634 (UMAP-matched) and 0.480 (DP), and on singly assigned leaves 0.448, 0.579, 0.634 and 0.636; every arm falls below perfect agreement on every book of the draw under each mask. Over the narrowed population the draw was repeated once, after a repair of its selection generator; the pre-repair draw held B04, B08, B13, B22, B36, B38, B51, B54, B55, B60, B61, B62, 4 of them in the locked draw (the initial whole-corpus draw is not printed), and agreement was measured on the locked draw only. The single-book result, on the robustness backbone at three seeds, reads 0.3098, 0.3744, 0.4143 and 0.2031 in the same arm order, as raw mean agreement, a different statistic on a different backbone; the registration records that on common-support leaves with its cluster-count correction the arms reorder, but that mask covers 18.1% of the book's leaves, below the registration's 20% coverage gate, so under the registration's own rule no ranking is read from it. The registration further defines an answer flip as a change of correctness and retrieval overlap over every retrieved node, where the rates above count changes of the chosen option and the overlap counts leaves only; it registers per-seed summary-token sums, a calibrated share of reassigned leaves and the spread of the first-layer cluster count, none of which this paper reports; and it asks that a quoted index-size ratio name its seed: the two summary-token ratios this paper quotes, $5.0\times$ and $1.29\times$, are seed-224 values.

The registered variance decomposition. The descriptive path has a second item: decompose the variance into book, seed and arm parts and answer, with a number, whether the difference between

Table 9: Per-book spread of accuracy across the 5 construction seeds, primary backbone, printed unsummarized as the registration requires. Each cell gives, for the book of its row and the arm of its column, the standard deviation (sd) and the range (largest minus smallest seed accuracy) over the 5 seeds. The census rows are the 54 confirmatory books whose per-arm medians Table 8 reports; the development rows are the registration's 6-book track, a different sample by a different rule, never pooled with the census. Columns follow the arm order of Table 8.

Book	GMM+BIC tree sd / range	Leiden tree sd / range	Leiden tree, UMAP-matched sd / range	DP-mixture tree sd / range
B00	0.0645 / 0.1579	0.0471 / 0.1053	0.0577 / 0.1579	0.0471 / 0.1053
B01	0.0297 / 0.0794	0.0261 / 0.0635	0.0309 / 0.0794	0.0481 / 0.1111
B02	0.0490 / 0.1200	0.0400 / 0.0800	0.0490 / 0.1200	0.0456 / 0.1200
B03	0.0426 / 0.0952	0.0476 / 0.0952	0.0583 / 0.1429	0.0852 / 0.1905
B04	0.0214 / 0.0563	0.0427 / 0.0845	0.0403 / 0.0986	0.0452 / 0.1127
B05	0.0975 / 0.2500	0.0822 / 0.2000	0.0447 / 0.1000	0.0500 / 0.1000
B07	0.0372 / 0.1053	0.0577 / 0.1579	0.0686 / 0.1579	0.0440 / 0.1053
B08	0.0456 / 0.0833	0.1264 / 0.3333	0.0373 / 0.0833	0.0950 / 0.2500
B09	0.1517 / 0.3000	0.1342 / 0.3000	0.1304 / 0.3000	0.2168 / 0.5000
B10	0.0418 / 0.1000	0.0354 / 0.1000	0.0418 / 0.1000	0.0837 / 0.2000
B11	0.0466 / 0.1071	0.0598 / 0.1429	0.0253 / 0.0714	0.0696 / 0.1786
B12	0.0000 / 0.0000	0.0000 / 0.0000	0.0000 / 0.0000	0.0000 / 0.0000
B13	0.0506 / 0.1333	0.0365 / 0.1000	0.0447 / 0.1000	0.0494 / 0.1333
B14	0.0686 / 0.1579	0.0600 / 0.1579	0.0600 / 0.1579	0.0577 / 0.1579
B16	0.0548 / 0.1000	0.0000 / 0.0000	0.0000 / 0.0000	0.0548 / 0.1000
B17	0.0496 / 0.1304	0.0238 / 0.0435	0.0364 / 0.0870	0.0943 / 0.2174
B18	0.0639 / 0.1429	0.1174 / 0.2143	0.0505 / 0.1429	0.0639 / 0.1429
B19	0.0381 / 0.0896	0.0309 / 0.0746	0.0279 / 0.0746	0.0267 / 0.0597
B20	0.1012 / 0.2632	0.0235 / 0.0526	0.0440 / 0.1053	0.0600 / 0.1579
B21	0.0822 / 0.2000	0.0447 / 0.1000	0.0652 / 0.1500	0.0418 / 0.1000
B22	0.0349 / 0.0833	0.0295 / 0.0833	0.0228 / 0.0417	0.0685 / 0.1667
B23	0.0349 / 0.0833	0.0373 / 0.0833	0.0228 / 0.0417	0.0349 / 0.0833
B24	0.0471 / 0.1333	0.0000 / 0.0000	0.0298 / 0.0667	0.0365 / 0.0667
B25	0.0345 / 0.0690	0.0308 / 0.0690	0.0393 / 0.1034	0.0154 / 0.0345
B26	0.1095 / 0.2000	0.0894 / 0.2000	0.1140 / 0.3000	0.0707 / 0.2000
B27	0.0161 / 0.0294	0.0383 / 0.0882	0.0161 / 0.0294	0.0294 / 0.0588
B28	0.0353 / 0.0968	0.0530 / 0.1290	0.0228 / 0.0645	0.0353 / 0.0968
B29	0.0000 / 0.0000	0.0000 / 0.0000	0.0000 / 0.0000	0.0373 / 0.0833
B31	0.0273 / 0.0658	0.0379 / 0.0921	0.0341 / 0.0789	0.0300 / 0.0789
B32	0.0000 / 0.0000	0.0000 / 0.0000	0.0000 / 0.0000	0.0000 / 0.0000
B33	0.0548 / 0.1500	0.0418 / 0.1000	0.0570 / 0.1500	0.0671 / 0.1500
B34	0.0463 / 0.1034	0.0289 / 0.0690	0.0244 / 0.0690	0.0189 / 0.0345
B36	0.0000 / 0.0000	0.0321 / 0.0909	0.0249 / 0.0455	0.0380 / 0.0909
B37	0.0708 / 0.1765	0.0246 / 0.0588	0.0360 / 0.0882	0.0360 / 0.0882
B38	0.0183 / 0.0333	0.0506 / 0.1333	0.0548 / 0.1333	0.0494 / 0.1333
B39	0.0391 / 0.0714	0.0299 / 0.0714	0.0160 / 0.0357	0.0479 / 0.1071
B40	0.0789 / 0.1765	0.0671 / 0.1765	0.0263 / 0.0588	0.0720 / 0.1765
B41	0.0761 / 0.1818	0.0407 / 0.0909	0.0000 / 0.0000	0.0407 / 0.0909
B42	0.0542 / 0.1212	0.0271 / 0.0606	0.0429 / 0.1212	0.0332 / 0.0606
B44	0.0283 / 0.0800	0.0219 / 0.0400	0.0490 / 0.1200	0.0632 / 0.1600
B46	0.0447 / 0.1000	0.0000 / 0.0000	0.0000 / 0.0000	0.0548 / 0.1000
B47	0.0358 / 0.0800	0.0335 / 0.0800	0.0219 / 0.0400	0.0456 / 0.1200
B49	0.0380 / 0.0909	0.0203 / 0.0455	0.0518 / 0.1364	0.0498 / 0.1364
B51	0.0000 / 0.0000	0.0000 / 0.0000	0.0000 / 0.0000	0.0000 / 0.0000
B53	0.0593 / 0.1364	0.0455 / 0.0909	0.0593 / 0.1364	0.0689 / 0.1818
B54	0.0373 / 0.0833	0.0801 / 0.2083	0.1271 / 0.2917	0.0977 / 0.2083
B55	0.0798 / 0.1579	0.0440 / 0.1053	0.0288 / 0.0526	0.0288 / 0.0526
B56	0.0743 / 0.1935	0.0177 / 0.0323	0.0177 / 0.0323	0.0489 / 0.1290
B57	0.0422 / 0.1034	0.0189 / 0.0345	0.0244 / 0.0690	0.0690 / 0.1379
B58	0.0447 / 0.1000	0.0447 / 0.1000	0.0000 / 0.0000	0.0447 / 0.1000
B59	0.0476 / 0.0952	0.0583 / 0.1429	0.0398 / 0.0952	0.0852 / 0.1905
B60	0.0440 / 0.1053	0.0372 / 0.1053	0.0440 / 0.1053	0.0440 / 0.1053
B61	0.0745 / 0.1667	0.0589 / 0.1667	0.0456 / 0.0833	0.0373 / 0.0833
B62	0.0000 / 0.0000	0.0000 / 0.0000	0.0407 / 0.0909	0.0498 / 0.0909
Development track, 6 books by length rank; not pooled with the census				
B15	0.0272 / 0.0577	0.0161 / 0.0385	0.0349 / 0.0769	0.0417 / 0.0962
B35	0.0953 / 0.2609	0.0714 / 0.1739	0.0532 / 0.1304	0.0496 / 0.1304
B43	0.0140 / 0.0312	0.0191 / 0.0469	0.0178 / 0.0469	0.0204 / 0.0469
B48	0.0322 / 0.0882	0.0509 / 0.1176	0.0720 / 0.1765	0.0208 / 0.0588
B50	0.0400 / 0.0800	0.0219 / 0.0400	0.0335 / 0.0800	0.0593 / 0.1600
B52	0.0667 / 0.1667	0.0149 / 0.0333	0.0149 / 0.0333	0.0527 / 0.1333

arms is larger than seed noise. Table 10 is that decomposition, a fully crossed random-effects model over arm, book and seed with a single observation in each of the 1,080 census cells, solved by moments. On the accuracy scale the between-book standard deviation is 0.1646, and the residual, which here also carries the three-way interaction, is 0.0536. The arm part is 0.0010, and the seed part is estimated below zero. The registered question compares those last two, which the adjudicated artefact does as their ratio, and a ratio whose denominator is not positive is undefined as a variance ratio, so we print the components and no ratio. The design also bounds how well either could have been pinned down: with these degrees of freedom, a component read off a single mean square carries a relative standard error of at least about 0.82 for the arm part and 0.71 for the seed part, and each of ours combines four of them.

Table 10: The registered variance decomposition of accuracy over the census, primary backbone. A fully crossed random-effects model over arm, book and seed with a single observation in each cell, so the three-way interaction cannot be separated from error and the last row carries both. Estimates are by moments, and a moment estimate can come out negative; the truncated column replaces a negative estimate by zero, and neither column is promoted over the other. The last column is the square root of the truncated estimate, on the accuracy scale. Degrees of freedom and sums of squares are checked to add up before the table is written.

Component	d.f.	Mean square	Estimate	Truncated	s.d.
Arm	3	0.004496	0.000001	0.000001	0.0010
Book	53	0.547409	0.027095	0.027095	0.1646
Seed	4	0.001339	-0.000003	0.000000	0.0000
Arm × book	159	0.005356	0.000496	0.000496	0.0223
Arm × seed	12	0.001730	-0.000021	0.000000	0.0000
Book × seed	212	0.003037	0.000040	0.000040	0.0063
Residual (arm × book × seed, and error)	636	0.002876	0.002876	0.002876	0.0536

The random-intercept model beside it. The registration's second path fits accuracy with a random intercept for book and one for seed, the arm effect withdrawn from it by the registration itself. Table 11 gives it at the reading the formula states, at the arm-averaged reading, and restarted from below and from above its own solution: 4 fits in all. Neither reading reports a clean fit: the stated reading does not report convergence, and the arm-averaged one reports it while emitting the same numerical warnings, which is why the restarts are here at all. Its book standard deviation at the stated reading is 0.1645, which the decomposition puts at 0.1646, and its seed standard deviation stays at or below 0.0012 across every fit. The largest disagreement among the restarts is 0.0049 on the variance scale, which is 0.0155 on the standard-deviation scale the rest of this paragraph uses, and it falls on the restart from below. Both paths yield a large estimated between-book component and a small estimated corpus-level seed component; given the zero-boundary estimates, the convergence reports and the numerical warnings above, they do not establish that seed effects are absent.

Table 11: The registration's second path: a random intercept for book and one for seed, with the arm effect withdrawn by the registration. Each row is a separate fit. The first is the reading the formula states, one row per arm, book and seed; the second averages over arms before fitting, which changes what the residual means and is therefore a sensitivity and not a replacement. The last two restart the first fit from below and from above its own solution. Book, seed and residual are standard deviations on the accuracy scale. Values are the fitting routine's output as returned and are not checked against the maximum of the likelihood it fits.

Fit	Rows	Book s.d.	Seed s.d.	Residual s.d.
Cell level, one row per arm, book and seed	1,080	0.1645	0.0001	0.0573
Arm-averaged, one row per book and seed	270	0.1679	0.0006	0.0273
Cell level, restarted at $0.25\times$ the solution	1,080	0.1490	0.0012	0.0576
Cell level, restarted at $4.0\times$ the solution	1,080	0.1648	0.0002	0.0573

The path we do not report. The registration also names a bootstrap over seed means and asks that it be printed beside the other paths. We do not report it, a departure from that request, because the registration's own text does not fix what it is a bootstrap of. The line that names the path

asks for a value; the name the path is filed under is that of a paired difference between arms, and a seed-averaged difference between arms is what another section of the same document declares unregistered, which is also the ground on which that document withdraws the arm effect from the model path a few lines earlier. Reading the path either way is a choice, and we do not make it after seeing the data. The registration fixes the descriptive path as the headline in advance, and the per-book spreads and the decomposition above are that path.

An exploratory interval beside it. The question behind that path is still a fair one, and the released shards fix its plain answer, so we print it and label it: average each census book's accuracy over the 5 construction seeds, take the paired difference between arms over the 54 books, and put on it the book-level t interval the primary analysis uses (53 degrees of freedom). The DP-mixture arm minus Leiden is +0.0088 [-0.0068, +0.0244], minus GMM+BIC +0.0074 [-0.0014, +0.0163], and minus the UMAP-matched Leiden arm +0.0029 [-0.0115, +0.0173]. Zero lies inside three of the three intervals, and the largest upper limit, +0.0244, lies below the registered target $\delta^* = 0.034$. This interval is exploratory, and we treat it as unregistered because the registration's words point both ways; it is conditional on these 5 seeds, it spends no α, and it neither replaces nor sharpens the registered verdicts above; it is printed because withholding a descriptive interval the released shards determine would be selective reporting.

D PRUNING THE DP ARM'S SUMMARY LAYER

A post-run diagnostic (registered as a post-run measurement; run over all 60 books) prunes the DP arm's summary nodes at five pre-specified levels (down to the Leiden arm's node count, its token sum, and 25/50/75% retention) and re-runs retrieval on the pruned index. Table 12 reports all five points together, as registered. The read-out is mechanical, per the metric's construction (§4): verbatim gold-evidence recall is scored on the assembled reader context, leaf and summary text alike, and removing summary nodes frees retrieval slots for leaves; in this pruning sweep no decrease was observed on any book at any level, while the summary share drops. Monotonicity is not asserted in general: it would require nested retrieved leaf sets and summaries that contribute no unique verbatim hit, conditions this sweep checks only through the observation itself. The read-out bounds verbatim evidence delivery; it is not evidence that pruning improves answer accuracy, and no accuracy claim is licensed by it. Search latency is flat across the sweep; those timings were taken while the backbone robustness check run shared the host, so they are read as indicative only.

Table 12: Pruned-index diagnostic on the DP arm, book mean over 60 books (each book weighted equally): summary share of retrieved slots, conditional verbatim gold-evidence recall, and the book mean of per-book median (p50) search latency. All five registered pruning points are reported together.

Index	Summary share	Verbatim recall	Search p50 (ms)
Unpruned DP arm	0.2499	0.0770	13.4
Pruned to Leiden node count	0.0812	0.0845	13.5
Pruned to Leiden token sum	0.0809	0.0845	13.4
Retain 25% of summaries	0.1007	0.0842	13.5
Retain 50%	0.1655	0.0821	13.5
Retain 75%	0.2130	0.0799	13.5

E CHAPTER-BOUNDARY CONTROL

This control is chapter-boundary-constrained contiguous grouping at layer 0: within a chapter the leaves are cut into the fewest contiguous blocks no larger than the contiguous control's block size, with sizes differing by at most one (the contiguous control cuts fixed-size blocks), every layer above the first is identical to the contiguous control, and layer-0 groups cannot cross an author-marked chapter boundary. It is a constrained baseline, not a replacement of the induced hierarchy by document structure.[1] Its effective range is a registered mandatory disclosure: chapter markers were

[1] Quantitatively, as the registration asks this claim to be footnoted: over the 38 refoldable chapter-arm trees behind the registered anchor (36 NovelQA books and 2 QuALITY articles), the induced layer-1 group count is

usable on 37 of 60 books (median 39 boundaries per usable book, range 5–172); the leaf-position gate then passed 36 of them and rejected 1, so the arm acts on 36 books and falls back on 24 (40.0%). On 24 of 60 books (40.0%) the arm has no chapter structure that survives both gates and falls back to the plain contiguous construction by design. On those books the two arms are not merely similar but identical: comparing the built trees node by node, their groupings agree at every layer, so the whole-corpus row of Table 13 compares chapter markup against contiguity on 36 books and reproduces the contiguous control exactly on the other 24. The whole-corpus gap between those two columns is therefore attenuated by construction and must not be read as a 60-book effect. The subset row of Table 13 restricts to the 36 chapter-parsed books, with every comparison arm restricted to the same books. The chapter-parsed subset is not a random sample (it is the books with clean markup, which may correlate with genre, era, and typography), so the subset row is read only against its matched row, never alone.

Three detector choices, none of them registered. The detector carries a coverage threshold, a rule variant that was rejected, and a heading-to-leaf mapping strategy adopted on an aggregate after the fact. Sweeping the coverage threshold across 5 values moves the usable count monotonically: 0.40, 0.50, 0.60, 0.70, and 0.80 give 40, 38, 37, 35, and 34 usable books, and the value in force is 0.60. A more permissive variant of the first-occurrence rule reaches 43 usable books; it was rejected because the line-shuffled negative control still left 10 of 60 books usable under it, against 0 under the rule in force. Of the 3 heading-to-leaf mapping strategies computed, which recover 1722, 1698, and 1690 of 1724 headings, the longest-common-subsequence strategy was adopted: the largest of the three, chosen on that aggregate rather than in advance. The arm is registered as descriptive and enters no registered comparison, so none of these choices moves one, though the registered moderator hypothesis below is computed on the books and chapter boundaries they fix; they are stated because a threshold that was selected and a strategy adopted on an aggregate are degrees of freedom a reader cannot otherwise see.

Table 13: Chapter-boundary control with matched comparisons (registered reporting unit: the full-corpus value, the subset value, and the same-book comparison values appear together; the subset value is never cited without its matched comparisons). Unweighted book-mean accuracy. The remaining 24 books (40.0%) have no chapter structure that survives both gates and fall back to the plain contiguous construction (Appendix E), so this is a partial-coverage control rather than a chapter-boundary arm over the whole corpus.

Scope	Chapter	Flat dense	GMM+BIC	Leiden	DP mixture	Contiguous
All 60 books	0.5837	0.5710	0.5606	0.5674	0.5770	0.5775
Chapter-parsed 36 books	0.6136	0.5978	0.5872	0.5943	0.6017	0.6031

A registered moderator hypothesis attaches to this appendix's structure axis: that books whose chapter token-mass is more concentrated show a larger DP−Leiden per-book delta, with a pre-fixed metric order (primary: chapter-mass Gini, registered sign positive; secondary: Hill tail index, registered sign negative, defined on the subset with estimator support). Neither interval separates from zero: the Gini–delta Spearman correlation is -0.2106 (book-bootstrap 95% CI $[-0.5466, +0.1560]$, $n = 36$) and the Hill–delta correlation is -0.0373 ($[-0.5113, +0.4450]$, $n = 20$). Under the registration's own wording rule these intervals license no directional statement in either direction; we report them as registered and move on.

F LENGTH-STRATIFIED DELTAS, FULL TABLE

Table 14 prints every cell of the registered length-stratified read-out: three per-book accuracy-difference terciles by book length (boundaries recorded on 2026-08-20, after the registration freeze of 2026-08-16, which the registered wording names as the recording time, and before the answer key was unsealed; this departure is disclosed here; each tercile cell holds 20 books) plus a Fisher-interval correlation between log book length and the per-book delta. The three comparisons against

2.20× the chapter-segment count at the median (range 1.00–11.89; 2.22 on the NovelQA books alone), so the chapter constraint leaves the induced partition at least as fine as the chapter structure and strictly finer on 34 of the 38.

the flat floor carry a registered direction, and 0 of 12 cells establish it; the DP−Leiden axis carries no registered direction and is printed for completeness.

Table 14: Length-stratified per-book deltas (cluster-robust 95% CIs within each tercile; corr rows give Pearson r over all 60 books with a Fisher interval). All cells printed; none is a gate.

Comparison	Cell	Estimate	95% CI
Leiden − flat	T1 short	$+0.0012$	$[-0.0217, +0.0240]$
	T2 mid	-0.0122	$[-0.0414, +0.0169]$
	T3 long	$+0.0004$	$[-0.0252, +0.0259]$
	corr	-0.0623	$[-0.3113, +0.1947]$
Leiden (UMAP-m.) − flat	T1 short	$+0.0091$	$[-0.0247, +0.0428]$
	T2 mid	-0.0081	$[-0.0435, +0.0273]$
	T3 long	-0.0032	$[-0.0444, +0.0381]$
	corr	-0.1264	$[-0.3685, +0.1318]$
DP − flat	T1 short	$+0.0202$	$[-0.0115, +0.0518]$
	T2 mid	-0.0045	$[-0.0548, +0.0459]$
	T3 long	$+0.0023$	$[-0.0490, +0.0535]$
	corr	-0.1516	$[-0.3905, +0.1064]$
DP − Leiden (no registered direction)	T1 short	$+0.0190$	$[-0.0064, +0.0444]$
	T2 mid	$+0.0078$	$[-0.0295, +0.0450]$
	T3 long	$+0.0019$	$[-0.0474, +0.0511]$
	corr	-0.1365	$[-0.3773, +0.1217]$

G QA BUDGET CURVE

The registered budget curve revisits the cached indexes of five arms at four reader-token budgets. Its reporting unit is the full grid: the registration forbids citing a subset of points or arms, and the cells are not comparable on budget alone because the retrieval depth changes at the largest point (top-40 at 7,400 tokens, against the top-20 of §3.4 elsewhere). Table 15 prints every cell; the 2,000-token column is the main run itself (Tables 1 and 21).[2]

Table 15: QA budget curve, unweighted book-mean accuracy: all cells (five arms × four budgets). The 2,000-token column is the main run; the 7,400-token column retrieves top-40 instead of top-20.

Arm	@1,000	@2,000 (main)	@4,000	@7,400 (top-40)
Flat dense	0.5683	0.5710	0.5710	0.5799
Leiden	0.5546	0.5674	0.5687	0.5625
Leiden, UMAP-matched	0.5637	0.5703	0.5743	0.5595
DP mixture	0.5587	0.5770	0.5704	0.5708
Fixed-block contiguous	0.5699	0.5775	0.5813	0.5710

Three disclosures accompany the grid. *First*, the flat arm's 2,000- and 4,000-token points are the same output: all 1,526 predictions are identical to the main run's, because top-20 retrieval over ∼100-token chunks fills roughly 2,000 tokens before a larger budget can bind (the main-run rows of Table 16: every flat-arm hit fits on 100.0% of questions). The flat arm's apparent flatness between those points therefore measures a retrieval ceiling, not reader behavior. *Second*, Table 16 reports, for every off-main cell, how many prompts actually moved in the budget's direction (bound) and how many of those changed the prediction: the tree arms rebind substantially at 4,000 tokens where the flat arm does not, and the change rate among bound questions is not constant across budgets. The same table carries the registered ceiling-reach share per cell, the main run included: at the smallest curve point the reader budget binds almost everywhere (every retrieved hit fits on at most about

[2] At the 7,400-token budget the first and last letter of the completion disagree on a small share of DP-arm records (8 of 1,526, $\approx 0.52\%$), so the registered last-letter rule and a first-letter rule would select different option letters. A disagreement is not by itself a misread: Appendix Z lists every one with its shape, and some of their completions name different letters in different places, where no mechanical rule picks one of the named letters over the others.

0.7% of questions in any arm), whereas at the points above the main run's budget the hits all fit on at least about 92.9% and 91.0% of questions, so the per-point top-k is what binds there. Figure 6 plots the accuracy grid against this table's ceiling column on one horizontal axis. *Third*, the 7,400-token flat-arm point rebinds every question only because the retrieval depth doubles there; its maximum prompt is 3,947 tokens, still under the nominal budget, so from the main run's budget upward the flat arm's effective budget axis is set by the retrieval ceiling (at 1,000 tokens the reader budget itself binds).

Two readings of Figure 6 belong in the text rather than in its caption. There are no error bars: per-arm accuracy intervals are not registered at any point, and the registered paired contrasts (Figure 2) are main-run contrasts, so no column of panel (a) carries a registered interval and the separation or overlap of points settles no statistical comparison. Panel (b) is the registered ceiling-reach share, the fraction of questions whose retrieved hits all fit under the budget, so that the point's top-k rather than the token budget decides what the reader saw. It answers what the accuracy panel alone cannot: above the main run's budget the token budget no longer binds on at least about 92.9% of the questions in every arm the sweep covers, and on at least about 91.0% at the largest point, so whatever the accuracy points do there is not a reading about how much context a reader can use.

Table 16: Prompt binding per off-main curve cell, recomputed from the per-query records (the changed column from the released answer-change flag) and anchored to the run ledger. "Ceiling" is the registered ceiling-reach share: the fraction of questions whose retrieved hits all fit under the budget, so the point's top-k rather than the token budget binds (primary backbone only); it is measured on the context tokens the budget governs, whereas median and max are whole-prompt tokens, so a cell can reach the ceiling while its longest prompt exceeds the nominal budget. The main-run rows carry only the ceiling share, since "bound" counts prompts that moved in the budget's direction relative to the main run and "changed" counts bound prompts whose prediction changed. The bound column gives a count with, in parentheses, its percentage of that cell's questions; the changed column's percentage is of that cell's bound prompts, not of its questions. The GMM+BIC arm was run at the main budget only, so it contributes the ceiling column at that point and no sweep row; its ceiling share is the share of its questions on which the registered retrieval depth, not the token budget, is what determined the context the reader saw.

Arm	Point	Ceiling (%)	Bound	Changed	Median tok	Max tok
Flat dense	@1,000	0.7	1,515 (99.3)	300 (19.8)	1,068	1,243
	@2,000 (main)	100.0	—	—	—	—
	@4,000	100.0	0 (0.0)	—	1,819	2,081
	@7,400	100.0	1,526 (100.0)	386 (25.3)	3,534	3,947
Leiden	@1,000	0.4	1,520 (99.6)	334 (22.0)	1,060	1,243
	@2,000 (main)	59.2	—	—	—	—
	@4,000	99.3	623 (40.8)	129 (20.7)	2,016	4,129
	@7,400	99.5	1,526 (100.0)	471 (30.9)	3,960	7,544
Leiden, UMAP-m.	@1,000	0.3	1,521 (99.7)	348 (22.9)	1,057	1,245
	@2,000 (main)	57.3	—	—	—	—
	@4,000	99.0	651 (42.7)	131 (20.1)	2,038	4,157
	@7,400	99.0	1,526 (100.0)	473 (31.0)	3,984	7,534
DP mixture	@1,000	0.3	1,521 (99.7)	341 (22.4)	1,048	1,252
	@2,000 (main)	27.3	—	—	—	—
	@4,000	92.9	1,110 (72.7)	267 (24.1)	2,596	4,267
	@7,400	91.0	1,526 (100.0)	567 (37.2)	5,034	7,637
Fixed-block contiguous	@1,000	0.5	1,519 (99.5)	312 (20.5)	1,056	1,259
	@2,000 (main)	50.9	—	—	—	—
	@4,000	97.4	750 (49.1)	144 (19.2)	2,110	4,212
	@7,400	97.2	1,526 (100.0)	472 (30.9)	4,124	7,634
GMM+BIC	@2,000 (main)	30.5	—	—	—	—

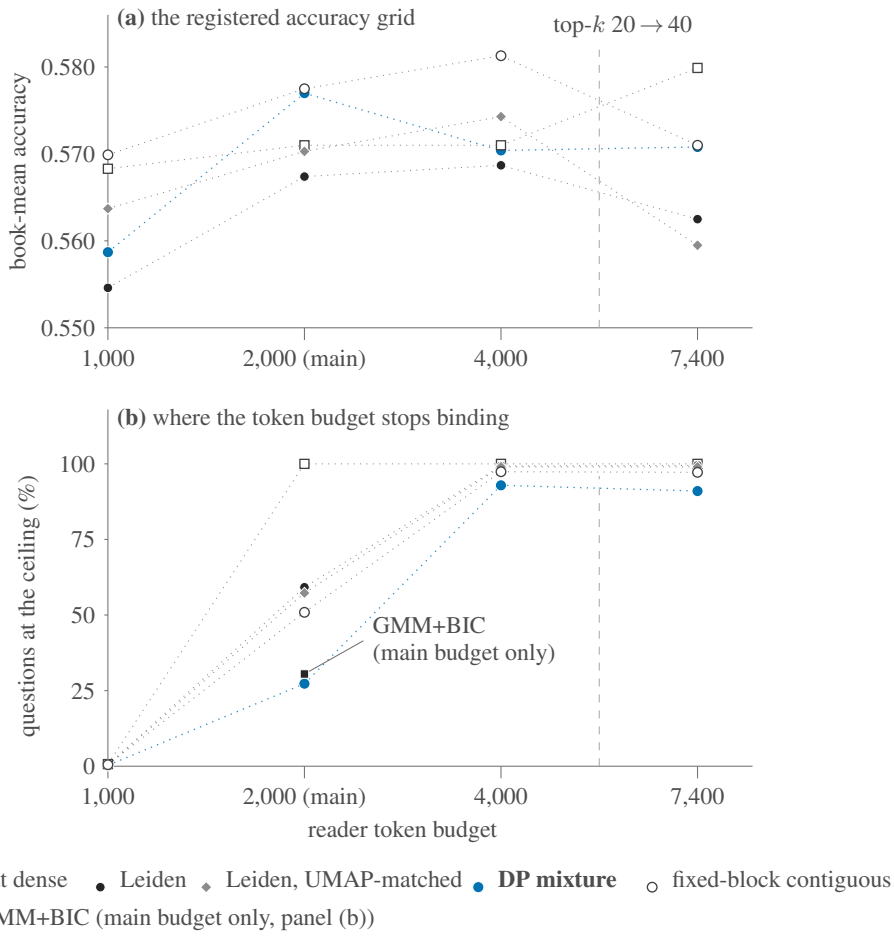

Figure 6: The registered budget grid of Table 15 drawn above the registered ceiling-reach share of Table 16, so that the upper panel can be read against the lower one. *Axes.* The horizontal axis is categorical: the registered budget points are equally spaced and are *not* to scale in tokens, so no slope here is a rate. Panel (a)'s vertical axis does not start at zero and covers a narrow window whose scale the printed ticks give, so a visible gap is a small difference drawn large. *Reading rules.* The dotted connectors join each arm's measured points and are reading aids; nothing is interpolated, smoothed, or fitted between them. The dashed rule marks the one place where two things change at once, since the largest point also raises retrieval depth from top-20 to top-40, so its cells are not comparable to the others on budget alone. Marks that fall within a marker width of each other overlap, and at the smallest budget point of panel (b) every arm does; the DP arm is drawn first everywhere so that its mark cannot cover another arm's, and the exact values are in the two tables. The GMM+BIC arm ran at the main budget only, so it contributes a ceiling cell and no curve and is drawn at that point alone, with a mark of its own. Every value drawn or named here comes from Table 15 or Table 16, from a cell or from that table's caption; the figure adds no quantity of its own.

H BUDGET–RECALL CURVE

The registered budget–recall curve rescores conditional gold-evidence recall (§4) after cutting each query's cached retrieval list at four reader budgets, a reconstruction of the packer, with no re-retrieval and no model calls. The instrument carries an honest bound, reported per cell: retrieval depth stays at the registered top-20, so *sat* is the share of queries whose full retrieval list was packed into the budget with nothing left out (the event the ceiling-reach share of Table 16 counts), and a saturated cell marks where this instrument stops measuring, not evidence that a larger budget cannot help. By construction the 2,000-token column reproduces the main run's evidence-recall read-out. Table 17 prints every primary-backbone cell of the six arms it scores; the fixed-block contiguous arm is not among them. The registration frames the curve as an artifact check: a recall gap at the registered budget that faded with more budget would point at the budget rather than at the clustering. Here the DP arm's recall deficit against GMM+BIC narrows from 0.0618 vs. 0.0661 at the registered budget to 0.0672 vs. 0.0676 at double the budget. The saturation column says how much of each cell the instrument still measures. It is a corrected read-out: an earlier rule counted a context as saturated when it was shorter than the budget, which the greedy packer makes almost always true, so the column now counts the packing event defined above; the recall cells are unchanged, and the superseded transcript is released beside the corrected one. At the two lower budgets almost no query packs its full list (saturation at most 1%), so those cells are pure truncations. At the registered budget the flat arm is fully packed but the tree arms are not: 25–59% of queries pack the whole list, with the DP arm near the low end (a share of 0.27), so for 41–75% of queries, depending on the arm, the budget still binds there. Only at 4,000 tokens is the instrument nearly exhausted for the tree arms (92–99% of queries pack their full list), and any movement beyond that budget would be carried by the unsaturated remainder alone: about 8% of queries in the least saturated 4,000-token cell (saturation is reported at two decimals and tallied over all $n_{\text{sat}} = 1{,}526$ questions; carrying that rate to the 1,519 evidence-eligible questions that enter the recall read-out gives about 122 questions, but only if the unsaturated share is the same in both populations, which we do not measure), with no interval attached. The summary of this appendix is therefore two-sided: at the registered budget the tree arms are budget-truncated for 41–75% of queries and the DP–GMM recall gap narrows as the budget doubles; at 4,000 tokens the registered top-20 depth, not the token budget, is what limits the instrument.

Table 17: Budget–recall curve: conditional gold-evidence recall, with the saturation share in parentheses (share of queries with at least one retrieved hit whose full retrieval list was packed into the budget; computed over all $n_{\text{sat}} = 1{,}526$ questions with a reconstructed context and non-empty retrieval, a different denominator from recall's). All cells (primary backbone): six arms $\times$ four budgets, recall over $n_{\text{eff}} = 1{,}519$ evidence-eligible questions. The 2,000-token column is the main run's evidence-recall read-out.

Arm	@500	@1,000	@2,000 (main)	@4,000
Flat dense	0.0517 (0.00)	0.0637 (0.01)	0.0731 (1.00)	0.0731 (1.00)
GMM+BIC	0.0424 (0.00)	0.0521 (0.00)	0.0661 (0.30)	0.0676 (0.94)
Leiden	0.0482 (0.00)	0.0595 (0.00)	0.0700 (0.59)	0.0711 (0.99)
Leiden, UMAP-matched	0.0470 (0.00)	0.0591 (0.00)	0.0692 (0.57)	0.0703 (0.99)
DP mixture	0.0424 (0.00)	0.0529 (0.00)	0.0618 (0.27)	0.0672 (0.93)
DP + confidence reranking	0.0416 (0.00)	0.0517 (0.00)	0.0610 (0.25)	0.0668 (0.92)

I COMPLEXITY SUBSETS

NovelQA tags each question with one of three complexity labels: single-hop (`sh`), multi-hop (`mh`), and detail (`dtl`) (Wang et al., 2025). The registration slices two estimands along this axis, both scored on the primary run and both descriptive. First, the primary contrast $\Delta(\text{DP} - \text{Leiden})$ per subset, registered as the pre-specified counterpart of a development-phase observation, precisely so that a subset signal could not resurface as an exploratory finding after a null primary result. Second, each of the three tree arms this second estimand covers (Leiden, UMAP-matched Leiden, DP mixture) against the flat baseline, $\Delta(\text{tree} - \text{flat})$ per subset, with a registered direction prediction: on the displacement account the tree's gain should be largest on multi-hop questions and smallest,

possibly negative, on detail questions. The registration binds the read-out in advance: without CI separation, no "as predicted" claim may be made.

That rule decides here. All three two-sided intervals for the primary contrast include zero (Table 18), and all nine one-sided lower bounds for the tree-versus-flat contrasts are negative (Table 19): no subset separates any tree arm from the flat baseline. At the level of point estimates only, the ordering is partially consistent with the prediction (the multi-hop delta is the largest of the three subsets for every tree arm, and the Leiden-UMAP-matched arm's detail delta is the one negative point estimate), but the DP arm breaks the "smallest on detail" half ($+0.0168$ on `dtl` vs. $+0.0154$ on `sh`). These orderings are noted; nothing more is claimed. No cell tests the ordering itself: the registered read-out asks, per subset s, whether $\Delta_s > 0$, whereas the prediction is a contrast *between* subsets, $\Delta_{\mathrm{mh}} - \Delta_{\mathrm{dtl}} > 0$. Neither implies the other: negative lower bounds in every subset leave the ordering untested, and positive deltas everywhere would not establish it either. The registration operationalized the prediction this way and we hold to it, so the ordering is described here and adjudicated nowhere. The symmetric caveat also holds: with the sample split three ways, these intervals are 1.5–$2.8\times$ the registered design target of the primary contrast in width (the one-sided tree cells are unbounded above), so the subsets are uninformative in both directions: they license neither the registered prediction nor a claim that no complexity-dependent gain exists. The registration made this read-out descriptive and alpha-free. The joint count across both registered directional families is stated in §4.

Table 18: Registered secondary read-out (descriptive, alpha-free). Primary contrast (DP − Leiden) by complexity subset: book-mean accuracy difference within each subset (per-book accuracy over the subset's questions, paired over books) with cluster-robust two-sided 95% CIs; n_{eff} is the number of books contributing questions of that type.

Subset	n_{eff}	Δ	95% CI
Single-hop	53	$+0.0024$	$[-0.0233, +0.0281]$
Multi-hop	54	$+0.0267$	$[-0.0212, +0.0746]$
Detail	57	$+0.0075$	$[-0.0399, +0.0548]$

Table 19: Registered secondary read-out (descriptive, alpha-free). Tree arm − flat dense baseline by complexity subset: book-mean accuracy difference within each subset with the one-sided lower 95% bound under the registered direction prediction (tree > flat). Every bound is negative; n_{eff} as in Table 18. Each cell weights every book carrying the subset equally, however few of its questions the subset holds, so the cells need not bracket the whole-corpus tree-minus-flat difference of §4.

Arm	Subset	Δ	Lower 95%
Leiden	Single-hop	$+0.0130$	-0.0245
Leiden	Multi-hop	$+0.0148$	-0.0218
Leiden	Detail	$+0.0093$	-0.0145
Leiden, UMAP-matched	Single-hop	$+0.0117$	-0.0259
Leiden, UMAP-matched	Multi-hop	$+0.0232$	-0.0146
Leiden, UMAP-matched	Detail	-0.0034	-0.0429
DP mixture	Single-hop	$+0.0154$	-0.0253
DP mixture	Multi-hop	$+0.0415$	-0.0136
DP mixture	Detail	$+0.0168$	-0.0175

J REGISTRATION INTEGRITY: THE RE-FREEZE EDIT AND ITS REHEARSAL

The registration in outline. During review the registration ships as an English rendering and an anonymized copy of the original-language file. The point estimate of the registered comparison is a census quantity for this fixed snapshot, conditional on the recorded outputs, while its interval treats the 60 books as exchangeable draws from a population of comparable books, a working model for inference beyond them that carries no variation from rebuilding the indexes or resampling model outputs (Appendix S); the conjunction replaced a single Leiden comparison at the re-freeze and, as an intersection–union test, needs no multiplicity correction (Berger, 1982). By registration the

H1 gate uses two-sided 95% intervals (directional error rate 0.025) and the pre-registered directional contrasts use one-sided 95% bounds (directional 0.05); the two are never pooled into one α, and the registration fixes some directional contrasts two-sided instead, among them the reranking comparison and the length strata. Registered power has two halves: at the registered design target $\delta^* = 3.4$ pp, the conjunction reaches 0.7201 power (below the 0.80 single-comparison target) under the symmetric common-delta scenario and 0.8085 under the registered ordering scenario in which the Leiden gate binds (both two-sided). Testing is fixed-sequence: the reranking comparison (§6) receives alpha only if H1 is supported, and the realized s that the registration thresholds is the between-book spread of the DP-minus-Leiden deltas of §4 alone. The registration also fixes a validity gate, per-arm floors, matched contrasts, two directional families (length strata, two-sided in every cell; complexity subsets, one-sided for trees against flat and two-sided for DP against Leiden), and a book-hierarchical secondary read-out, each adjudicated in §4, plus a descriptive option-position probe that spends no error budget (Appendix T).

The frozen protocol went through two generations. The initial freeze (1,572 lines, SHA-256 `5300b7efc2301123756cfe056ae7d9da3b93b564b8f695b620f87547790aa517`; not in the release) predates every confirmatory artifact but one tree, which the registered pre-freeze feasibility check built and the main run reused under the frozen configuration. One continuous unfreeze–edit–refreeze procedure (with the edit plan fixed before the window opened) produced the final protocol (2,198 lines, SHA-256 `8e40dfc4c5e57aa3d079d14e71587797c56b159350e70ee3577469b766d9211c`), the version every result of the first registration is scored against. The answer key remained sealed throughout: the edit window closed before unsealing, and no edit had access to any confirmatory score. Neither hash above is anchored by a third-party timestamp: the order of freeze, sealing, and unsealing rests on our version-control history and on the run ledgers, which stamp the protocol's hash on every scored query and which the confirmatory driver refuses to produce unless the protocol's status is frozen and the hyperparameter file matches the hash the protocol registers, all of them internal records; during double-blind review the protocol is available as an English rendering and as an anonymized copy of the original-language file, in which the token classes named in a table beside it (an advisor-role word, the registration's own venue history, a prior project name, a third-party form address, personal-funds amounts in the local currency, the same funds' US-dollar figures, and a time-zone offset, which occurs only in the second registration's deviation log) are replaced by uniform markers of their class, that table giving the count per marker; the English rendering carries every marker the protocol uses except the amount ones, stating the local-currency amounts in unnamed units and withholding the dollar figures, and the amount classes are declared, not marked, in a note beside it; the original-language file with its full commit history is released with the camera-ready version. Three generators read that rendering rather than the original, so that what they print re-derives from what a reviewer holds: the registered design target against which Appendix I scales the complexity-subset intervals, the power arithmetic of Appendix M (whose generator also opens files the release does not carry; Appendix S), and the paper's own claim register (not the registration) of Appendix Q. Two of them take values out of it by pattern; the third also hashes it, so what that one pins is the rendering's bytes, not the frozen original's. One of the first two opens the original as well when it is present and stops if the two disagree on the fields it uses; on the release the original is absent (the anonymized copy is a different file, which that generator does not open) and that comparison does not run. The quoted claim sentence keeps a pin of its own, and that hash is the same value in both documents because the sentence was already in English in the original; the prefix printed beside it is the original's hash as the rendering declares it. Declaring is not verifying: the check that reads both documents and compares them field by field runs on our side before release rather than on the release, and it is not one of the generator records above, so a reviewer cannot re-run it before the camera-ready version. Besides the cases recorded later in this appendix, two statements of the frozen first registration are inaccurate as written and stay unedited, because its bytes are pinned: it calls the adjacent-merging ablation of Ou & Lapata (2026) non-semantic, whereas that ablation decides each merge by semantic compatibility (Appendix A), and its two bounds on how far the conjunction threshold moves when its estimated inputs, a correlation and the projected spread ratio, are perturbed hold only after rounding. The registration README shipped with the release records both statements and lists the errata of its English rendering.

One further departure concerns accounting, not analysis. The confirmatory run closed within the frozen spend cap; the transfer check and its control, run after it, were charged to a later-raised project

budget, with the primary backbone's transfer runs on the confirmatory run's ledger, still within that cap, and the robustness-check backbone's transfer runs on separate ledgers of their own, and any spend figure we release is computed from the ledgers that bill it, all of which ship with the release manifest. The first registration had budgeted the transfer check inside that cap and makes raising the cap a new registration; the later budget was instead set by an operational decision of ours, taken when the projected spend of the run and the check together passed the cap. Some transfer cells hold records written under several code revisions, and in the first registration's transfer check the LongBench v2 arms of both backbones, together with the closed-book controls of the robustness-check backbone, were written from an uncommitted working tree, with a single exception among the primary-backbone LongBench v2 arm records; every record carries its own revision string, and the analysis that produced the appendix was run from a single committed revision. The primary and robustness-check cells are not exempt from the first of these: four of their arm ledgers carry more than one revision string, two of them for an arm the first registration's fixed-sequence comparisons read (GMM+BIC and confidence reranking; the reranker comparison is descriptive in this outcome, §6). That set is not the set of four arms whose builds failed and were healed by a resumed run, reported with the zero exclusion rate in Appendix L: the two sets are the same size and share two members, with neither contained in the other, so a change of revision string is not the trace of a resumption and its absence is not evidence that none occurred.

What changed between the revisions a ledger straddles. Table 20 reads the repository's own history for every ledger whose records carry more than one revision string and asks one question of each transition: does the diff reach the code that produces an answer? We do not curate that code by hand. The answer path is the import closure, one level deep, of the three modules a scored query passes through (per-query generation, context assembly, and the index), so prompting and parsing, the model call, the queries and the option order the reader sees, retrieval, the record schema, scoring and the partition sampler enter it whether or not we would have thought to list them. The caption names the eleven modules that yields, and the gate fails if that list and the import graph disagree. Two things bound what this settles. The closure is a superset of what any one run executes: one of the eleven is reached only through a function-local import, on a code path that no registered run executes and that the post-hoc rebuild of Appendix V.2 does, and is audited anyway. And one level is a choice: the four modules those eleven import in turn, the response cache and the seed manager behind the model call among them, sit outside it. Two of the six transitions leave all eleven untouched. Two reach exactly one module each: one widens the record schema's set of dataset names, and the other re-derives a call's end instant from a monotonic anchor when the operating system steps the wall clock. Two cannot be audited this way at all, because each has a side that is not a commit but a commit plus an uncommitted delta. In the transfer row the other side is committed, and it is the single exception the paragraph above discloses; in the rebuild row the records span two working states and the commit that followed them. For both, what a diff still recovers is the span between the committed bases, and the table prints what that span reaches. Of the 13 contrasts the probe of Appendix W prints, three straddle the transition that reaches the generation module, and none of the three the main text prints is among them. The probe's row of the table and that count rest on its tree-fetch ledger and per-query transcripts, neither of which is released. The audit is a statement about diffs between named revisions, not a proof that two runs were behaviorally identical.

Among the edits, those touching the primary predicate replaced a single Leiden comparison with the conjunction of §3.4. The change followed a third-party seam-only ablation that found no grouping rule dominating across datasets and Leiden trailing GMM+BIC (Ou & Lapata, 2026); in the same ablation an order-constrained semantic grouping variant (N-RAPTOR, which merges only adjacent units but decides each merge by semantic compatibility) led on NovelQA but trailed on DetectiveQA and LiteraryQA; our fixed-contiguous control shares only the order constraint, so its NovelQA result is a related observation, not a reproduction. Beating Leiden alone therefore risked being the easier claim; swapping comparators instead would have turned the halted QuALITY confirmatory comparison recorded in the registration (Appendix J.1) from a loss into a win (registered unweighted accuracies 0.6049 for GMM+BIC, 0.6176 for Leiden, and 0.6113 for the DP arm: a win over GMM+BIC and a loss to Leiden, so the conjunction fails on the data already seen); requiring both is strictly harder than either. That is what the two-scenario power disclosure stated earlier in this appendix quantifies, and the predicate change was fixed before any confirmatory outcome of this study existed. The registration also records that the comparator swap was proposed first and rejected after the QuALITY accuracies above were presented, and it records proposing the swap

Table 20: Every ledger whose records carry more than one revision string, grouped by the transition they straddle, and what that transition changes. The answer path is not a hand-made list: it is the one-level import closure of `harness/novelqa_answer.py`, `baselines/raptor.py` and `harness/raptor_tree.py`, taken at every revision this table names and at the revision this paper is built from, which yields those three together with `harness/prompts.py`, `harness/llm.py`, `harness/retrieval.py`, `harness/dataset_loader.py`, `harness/option_shuffle.py`, `harness/instrumentation.py`, `harness/metrics.py` and `harness/gibbs_partition.py`. A gate on our side re-derives all of it (the closure, the transitions, the ledgers in each row, the modules each diff reaches, and the counts printed above) from the code and the repository history on every run.

Transition	Ledgers that straddle it	Answer path, and what does change
`2131487b` → `05dfb0a7`	primary backbone: the GMM+BIC arm and the confidence reranking arm	none of the eleven. The transition changes five files: comments correcting a planning constant, a budget-recall-curve analysis block with its two test files, and the draft `harness/instrumentation.py`, and only it, in two added lines and no removed line: `detectiveqa` and `longbench_v2` join the record schema's set of permitted dataset names. The transition changes thirty-nine files, the rest of them transfer-check and analysis scripts with their tests
`201a6df9` → `535e7310`	robustness-check backbone: the cluster-count-matched GMM+BIC arm and the cluster-count-matched contiguous grouping arm	none of the eleven. The transition changes four files: an option-order probe added behind switches that default to off, in the confirmatory driver and in a new module (`harness/mcq_shuffle_paths.py`) that nothing on the answer path imports, with an audit script and a test file
`535e7310` → `aeaf54e6`	DetectiveQA transfer, primary backbone, the DP-mixture arm	`harness/novelqa_answer.py`, and only it, in seventy-six added lines and four removed lines: a monotonic anchor (`_end_instant`) that re-derives a call's end instant when the operating system steps the wall clock, and the field `clock_step_sec` that records the step; the per-query latency the efficiency table reads comes from that same instant. The four removed lines are the `datetime` import and the two `datetime.now` clock reads this replaces, and the constructor line the new field joins. The transition changes five files, the rest of them the probe driver's spend cap, the confirmatory driver's declaration of the new field, and two test files. The three contrasts the main text prints were all scored before it
`eb8ddd29` → `06935618`	stronger-reader probe: the tree-fetch ledger; the probe's condition set is split across this transition	*not auditable*: one side is a commit plus an uncommitted delta, so no diff reconstructs it. What is recoverable is the committed span between the two bases: it changes twenty-nine files and leaves all eleven untouched. The uncommitted delta is not in that span
`0a7c762e-dirty` & `535e7310`	LongBench v2 transfer, primary backbone, the Leiden arm	*not auditable*: two of the three sides are a commit plus an uncommitted delta, so no diff reconstructs them. What is recoverable is the committed span from the earliest base to the side that is a commit: it changes fifty-two files and reaches `harness/instrumentation.py`, and only it, in six added lines and no removed line, where the two slices of the second registration join the record schema's set of permitted dataset names. The uncommitted deltas are not in that span
`49d4151a-dirty` & `5697e43d-dirty` & `7831988a`	primary backbone, the seed-propagation rebuild: four seed shards, one for each added build seed, over the GMM+BIC, Leiden, UMAP-matched Leiden and DP-mixture arms, sixteen ledgers in all	

without presenting those accuracies as a procedural defect. The registration names the test family (an intersection–union test) under which each conjunct is tested at full α and no multiplicity correction applies (Berger, 1982). The predecessor version's multiplicity rule was different again: it labelled its rule Holm over the primary and the reranking comparison, the label was procedurally incoherent for the fixed levels it assigned, and the fixed-sequence structure outlined at the start of this appendix replaced it. The replacement is favorable to the primary hypothesis, it was made before any of this study's own NovelQA accuracy data existed, and the registration's change table records it with that direction stated.

One name in §3.4 is worth pinning for a reader holding the registration. The registered design target δ^* is the quantity the registration stores under the key `registered_mde_pp`, named there as a minimum detectable effect. It is not the conjunctive minimum detectable effect, which the registration records separately as the larger effect at which the conjunction would reach the single-comparison power target. The value itself is read off the registration's simulated power curve rather than solved in closed form: on the registered design at 60 books, simulated single-comparison power reaches 0.823 at δ^*, the smallest point printed on that curve clearing the 0.80 single-comparison target. The registration also records, by a different route, the effect at which that same single comparison attains the target exactly; it lies just below δ^*, so δ^* is the printed grid point above that effect rather than the effect itself, and the registration states no rule for preferring one to the other. Conjunctive power at the same δ^* is lower and scenario-dependent, and both of its registered sce-

narios are given earlier in this appendix. What the design is therefore unable to resolve, and what its abandonment verdict consequently does not speak to, is stated in Appendix S.

Because a hand-edited registration is a channel for silent drift, the edit itself was mechanized. All 29 edits (12 replacements, 17 insertions) were applied programmatically from a pre-rendered edit packet: most payloads parser-extracted from the packet; for the rest, the text each edit adds was transcribed into the harness by hand and is in the packet verbatim apart from indentation, so no edit was free-typed into the frozen file. A rehearsal harness applied the full packet to a *copy* of the frozen file and ran 112 automated checks (parse integrity, gate wording, threshold values, cross-references, accounting constants) before the real edit was permitted. The harness was itself mutation-tested: nine seeded corruptions of the packet (a dropped section, a stale gate predicate, altered prior and cap constants, a regressed paragraph) were each caught with a non-zero exit, and the unmutated packet passed all checks, a positive and negative control on the instrument, in the same spirit as the paper's other instrumented checks. After the edit, an invariance proof compared the two generations key-by-key: the model hyperparameters, all random seeds, and the verbatim primary-claim sentence are byte-identical across the re-freeze, and the pinned hash was verified with a one-byte-perturbation control (perturbed file fails the pin).

Where the registration text does not describe what the code does. An audit of the arms added at the re-freeze against the frozen protocol found five cases in the registration's own prose. We record them here rather than repair the frozen file: editing a freeze after the fact is the act a freeze exists to prevent, so disclosure is the remedy. (i) The registration says the GMM+BIC arm at DP capacity tags clamping and fallback "by the same convention" as the contiguous arm at DP capacity. It does so only for the fallback tag, which both arms raise by the same rule and which never overwrites an imported count. The clamp differs: the contiguous arm clamps its imported target to half the layer size and reports that clamped target as `k_eff`, while the GMM+BIC arm clamps to the layer size, reports the clamped target as `k_fit`, and uses `k_eff` for the realized cluster count instead. It also emits `k_fit` and `small_n`, which the contiguous arm never writes. (ii) The registration names the reported field `k_import`; the field the artifacts carry is `k_target`, and no artifact carries a field named `k_import`; the analysis transcript names the string only to record this same discrepancy. (iii) The same sentence defers to the tree-metadata convention section, but that section enumerates only three arms and the GMM+BIC arm at DP capacity is not among them; the re-freeze added five arms in all. (iv) That section lists `n_boundaries` among the chapter-boundary arm's keys; the arm's tree metadata carries the section's other keys but not that one, and the quantity it names is recoverable from the `n_segments` the arm does write. (v) The same passage says the GMM+BIC arm at DP capacity shares the UMAP pipeline of the GMM+BIC tree, which runs a global stage and then a cluster-local one. The committed factory is a single flat stage, matching that passage's own headline definition of the arm and the arm's purpose: a second stage would re-introduce the baseline's own cluster-count choices inside the matched capacity.

Three of the five, (ii)–(iv), are naming and cross-reference defects. Two are not: (i) is a difference in the clamp rule applied and in what the reported field means, and (v) a difference between the passage's cross-reference and the pipeline shape its own headline definition specifies; both concern descriptive arms only. A sixth case lies outside that audit's count: the registration's description of the chapter-boundary arm says the leaves inside a detected section are cut with the same size-based arithmetic as the contiguous arm, while the frozen definition in the same passage, and the code, split each section equally into the fewest blocks no larger than the contiguous arm's block size (Appendix E).

Each arm they touch is registered as descriptive, spends no alpha, and enters no success criterion, and every value this paper reports for those arms is read from the emitted diagnostics rather than from the registration's description of them. Each case was checked against the frozen file, and each case that names an emitted field was checked against the diagnostics the run actually wrote.

A further departure is from a registered duty rather than from the registration's description of the code. The parser-shape monitor commits us to printing every monitored completion verbatim (Appendix Z); none of its 122 completions is printed. Beside the per-row correctness the released ledgers carry, a printed completion would carry information about the benchmark's answer key, which is request-gated and not ours to redistribute, and the first registration keeps gold answers out of the paper and out of the model outputs we release; every row keeps its identifier and shape

columns, and the table re-derives from the released records. The frozen file is not edited; the duty is recorded here as unmet for every row. The first registration's closed-book gate also asks for the per-book closed-book distribution, identifying memorization-saturated books; this paper prints its range and median (§4) and identifies no such book, another duty left unmet, although the monitor aggregate the release ships carries the per-book values.

Three more, outside that population. The audit just described ran over the arms added at the re-freeze. Three further cases of the same kind sit outside it, and we state them apart because they touch the confirmatory machinery rather than descriptive arms. The first two are the same defect: a revision updated the binding field and left older text standing beside it. The first is in the section that maps outcomes to claims, whose four cell labels still key the gate of every cell on the single Leiden comparison rather than on the conjunction: three of them then split on the UMAP-matched selector once that comparison is said to pass, and the fourth is that comparison failing. The re-freeze made the predicate conjunctive and says so in the criterion's own definition, which states the conjunctive form, requires both conjuncts reported with intervals and the binding one named, and warns in that same place that changing the hypothesis wording while leaving the criteria untouched would leave the file self-contradictory with a single gate still operating. The revision did reach that section: the abandonment cell carries the conjunctive threshold the same revision introduced. It did not reach the label strings. The analysis code implements the conjunctive gate, with the second conjunct a required argument so that a call site left on the predecessor signature raises rather than falling back silently to one gate. Nothing reported here turns on the discrepancy: the first conjunct did not clear its bound (§4), and the stale label and the implemented gate send that outcome to the same abandonment cell under the same threshold, the implemented label additionally naming which conjuncts bound. This is a different residue from the one in the registered claim text quoted in Appendix Q, which names the same pre-re-freeze comparator for a reason of its own.

The second case is the predecessor spend cap. Outside the change table and the dated recomputations, which record the earlier figure correctly as earlier, six places of standing text still name it as current: the budget abort condition, two clauses of a single freeze precondition, the reason given for running one seed, a note recording an arm rejected on cost, and one sentence inside the budget section itself. The budget section's cap field, its headroom arithmetic and the change table all carry the current figure, and the abort condition delegates its own cost arithmetic to that section, though the parenthesis that delegates also repeats a headroom figure later revisions superseded. The cap this paper reports (§3.4) is the current one, and the predecessor figure appears nowhere in this paper.

The third case is a mismatch between the registration and the code rather than within the registration: the confidence-reranked arm's registered scoring limitation states that candidates with negative cosine are excluded from the candidate set before re-ranking, whereas the deployed code takes the top-k slice on the reranked score first and only then drops candidates whose raw cosine is non-positive (Appendix K). The two orders return different sets only when an admissible node with a confidence of exactly zero is pushed out of the slice by a lower-indexed node tied with it at score zero, which the filter then drops; the released tests pin one such input, and how often it arose in the main run was not measured. We record the three cases because a reader holding the registration will meet that text, and because the frozen file had already met the failure mode of the first two: two revisions before this one it found itself naming two different arm orders after an edit updated one home and not the other, and consolidated them for that reason.

A premise of the second registration. The second registration registers its short non-fiction slice as a negative control on the premise that those documents fit a context window whole. For this study's readers the premise does not hold: each reads a context of at most 8,192 tokens, and the slice's documents run from 12,422 to 58,173 tokens by the index's token count. The frozen text stays unedited.

J.1 The halted QuALITY confirmatory attempt

The confirmatory experiment in this paper was preceded by a pre-registered confirmatory attempt on QuALITY (Pang et al., 2022), on the primary backbone, over the 115-article development split that the benchmark's withheld test labels leave as its usable evaluation set. We unblinded it. Its registered sanity hypothesis, that a hierarchical tree beats the no-tree flat baseline, failed on every tree arm the

sanity family covered, each one-sided lower bound falling below zero; the run was halted at that gate, and the primary and secondary hypotheses were left unadjudicated. That is a confirmatory result, not a development diagnostic, and the registered accuracies quoted earlier in this appendix are its accuracies. We report it because a reader is entitled to ask whether the datasets in this paper were a free choice.

Three aspects of the research history affect its interpretation.

First, the halt fell where a registered rule placed it: the registered ordering adjudicates the sanity gate ahead of the primary hypotheses, so the halt came before the primary hypotheses were adjudicated, although that analysis had already printed every arm's accuracy at the sanity stage (those of GMM+BIC, Leiden and the DP arm are quoted earlier in this appendix), as the successor registration records. What was not registered is the branch that then applied. The rule routed a sanity failure to a suspected pipeline defect and required a recheck rather than a stop; the recheck found no defect, and no registered branch covered a halt with a healthy pipeline. That continuation was decided outside the registration, and the successor registration lists the creation of the missing branch among its changes.

Second, this study is a new registration, frozen before any of its own accuracy data, and not a reanalysis of the halted run. The leaf chunking, the retrieval budget, the seeds and the build and retrieval hyperparameters carried over unchanged, hash-pinned in the new registration and declared there identical to the halted run's own run configuration. The falsification conditions and the registered effect size did not carry over, and every change, those included, is enumerated with its direction in the change table the successor registration carries.

Third, the mechanical reason for the change of dataset was stated before the change: a fixed retrieval budget that is a large fraction of a QuALITY article is a small fraction of a book, so the summary nodes that displaced the leaves they summarize at article length cannot displace them at book length. That reason does not make the move outcome-independent, and the registration does not claim it does: its change table records the change of dataset as an outcome-dependent decision and withdraws the word outcome-independent as an overclaim. What limits what the move could buy is that the displacement argument bore on the sanity comparison alone, an expectation the tree-versus-flat results of §4 did not bear out, and that within QuALITY the primary pair showed no positive trend with document length, so the primary comparison remained a genuine open bet at book length rather than a foregone conclusion. That trend was measured after the decision to move and registered before any accuracy observation on the new dataset. What the bet returned is what this paper reports, including where it returned nothing.

J.2 What the release carries, and what it withholds

The accuracy statistics of the arms and sweeps outside the registered set are the same aggregation of the same released ledgers, condition by condition, except the reader probe of Appendix W and the second registration's reader run reported beside it, and the rebuilt trees of Appendix V.2, whose accuracy figures and intervals are copied from the analysis summaries the release ships, and the halted QuALITY attempt, whose accuracies are quoted from the registration that records it. That probe's per-record parse file does ship, with each record's condition, book, question and recorded correctness, so its accuracies can be re-aggregated from that file, as Appendix Y does for its point deltas; the probe's per-query transcripts and its tree-fetch ledger stay out of the release. The rebuild ships no per-record file: its per-query records stay out of the release, so its contrasts are copied rather than re-aggregated. Every other digit a generator prints is re-derived from, or read as recorded from, files the release carries, apart from a small set of auxiliary read-outs whose inputs are governance records we do not release or the benchmarks' own public files, which Appendix S names with what rests on them, the output of the generators that Appendix S counts as stopping on a file the release does not carry, the development-track answer-flip rates, whose input the release withholds (Appendix S), and the revision counts of Table 20, which come from the code repository's history. That reaggregation takes each question's recorded correctness as given, because no released ledger carries an answer key: the release drops every gold field from its records. A record's chosen letter, read beside its recorded correctness, would give the key's letter for that question, so the released records withhold the choice, as a letter and as an option index, and the completion text of every record whose completion parsed. The answer-change flag ships where the budget-curve table

needs that comparison. It names no option: beside the correctness the records already carry, it adds only whether the wrong answer is the same in both conditions. Completions that the record labels unparsed do ship in the registered run's ledgers, together with the parse stage of every answered record, so the parser-shape monitor's unparseable counts re-derive. Re-deriving those labels therefore needs the benchmark's own key, and access to it differs: the confirmatory benchmark's key is request-gated and non-redistributable, while the transfer benchmarks' keys are public, so their gold labels can be recovered by rejoining the released records to the public benchmark on the released question identifiers; rescoring a record independently needs its chosen letter as well, which the released records withhold. Build caches that embed benchmark source text are withheld where the benchmark's terms restrict redistribution: regenerating the evidence-recall quantities (the appendix tables that compare retrieved contexts with gold evidence spans) additionally requires those caches together with the benchmark's evidence annotations, and reproducing model outputs and tree construction is a separate, resource-dependent step. Result numbers written by hand in digits, rather than printed by a generator, are checked by the build gate against an allowlist that ships with the release (`checks/prose_numbers_allow.tsv`) and records a source for every row; some of those sources are internal records we do not release, among them the development perturbation sweep of Appendix O.1 and the development-phase sweep of Appendix K.

K SEAM IMPLEMENTATION DETAILS

The assignment-confidence scalar. The membership threshold and the scalar c_i (Equation 1) are stated in §3.3, with the interpretation the scalar is and is not given.

The prior, the fitter, and what a mixture over UMAP coordinates is. The truncated variational Dirichlet-process mixture carries stick-breaking weights $\pi_k = v_k \prod_{j<k}(1 - v_j)$ with $v_k \sim \mathrm{Beta}(1, \alpha_0)$ and concentration $\alpha_0 = 1$, α being reserved for the Type-I error rate. In the fits reported here the truncation $T = \min(50, \max(2, \lfloor n/10 \rfloor))$ for n input vectors is an upper bound on the reported K, and the fitted count frequently exhausts it (counts below). Two cautions travel with the model class, as background rather than a warrant for these fits: the DP posterior on the component count is not consistent for a true count (Miller & Harrison, 2013), so we claim no true number of narrative topics; what has been shown to converge in this model class is the mixing measure (Nguyen, 2013). Inference is mean-field variational Bayes with full covariances, matching the baseline's scikit-learn 1.8.0 (Pedregosa et al., 2011) `GaussianMixture` default (the version is pinned in the released `requirements.txt`); the swap changes two of the fitter's numerical settings, both fixed before any run (covariance regularizer 10^{-4} against the library default 10^{-6}; iteration cap 500 against the baseline's 100); the base-measure hyperparameters are library defaults, two of them read from each book's own embeddings (Appendix C). A Gaussian mixture over UMAP coordinates is a working model of the reduced space, not a generative claim about it (UMAP preserves neighborhood structure, not density; Narayan et al., 2021); the rationale is harness parity with RAPTOR plus the matched-representation controls of §3.2 (retry ladder and frozen random states below).

Numerical fallback and seeds. When a fit at truncation T fails numerically, the implementation retries a fixed ladder $T \to \max(2, \lfloor T/2 \rfloor) \to 2$ rather than silently falling back to a different model family. All random states reuse the frozen RAPTOR seed. The remaining pipeline is byte-identical in code to the GMM+BIC baseline's; the intermediate arrays it produces differ by arm.

The convergence check is asymmetric, and we measured whether it mattered. The ladder has no counterpart in the parametric baseline: RAPTOR's reference GMM performs no convergence check, so the DP arm can react to a fit that fails to converge where the baseline would carry such a fit forward unremarked. We therefore replayed every GMM fit the baseline made in its cached trees: 98,670 fits over 60 books on the primary backbone and 98,874 over 60 on the backbone robustness check. Of these, 0 reported non-convergence, and the slowest stopped 9 iterations short of the solver's iteration cap. Each layer's input is read from the cached tree itself; at layer 0 the reduction is the proposed arm's stored one, reused under the two arms' bit-identical clustering inputs and their character-identical reducer call, so these are the baseline's own fits and not another pipeline's; on 117 of the 120 book–backbone pairs the replay also reproduces every cached layer's size exactly, and on the other 3 the layer-0 partition differs by at most 24 clusters. The replay does not separate

the local reducer's nondeterminism (Appendix S) from the reuse of the stored global reduction as the source of that difference, and no fit in those books reported non-convergence either. The stored reduction is read from the proposed arm's build sidecar. The registration attaches, to statements citing sidecar figures, the fidelity of parent membership rebuilt from the sidecars, measured on one NovelQA book's layer refits: symmetric difference 0 on the top three layers and 0.3665% (258 of 70,389) at layer 0; this replay reads only that reduction from the sidecar and rebuilds no parent membership. The asymmetry is real; in these trees it had no occasion to act.

Membership bookkeeping. A component counts as active when it carries the argmax responsibility of at least one node. Nodes below threshold everywhere join their argmax cluster, and the orphan rate is logged.

The deployed ranking order. The confidence-reranked arm multiplies the collapsed-tree cosine score of each node by c_i; nodes without a stored confidence (the unclustered top layer and the root) use $c_i{=}1$. The deployed order takes the top-k slice by whichever score the arm ranks on (the raw cosine elsewhere, the confidence-weighted product here) and then drops candidates whose *raw* cosine is non-positive, so admissibility is judged on the unweighted similarity and no node is judged inadmissible for low confidence alone; ranking ties break by node index under a stable sort. The slice being taken on the reranked score is why the retrieved set can differ from the DP arm's at all (§6). Equation 1 is clipped to $[0, 1]$ with $0 \log 0 := 0$.

How often the truncation binds. In a development-phase sweep, 74.1% of 108 fits used every allowed component; the post-run seed study (Appendix S) finds the count at the ceiling on three of six audited books at every seed, and between 31 and 49 on the rest.

Two-stage simplification. In two-stage builds the stored confidence reflects the global-stage assignment only; local-stage refinement uncertainty is recorded in diagnostics but not consumed, a disclosed simplification.

Where the posterior is stored. The index itself carries one derived scalar per clustered node, the confidence c_i of Equation 1; the full responsibility matrix of every fit, together with the fitted variational parameters, is written at build time to a sidecar file beside the cached tree; the posterior diagnostics of §6 read a gated replay of that file, re-fitted in a separate process and accepted only when every gated field matches. A tree loaded from cache is not re-fitted, so the sidecar is written once, at the fresh build; every DP-mixture tree of the confirmatory run has one.

L Registered secondary read-out and the control table

Read-outs the body summarizes. At the registered top-20 depth every summary slot displaces a leaf slot, so gold-evidence recall is verbatim by construction and bounds evidence delivery, not clustering quality (Appendix O): the flat floor is highest on recall, and an oracle-evidence reader over gold spans beats every arm. Pruning the DP arm's summary layer at five registered levels lowered gold-evidence recall on no book (Appendix D). Giving GMM+BIC the DP arm's per-layer cluster counts leaves an interval still admitting a gain as large as the nominal DP–GMM gap (itself unestablished, §4). The representation-matched Leiden variant (both UMAP stages) sits at 0.5659. The two no-tree rows of Figure 2 (document-order repacking, dense minus lexical) are read here.

Book-hierarchical partial-pooling secondary. A registered secondary read-out fits the same per-question outcomes in a book-hierarchical partial-pooling model (non-centered nested logistic over the 60 books, 3,052 observations; frozen priors and sampler): the posterior probability that the DP−Leiden effect is positive is $\Pr(\delta > 0 \mid \mathcal{D}) = 0.625$, with posterior median $+0.0253$ (logit scale) and 90% equal-tailed credible interval $[-0.1051, +0.1515]$; the registered convergence gate passes ($\hat{R}_{\max} = 1.005$, minimum bulk-ESS ≈ 1156). The registration attaches no threshold to this probability: it is a partial-pooling description of the same primary contrast, not a second gate, and it leads to the same qualitative reading as the cluster-robust interval of §4.

Registered disclosure rows and trivial floors. Three record-level failures are counted separately, and all three are already scored as wrong answers wherever accuracy is reported under the registered rule, so none of them names a population the accuracy read-outs set aside. An *error* record carries a non-null infrastructure or generation error, so no answer was produced; an *unparseable* record produced text from which no option letter could be recovered (no answer step, no chosen index, or no parse stage); a wrong answer is a recovered option that is not the gold one. Table 25 prints the error and unparseable counts separately for the transfer cells, beside the count of empty retrievals. In that sense the worst-arm error rate is 0.0000, the worst-arm empty-retrieval rate 0.0000, and the exclusion rate 0.0000. The registration asks for the per-arm values beside the worst case: the error rate is 0.0000 (closed book), 0.0000 (flat dense), 0.0000 (GMM+BIC), 0.0000 (Leiden), 0.0000 (Leiden, UMAP-matched), 0.0000 (DP-mixture) and 0.0000 (reranking); the empty-retrieval rate is 0.0000, 0.0000, 0.0000, 0.0000 and 0.0000 over the 5 tree arms (GMM+BIC, Leiden, Leiden UMAP-matched, DP-mixture, confidence reranking; the order of Table 1, kept here and below); and the summary share of retrieved evidence, whose minimum the gate reads, is 0.2523, 0.1028, 0.1130, 0.2745 and 0.2913. Multiple choice (four options on 1516 questions and two on 10, the latter kept by registered rule and the reason the random floor is not exactly a quarter) puts trivial floors at 0.2516 (random), 0.2510 (constant first option), and 0.2523 (majority letter).

The zero exclusion rate is a statement about the records, not about the builds: it says that no book is missing from any arm. Some builds failed and were resumed: ten books in the primary cell, over 3 arms (GMM+BIC, two-stage representation-matched Leiden variant and chapter-boundary), and one in the robustness-check cell (cluster-count-matched contiguous grouping) recorded a tree-build error that a later resumed run healed. The registered exclusion rule addresses a failed build and is silent on one that a resumed run healed; we keep these books because their records are complete, so all 60 books stay in the kept set and no comparison in this paper loses a book to these errors. Of the arms involved, the number a confirmatory comparison reads is 1 in the primary cell (GMM+BIC) and 0 in the robustness-check cell (none). We state this rather than leave it to be found: the released per-query ledgers carry the per-record timestamps the resumptions are visible in.

Registered outlier sensitivity, both cells. The registered sensitivity co-report re-quotes each interval without the pre-specified outlier book (the corpus's longest). For the primary cell, excluding it, the DP$-$Leiden difference is $+0.0102$ ($[-0.0111, +0.0315]$, $n = 59$), materially unchanged; the verdict is computed on the full kept set either way. For the floor cell, excluding the same book, the DP-minus-flat lower bound is -0.0138 (point $+0.0071$, $n = 59$), leaving the cell unchanged.

Per-contrast intervals for the anatomy. The DP$-$contiguous difference is -0.0005 ($[-0.0255, +0.0245]$). Against the cluster-count-matched contiguous tree it is -0.0019 ($[-0.0254, +0.0216]$). The cluster-count-matched GMM arm differs from its unmatched form by $+0.0003$ ($[-0.0207, +0.0212]$, spanning zero), and the DP$-$matched difference is $+0.0161$ ($[-0.0068, +0.0390]$). Against the two-stage representation-matched Leiden variant, DP$-$matched is $+0.0111$ ($[-0.0097, +0.0318]$). The document-order repacking shift is $[-0.0289, +0.0114]$ and the lexical-floor gap $[-0.0189, +0.0333]$.

Registered descriptive controls. Table 21 lists the registered descriptive controls other than the chapter-boundary control (Table 13) beside the confirmatory comparators, and Figure 2 draws the same intervals on one axis.

The two descriptive read-outs. Neither read-out carries a registered direction; descriptively, the summary share of retrieved slots rises from 0.1028 (Leiden) through 0.2523 (GMM+BIC) to 0.2745 (DP), while conditional recall of gold evidence spans falls from 0.0700 through 0.0661 to 0.0618. The flat floor, which retrieves only verbatim leaves, sits at 0.0731. And every arm sits below the observed oracle-evidence reference on the questions that carry gold spans: the deltas run from -0.1803 (DP, $[-0.2145, -0.1460]$) to -0.1978 (reranked) among the confirmatory tree arms, with the flat floor at -0.1845. These are the Table 7 cells and carry that table's common support: the 7 questions the oracle cannot serve are dropped from the condition and from the oracle alike. Per-arm resource axes (build time, index size, summary tokens, latency) are tabulated in Appendix U; Appendix V declares post-hoc exploratory read-outs of these outcome records.

Table 21: Registered controls beside their comparators. The upper block repeats the registered arms' book-mean accuracies from Table 1 together with the registered success criteria they enter (SCn labels are the registered success-criterion identifiers: SC0 is flat dense minus closed book, registered one-sided, so its upper bound is open; SC1, DP minus Leiden, and SC4, DP minus GMM+BIC, are the two conjuncts of the primary predicate; SC2 requires each of the GMM+BIC, Leiden and DP trees to clear the flat dense floor on a one-sided test (§4); SC3 is confidence-reranked minus DP (§6); SC5 is DP minus UMAP-matched Leiden; SC6 is UMAP-matched minus raw Leiden (§4); SC7 is UMAP-matched Leiden minus flat dense, one-sided (§4); only SC0, SC1, SC4 and SC5 have rows here), so that every control in the lower block sits in the same table as the arm it is registered against; no alpha is spent on the lower block. "DP" abbreviates the DP-mixture arm; all accuracies are book means. Intervals are cluster-robust 95% over books.

Arm / control	Acc	Comparison	Δ [95% CI]
Confirmatory comparators			
Flat dense	0.5710	this − closed book (SC0)	$+0.1865$ $[+0.1436, +\infty]$
GMM+BIC tree	0.5606	DP − this (SC4)	$+0.0164$ $[-0.0050, +0.0378]$
Leiden tree	0.5674	DP − this (SC1)	$+0.0096$ $[-0.0114, +0.0306]$
Leiden tree, UMAP-matched	0.5703	DP − this (SC5)	$+0.0067$ $[-0.0166, +0.0300]$
DP-mixture tree	0.5770	—	—
Registered descriptive controls			
Fixed-block contiguous (non-semantic)	0.5775	DP − this	-0.0005 $[-0.0255, +0.0245]$
Contiguous at DP capacity	0.5789	DP − this	-0.0019 $[-0.0254, +0.0216]$
GMM+BIC at DP capacity	0.5609	DP − this	$+0.0161$ $[-0.0068, +0.0390]$
		this − GMM+BIC	$+0.0003$ $[-0.0207, +0.0212]$
Leiden, two-stage-UMAP matched	0.5659	DP − this	$+0.0111$ $[-0.0097, +0.0318]$
Document-order repacking (flat set)	0.5622	this − flat dense	-0.0088 $[-0.0289, +0.0114]$
Flat BM25 (lexical floor)	0.5638	flat dense − this	$+0.0072$ $[-0.0189, +0.0333]$

L.1 Stronger flat controls added by the second registration

The second registration adds both of its registered flat retrievers to the frozen comparison, the cross-encoder at a base and, outside the registration, also at a large capacity, 3 conditions in all, and holds everything else fixed: the primary backbone as writer and reader, the same 1,526 questions over the 60 books, and the dense flat index of the registered floor as the first stage. The conditions are rank fusion of the dense and lexical flat indexes, and the dense index reranked by a cross-encoder (bge-reranker) at a base and at a large capacity. These comparisons sit outside the registered SC2 and do not re-judge it; the registration fixes in advance how a stronger flat is written up either way: if it beats a tree, SC2 stands and a separate paragraph says that adding a stronger flat outside the registration widens the gap, and if it loses, the favourable side is not cited alone. Table 22 gives each condition's accuracy in both registered units beside its paired book-level difference against the dense flat floor, the Leiden index and the DP index, with cluster-robust two-sided 95% intervals over books. In point estimate, 3 of the 3 conditions sit above the dense floor, 3 above the Leiden index and 3 above the DP index; of the 9 intervals, 1 excludes zero (rank fusion against the Leiden index). This is why the chapter-boundary control's superlative in §4 is scoped to the first registration's arms.

Because each stronger flat sits above the dense floor in point estimate, adding it outside the registration shifts every tree-minus-flat point estimate toward flat; SC2 stands.

Table 22: Stronger flat controls at the frozen reader, primary backbone, 60 books and 1,526 questions. Accuracy in the registered book mean and the co-reported question mean; each difference is condition minus the named reference in the book mean, paired over books, with a cluster-robust two-sided 95% interval.

Condition	Acc (book mean)	Acc (question mean)	Reference	Δ	95% CI
Rank fusion, dense + lexical	0.5922	0.5813	dense flat floor	0.0212	[−0.0006, 0.0430]
			Leiden index	0.0247	[0.0046, 0.0449]
			DP index	0.0152	[−0.0085, 0.0388]
Cross-encoder rerank, base	0.5789	0.5655	dense flat floor	0.0079	[−0.0134, 0.0292]
			Leiden index	0.0115	[−0.0110, 0.0340]
			DP index	0.0019	[−0.0254, 0.0292]
Cross-encoder rerank, large	0.5853	0.5865	dense flat floor	0.0143	[−0.0157, 0.0443]
			Leiden index	0.0179	[−0.0111, 0.0468]
			DP index	0.0083	[−0.0216, 0.0382]

M TRANSFER CHECK: EVERY CELL, BOTH DATASETS, BOTH BACKBONES

The transfer datasets. As a transfer check, separate from the primary analysis, we use the English subset of DetectiveQA (Xu et al., 2025) (600 questions over 62 detective novels, each indexed in full, a departure from the benchmark's truncation protocol; Appendices M and N) and a 37-question single-document, non-fiction, medium-and-long-length slice of LongBench v2 (Bai et al., 2025) that our own filter, not the benchmark, labels non-narrative. Both are analyzed under one committed analysis revision, though whether every record was generated by behaviorally identical code cannot be verified from the revision strings alone (Appendix J).

What the transfer check is. The transfer check asks whether both seam differences of the primary study, DP−Leiden and DP−GMM+BIC, keep their direction away from the primary benchmark, under a genre shift and a shift out of fiction. It is a descriptive co-report registered apart from the primary hypothesis: the first registration defines no hypothesis, threshold or α on these data, no cell carries a label of any kind, and nothing here is pooled with the NovelQA results or across backbones (the second registration's negative control names this LongBench slice as its long side, Appendix N). The four registered tree arms (GMM+BIC, Leiden, UMAP-matched Leiden, and the DP mixture; the confidence-reranking arm of Table 1 is not in the transfer set) are run on 2 datasets whose roles differ. DetectiveQA (Xu et al., 2025) is a genre shift at book scale (600 registered questions over detective novels; Appendix B), the closest move from NovelQA. The LongBench v2 slice (Bai et al., 2025) is a domain and structure shift: 37 single-document non-fiction questions, a single question per document. Each dataset is run under both backbones, the primary backbone and the backbone of the backbone robustness check (§3.4), each building and reading its own trees, so the table has 16 cells; the (dataset, backbone) populations are reported side by side and never averaged. The registered transfer arm set has no flat dense floor, so these cells say nothing about hierarchical versus flat retrieval, which is the primary study's question (§4).

What the LongBench axis can resolve, disclosed before the data. With a single question per document, each per-document difference between arms takes a value in $\{-1, 0, +1\}$ and the cluster-robust interval coincides with a paired per-document t. Before any transfer record was scored we wrote down what that implies. If the arms differ in *correctness* on a share p of the documents – $p = \Pr[Y_a \neq Y_b]$, which is not the share on which they pick different option letters, since two different wrong letters contribute nothing to an accuracy difference – the 95% half-width, if the disagreements are balanced in sign (otherwise the interval is narrower), is about $2.03\sqrt{p/37}$: 0.105 at $p = 0.10$ and 0.183 at $p = 0.30$, which is 5.4 times the registered design target of 0.034; resolving 0.034 at $p = 0.30$ would take about 1,067 documents, holding that t quantile fixed. The wording rules that therefore bind this axis, whatever it shows, are these: its interval is never read as the absence of an effect; its point estimate is a descriptive observation of direction and, wherever its size is quoted, the half-width is quoted beside it; it is never placed in a table with NovelQA numbers; and no statement that the seam effect does or does not carry over to non-fiction rests on this axis alone.

DetectiveQA is indexed in full, against its published protocol. The benchmark's published protocol truncates each book; its figure caption reads, verbatim, "The novel content for each question is truncated before the answer appears." (Xu et al., 2025). We indexed the full text of every book,

so what is measured here is transfer under the condition that the answer passage is in the corpus, not the benchmark's published setting, and we read no number below as a result on the benchmark's own task. For 323 of the 600 questions (a share of 0.5383) at least a tenth of the book follows the answer position, text the published protocol would have withheld; the residual fraction after the answer, defined as $1 - \mathrm{ap}/N$ for answer position ap and paragraph count N, averages 0.2854 (median 0.1673) under the snapshot's own paragraph count and 0.2852 (median 0.1684) under the paragraph count of the text we indexed, counts that differ on 502 questions without moving the quoted share (323 under the snapshot's count). The number of questions whose stated position lies past the end of the text is 1 under our count and 0 under the snapshot's; such questions stay in the averages. The answer-position field is read by the dataset adapter and is consumed nowhere in the indexing, retrieval or scoring path; it stays adapter-local and is not written into the per-query records, whose schema is locked, and no question was filtered or weighted by it.

Exclusion rule, fixed before scoring. A book whose tree did not build in any arm is excluded from every arm of the same (dataset, backbone) cell, so that each seam difference is a difference over a single set of books. A book counts as not built only when both conditions hold: the build-error sidecar has a row for it in that arm, and that arm's record file holds none of its questions; a book with a sidecar row but complete records was healed by a resumed run and stays in, disclosed as recovered, unless another arm's build error excluded it from the cell. Where a cell excludes a book, Table 24 carries, in the same table, the excluded books, the re-adjusted book count of each contrast, and the leave-one-book-out hull of each interval over the books that remain. Excluded books, by cell: DetectiveQA under the primary backbone, none; DetectiveQA under the robustness-check backbone, none; LongBench v2 under the primary backbone, none; LongBench v2 under the robustness-check backbone, none.

Closed-book control. A closed-book control, in which the reader answers from the question and options alone, was run in every cell. It sits outside the registered arm set, was specified before any transfer record was scored, and is reported as a contamination floor: not as an arm, and not as a way to separate retrieval's contribution. Its book mean per cell, its own record counts, and each arm's margin over it (the arm's book mean minus the control's book mean) are in Table 27, kept apart from the registered tables and read beside the trivial floors of Table 26; on the LongBench rows the margin is not quoted as a magnitude, for the reason given above.

How to read the tables. Table 23 gives every arm's accuracy in every cell; Table 24 the seam differences on the registered book-mean estimand with their cluster-robust intervals, book counts, exclusions and leave-one-book-out hulls; Table 25 the per-arm record counts that the registration asks to be shown beside accuracy; Table 26 the completeness of each cell against the registered totals and its trivial floors. The sign of each seam interval is to be compared directly with the NovelQA intervals of §4; we attach no reading to agreement or disagreement, and a sign that differs from NovelQA is itself the result. On the LongBench rows the rules above apply.

Table 23: Transfer check, accuracy by cell: unweighted book mean, with the question-pooled mean in parentheses. Each column is one (dataset, backbone) population after the symmetric exclusion above, scored by every arm; columns are never pooled. Primary = the primary backbone; Robustness check = the backbone of the backbone robustness check. The per-cell trivial floors and the population of each column are in Table 26.

	DetectiveQA		LongBench v2, non-narrative slice	
Arm	Primary	Robustness check	Primary	Robustness check
GMM+BIC tree	0.4715 (0.4717)	0.4810 (0.4833)	0.2432 (0.2432)	0.4324 (0.4324)
Leiden tree	0.4628 (0.4783)	0.5208 (0.5133)	0.3514 (0.3514)	0.2973 (0.2973)
Leiden tree, UMAP-matched	0.4733 (0.4867)	0.4922 (0.4983)	0.2432 (0.2432)	0.3514 (0.3514)
DP-mixture tree	0.4825 (0.4900)	0.5021 (0.4983)	0.2703 (0.2703)	0.2703 (0.2703)

N TWO FURTHER SLICES, REGISTERED SEPARATELY BY THE SECOND REGISTRATION

What these slices are, and what they are not. The second registration adds 2 LongBench v2 slices (Bai et al., 2025) to the transfer check. It registers neither as further sample for the Long-

Table 24: Transfer check, seam differences on the registered book-mean estimand: point, two-sided cluster-robust 95% interval (cluster = book, or document for LongBench v2, whose rows are therefore a paired per-document t; see the half-width disclosure above), degrees of freedom, and the number of books each contrast is computed over after the exclusion rule. The note line under each cell names the excluded books and, where a book was excluded, the leave-one-book-out hull of each interval (the range of the point and the envelope of the interval over the folds that each leave out a single book); the hull is skipped, with 0 folds, where fewer than three books remain, and no interval is formed, point only, where fewer than two remain; each note line carries the reason recorded by the analysis.

Cell	Contrast	Point	Lower	Upper	df	Books
DetectiveQA, primary	DP−Leiden	+0.0197	−0.0135	+0.0528	61	62
	DP−GMM+BIC	+0.0110	−0.0264	+0.0484	61	62
Excluded books: none. DP mixture minus Leiden: hull n/a (no book excluded). DP mixture minus GMM+BIC: hull n/a (no book excluded).						
DetectiveQA, robustness check	DP−Leiden	−0.0187	−0.0527	+0.0153	61	62
	DP−GMM+BIC	+0.0210	−0.0143	+0.0564	61	62
Excluded books: none. DP mixture minus Leiden: hull n/a (no book excluded). DP mixture minus GMM+BIC: hull n/a (no book excluded).						
LongBench v2, primary	DP−Leiden	−0.0811	−0.1733	+0.0112	36	37
	DP−GMM+BIC	+0.0270	−0.1197	+0.1738	36	37
Excluded books: none. DP mixture minus Leiden: hull n/a (no book excluded). DP mixture minus GMM+BIC: hull n/a (no book excluded).						
LongBench v2, robustness check	DP−Leiden	−0.0270	−0.1228	+0.0688	36	37
	DP−GMM+BIC	−0.1622	−0.2868	−0.0376	36	37
Excluded books: none. DP mixture minus Leiden: hull n/a (no book excluded). DP mixture minus GMM+BIC: hull n/a (no book excluded).						

Table 25: Transfer check, per-arm record health: error records (no answer, or a recorded error) / unparseable answers / empty retrievals, per cell, over the same post-exclusion population as Table 23.

	DetectiveQA		LongBench v2, non-narrative slice	
Arm	Primary	Robustness check	Primary	Robustness check
GMM+BIC tree	0/0/0	0/0/0	0/0/0	0/0/0
Leiden tree	0/0/0	0/0/0	0/0/0	0/0/0
Leiden tree, UMAP-matched	0/0/0	0/0/0	0/0/0	0/0/0
DP-mixture tree	0/0/0	0/0/0	0/0/0	0/0/0

Bench slice registered first, whose cells are in the tables of Appendix M. The user-guide slice takes its 28 questions from a domain under which the slice already registered has no questions at all, with a single question per document. The short non-fiction slice takes the same non-fiction filter at its short length band, 66 questions. It is registered as a negative control and as nothing else, on the registration's premise that these documents fit a context window whole, so that the condition a hierarchy is built for would be absent; that premise does not hold for this study's readers (Appendix J). The registration predicts that the hierarchy's advantage disappears. Neither slice is summed with the other or with the long results. Each slice is read under both backbones, each building and reading its own trees, so Table 28 has 4 columns.

What the constant-answer floors say. We print the constant-answer floors beside the arms (neither registration asks for them on these slices), and on the user-guide slice they carry the read-out. The two floors coincide on this slice, the first option being also the letter most often correct, at 0.3929, and no arm on either backbone reaches it; the highest arm accuracy on that slice is 0.3214. We report that comparison rather than the ranking among the arms, because a set of arms that all sit below a constant rule describes the reader on this slice and not the index. The arms are also not separated from each other there: the contrasts that print as zero in Table 29 are exact ties rather than rounded ones, and on both backbones the DP mixture and the Leiden arm answer the same number of user-guide questions correctly. The short non-fiction slice does not have this property. Its floor is 0.3485, and all 8 arm readings of that slice sit above it, by +0.0152 to +0.1667; the smallest of those margins is a single question, which is the resolution this population supports. Its rows are read against the closed-book control in the same table.

The contrasts, and how many of them there are. Table 29 gives 8 contrasts of these slices, none of them registered, each at the 95% level. Exactly 1 of them has an interval that excludes zero: the DP mixture over Leiden on the short non-fiction slice under the robustness-check backbone, $+0.1212$ with interval $[+0.0199, +0.2225]$. On that cell the Leiden arm scores 0.3939, the closed-book control's score there, so the point estimate of this contrast is also the DP mixture's

Table 26: Transfer check, population and completeness of each cell against the registered totals, and its trivial floors: books registered, loaded by the adapter, and kept after the exclusion rule; questions loaded by the adapter (registered total in parentheses), scored, and excluded through the exclusion rule; duplicate-question records tagged by the loader (kept, tagged only); books excluded and books recovered after a build error (the recovered books are named in the released per-book build sidecars); rows in the build-error sidecar, with the prefix of its pre-scrub sha256 (none where no sidecar was written). Floors are over the same post-exclusion population: random guess at the real option arity, the constant first option, and the majority gold letter (given).

	DetectiveQA		LongBench v2, non-narrative slice	
	Primary	Robustness check	Primary	Robustness check
Books registered / loaded / kept	62 / 62 / 62	62 / 62 / 62	37 / 37 / 37	37 / 37 / 37
Questions loaded (registered)	600 (600)	600 (600)	37 (37)	37 (37)
Scored / excluded	600 / 0	600 / 0	37 / 0	37 / 0
Duplicate-tagged (kept)	5	5	0	0
Books excluded / recovered	0 / 1	0 / 0	0 / 1	0 / 0
Sidecar rows (sha256 prefix)	1 (`5ce07230cb0c`)	0 (`none`)	1 (`7aaba34017f9`)	0 (`none`)
Random floor	0.2500	0.2500	0.2500	0.2500
First-option floor	0.2333	0.2333	0.1351	0.1351
Majority-letter floor (letter)	0.2617 (B)	0.2617 (B)	0.3514 (D)	0.3514 (D)

Table 27: Closed-book contamination control, outside the registered arm set and read as a floor, not as an arm: the control's book mean with the question-pooled mean in parentheses, its error and unparseable counts, and, for the DetectiveQA cells only, each arm's margin over it (arm book mean minus control book mean; no interval). On the LongBench axis the margin is not quoted as a magnitude (half-width disclosure above); the floor itself is given.

	DetectiveQA		LongBench v2, non-narrative slice	
	Primary	Robustness check	Primary	Robustness check
Closed-book floor, book mean (pooled)	0.3289 (0.3200)	0.3253 (0.3233)	0.3243 (0.3243)	0.3514 (0.3514)
Control error / unparseable	0 / 2	0 / 0	0 / 0	0 / 0
Margin, GMM+BIC tree	$+0.1426$	$+0.1557$	not quoted	not quoted
Margin, Leiden tree	$+0.1339$	$+0.1954$	not quoted	not quoted
Margin, Leiden tree, UMAP-matched	$+0.1444$	$+0.1669$	not quoted	not quoted
Margin, DP-mixture tree	$+0.1536$	$+0.1767$	not quoted	not quoted

margin over a reader given no document at all. The two are matched in total only: they answer the same number of questions correctly, and the summary these numbers come from does not say which questions those are. The control carries no interval of its own, so the point estimates coincide and the interval does not transfer. It is the only arm on that cell the control matches; the other three sit above it. At this level and over this many intervals, 0.40 of them would be expected to exclude zero if every underlying difference were zero. No correction is applied and none is registered: these cells define no hypothesis, carry no error budget and enter no verdict, so the count is reported rather than tested. The registered negative control is read the same way. Nothing here licenses a claim that the hierarchy helps on short documents, and nothing here is pooled with the primary benchmark.

The comparison this control was registered on, and why it is absent. The registration states the negative control as a comparison between a tree arm and a flat arm, that is retrieval over the passages with no tree: on the short non-fiction slice the difference between the two should be smaller than it is on the long slice. Neither slice has a flat arm. The registered transfer arm set is four tree arms; the cost model of the same document buys eight configurations per document on each of these two slices, four arms under two backbones, and no flat condition is among them; and the long slice registered first has no flat arm either. The quantity the control names is therefore not computable from what was run, on either side of the comparison it asks for, and we report that rather than substitute something for it. The two halves of the registration do not agree here: the section that states the control asks for a comparator that the section pricing the runs does not buy. That is a property of the registration, and we disclose it as one.

A weaker question, reported under its own name. Table 28 carries a closed-book row for every cell, and an arm's distance from that row asks whether the reader gains anything from being given the document at all. That is not the registered comparison and does not stand in for it: closed book is no retrieval, while a flat arm is flat retrieval, so a margin over closed book cannot separate a hierarchy from a flat index. Read that way, the arms of the short non-fiction slice sit at or above the closed-book control in every cell, and on the user-guide slice under the primary backbone every arm

Table 28: The two further registered transfer slices. Each column is one slice as read under a single backbone; the columns sit side by side and are never pooled. The top rows are the registered tree arms. Closed book is the contamination control: the reader answers from the question and its options with no document at all. The control is not one of the registered arms; it is a contamination floor and not a retrieval floor. Best letter is the accuracy of always answering the single option letter that is most often correct on that slice, with the letter in parentheses; first option is the accuracy of always answering the first option. Accuracies are question-pooled over the slice's scored population.

	User guide		Non-fiction, short	
	primary	robustness check	primary	robustness check
GMM+BIC tree	0.3214	0.2143	0.4091	0.4697
Leiden tree	0.2857	0.1786	0.3636	0.3939
Leiden tree, UMAP-matched	0.2143	0.1429	0.3939	0.4697
DP-mixture tree	0.2857	0.1786	0.3939	0.5152
Closed book	0.3571	0.1071	0.3485	0.3939
Best letter	0.3929 (A)	0.3929 (A)	0.3485 (C)	0.3485 (C)
First option	0.3929	0.3929	0.1515	0.1515
Questions	28	28	66	66

Table 29: The contrasts of the two further slices (none of them registered). Each point is the mean over documents of the per-document accuracy difference, each document weighted equally, and each interval is the cluster-robust interval at the level the run recorded. With a single question per document, a per-document difference takes a value in $\{-1, 0, +1\}$ and the interval coincides with a paired per-document t. Rows are reported side by side and never averaged.

Slice	Backbone	Contrast	Point	95% interval	Books
User guide	primary	DP mixture − GMM+BIC	-0.0357	$[-0.1642, +0.0928]$	28
User guide	primary	DP mixture − Leiden	0.0000	$[-0.1492, +0.1492]$	28
User guide	robustness check	DP mixture − GMM+BIC	-0.0357	$[-0.1642, +0.0928]$	28
User guide	robustness check	DP mixture − Leiden	0.0000	$[-0.1055, +0.1055]$	28
Non-fiction, short	primary	DP mixture − GMM+BIC	-0.0152	$[-0.1162, +0.0859]$	66
Non-fiction, short	primary	DP mixture − Leiden	$+0.0303$	$[-0.0835, +0.1441]$	66
Non-fiction, short	robustness check	DP mixture − GMM+BIC	$+0.0455$	$[-0.0453, +0.1362]$	66
Non-fiction, short	robustness check	DP mixture − Leiden	$+0.1212$	$[+0.0199, +0.2225]$	66

sits below it. Both readings are descriptive, neither is the registered control, and neither licenses a claim about hierarchy on documents of these lengths.

O DIAGNOSTIC PROCEDURE AND PRE-RUN CHECKS

O.1 THE AUDIT, STEP BY STEP

The audits named in §1 were not run side by side; they were run in an order, and the order is the part that carries beyond this index. We set it out as a procedure, with the same scope the results have: each step is a way of arranging evidence, not a finding.

Step 1: price the swap before reading it. Hold the pipeline's interface fixed, vary only the grouping operator, and report each index's build cost beside its accuracy (Figure 3, Figure 2) and its variation across rebuilds beside both (§5). Read the interval against a target fixed in advance rather than the sign of the difference: here the registered δ^* sits above one comparator's interval and inside the other's, and those are different statements (§4).

Step 2: make the hierarchy face the cheap controls first. Before a difference is attributed to the grouping rule, put the same retrieval budget against references that store no summaries at all and against controls that group without semantics (contiguous blocks, document structure, and a cluster-count-matched variant of the arm itself; §4). Flat retrieval and the simple structural controls are inexpensive references, so running them before the attribution, rather than after a reviewer asks, is affordable; the count-matched controls answer a different question and can cost as much as the semantic index they bound (Appendix U). What they establish is co-variation among implementation factors, not mechanism-level attribution.

Step 3: name the variation a stored confidence is stable under, then check it where retrieval consumes it. A confidence quantity written beside the index is only as meaningful as the source of variation it survives, and stability under one source does not transfer to another: within-fit reassignment stability did not imply stability across refits (§6). Whatever the diagnostic says, the quantity has to be evaluated at its registered consumer (for us a fixed multiplicative reranker) because that is the only place retrieval actually reads it (§6).

The three practices of §7 map onto these steps: the first covers Step 1 with Step 2, the second is the rebuild clause of Step 1, and the third is Step 3.

How gold-evidence recall is computed. A gold span counts as recovered only if it appears as a substring (casefold, whitespace-collapsed) of the assembled reader context. The assembled context interleaves bracketed block labels between passages and the matcher does not strip them, so *any* span straddling two retrieved chunks scores as a conservative miss, adjacent or not; we confirmed on the confirmatory records that no counted hit was straddle-derived. The release carries how many snippets occur only across a chunk boundary (next paragraph), not that check's result. Pruning summary nodes did not lower this recall on any of the 60 books at any of five pruning levels (Appendix D); that observation does not establish that summaries contribute no unique verbatim matches.

A recoverability reference for the metric. The metric is a verbatim match against the assembled context, and most gold snippets cannot be matched verbatim from the leaves of our extracted corpus, so reporting the recalls of Table 7 without that reference invites the reader to over-read them. Of the 2573 gold snippets in the confirmatory band, 77.8% do not occur verbatim anywhere in the harness's own extracted book text (whether the gold was annotated against a different edition or the difference arises in text extraction we do not determine), and a further 2.0% occur only across a chunk boundary, where the paragraph above makes them unmatchable from leaves; only 519 of 2573 snippets (0.2017) sit wholly inside a single leaf. This is a source-and-chunk recoverability reference for verbatim leaf retrieval under our extraction, normalization, and chunking, not a universal ceiling for contexts that include generated summaries. Normalized by that reference, the per-arm recalls of Table 7, as printed there, are 30.2–36.2% of it: flat dense 36.2%, GMM+BIC 32.8%, Leiden 34.7%, UMAP-matched Leiden 34.3%, DP mixture 30.6%, and confidence-reranked 30.2% (every row of Table 7 that carries a recall cell); the raw recalls stay in that table, and the rescaling changes no comparison drawn from it, since the reference is a property of the gold text and the leaf chunking, both of which every arm shares.

Pre-run diagnostics (no additional LLM calls) and frame reproducibility. Wherever this paper calls a diagnostic "pre-registered", it means fixed in a committed protocol document before the

corresponding outcome was observed; those documents are, apart from the two pre-run diagnostics' combining rule (which that registration fixes), separate from the NovelQA registration of §3.4, which contains no decision table for the diagnostics of §6. The headline range of §6 (ARI 0.64–0.79, mean 0.72) is defined as follows: each value is one book's ARI between a fit on row-permuted input and the fit on the original ordering, taken on the DP-mixture stage with the reduced embeddings held fixed, the setting in which the two fits' log-likelihoods are comparable; the permutation control of §6 refits the reduction as well and scores lower. The eighteenth book of that section, the one in the inconclusive band, is also the one book on which a permuted restart, not a Gibbs candidate, holds the highest mass.

Is the mean-field approximation to blame? With conjugate priors the collapsed mass of a partition under a diagnostic model (marginal likelihood times a Chinese-restaurant-process prior, the untruncated limit of the build's truncated stick-breaking prior, not the same object) is computable exactly, without any variational update. On 17 of 18 books it concentrates within the candidate catalogue (margins and the inconclusive band fixed in advance, in the paragraph below), yet the highest-mass candidate is the variational fit's own labeling in *none* of the 18 books: a collapsed Gibbs candidate wins on 17 of 18, by 141.5–621.8 nats, at ARI 0.39–0.62 from the variational solution. These are point masses at a finite catalogue of partitions, not total posterior mass, so they settle neither whether the posterior over partitions is narrow nor how much of the narrowness the approximation adds. No index in the registered analysis is built from a Gibbs partition, and the post-hoc rebuild that builds one (Appendix V.2) carries no registered status; two pre-run diagnostics requiring no additional LLM calls and a frame reproducibility check point the same way and license no causal or superiority claim.

The collapsed-mass diagnostic, as frozen. Every rule below was fixed in that diagnostic's specification before that diagnostic scored any book; we restate them because neither the summary in §6 nor the paragraph above is enough to reproduce the comparison. The scored quantity is the exact collapsed log-mass of a *given* partition c under a conjugate model, $\log w(c) = \log P_{\text{prior}}(c) + \sum_k \log m_{\text{NIW}}(X_k)$, with a Chinese-restaurant prior at concentration 1 and a normal–inverse-Wishart term at $\kappa_0 = 1$, $\nu_0 = 10$ (the corner the variational fitter itself uses; the $\nu_0 = 12$ convention is co-reported and does not move the verdict). The build's truncated stick-breaking prior is carried alongside at truncation 50. Collapsed-Gibbs chains generate candidate partitions, but nothing is sampled from the collapsed posterior at scoring time and no chain is treated as a converged estimate of posterior mass: these are point masses at named partitions, which is why they bound nothing about total posterior mass. The catalogue scored per book is generated from the variational fit's own labeling, 8 draws from that fit's variational posterior, 2 collapsed-Gibbs harvests (500 sweeps at seeds 224 and 225, used as a candidate generator with no mixing claim attached), 16 row-permuted restarts with the reduction held fixed, full-arm variants of those that refit the reduction as well, refits under the 10^{-7} perturbation, k-means and Ward solutions at $k \in \{5, 10, 15, 20, 30\}$, and small-truncation variational fits at $T \in \{5, 10, 20\}$; after candidates that reduce to the same canonical labeling are merged, 44–45 distinct partitions per book remain. A candidate *other than the best one* is *far-away* when its adjusted Rand index against *the best candidate's* labeling is at most 0.90 (reported at 0.85 and 0.95 as well). The variational fit's own labeling is one of the ranked candidates here, not the reference, and it is itself far-away on every book. A book is *concentrated* when every far-away candidate falls below the best one's mass by at least ten orders of magnitude, *flat* when the margin stays within two orders of magnitude, and in the *inconclusive band* between the two. Positive and negative controls on synthetic mixtures at three sizes had to pass before any book was scored. Two diagnostics requiring no additional LLM calls (measured before the main run, with their combining rule, not their outcome, pre-registered) point the same way while licensing no causal or superiority claim. On the eight confirmatory books with the fewest leaves, a collapsed-Gibbs mixture over the same reduced embeddings concentrates on fewer active components than the truncated variational fit reports (a statement about those fits under the diagnostic prior used there, not about the exact posterior of the registered model), and the pre-registered label that would have read this optimistically was reached on zero books. A posterior-predictive check over all 60 books places the hard-assignment component count in the lower tail of the model's own predictive distribution wherever it is extreme at all. Writing $p = (c + 1)/(R + 1)$ for a cell in which c of the $R = 200$ replicates lie at least as far out as the observation, and calling the cell extreme when $p \leq 0.01$ (that is, $c \leq 1$; a level at which these counts equal those the registration reports, though it states no level),

the count is extreme in the lower tail on 6 of 60 books and in the upper tail on none (0 of 60).[3] Both are consistent with the truncated variational approximation spreading mass widely over active components; the reading is registered as an exploratory diagnostic, not a test; it spends no error budget and touches no gate.

Does the diagnostic frame itself reproduce? The exact-mass computation reproduced bit-identically on every book when repeated within a process, and the within-process instabilities disclosed in Appendix S did not manifest in these runs. We also repeated the whole diagnostic in a separate process on a later date, with the population pinned book by book rather than reselected: every per-book and aggregate quantity reported here came back identical, and the only field that differed anywhere in the output was the wall-clock seconds. That population is the 18 development sidecars scored here together with the one that this diagnostic's protocol, which is not among the released registrations, excludes by rule. The specification that froze this diagnostic does not enumerate them, which is a gap in that protocol and not in the record: the released development driver selects that subset by a deterministic rule – a fixed minimum-length filter, a fixed permutation seed and a fixed question budget over the public benchmark – so the identifiers are regenerable from the release, and regenerating them that way returns exactly the population the diagnostic ran on. The diagnostic runs on development books, but the code path and frozen constants are the ones the confirmatory index builder uses; nothing in the diagnostic is specific to the development corpus.

The perturbation axis and its zero-information control. The 10^{-7} perturbation scale of §6 is representational, not semantic (a one-book encoder-swap cell: Appendix V), and is not the yardstick the body argues from. A zero-information control (row-order permutation, which leaves the point cloud unchanged but, unlike the row-permuted refits of §6, also refits the reduction) is not separable from the 10^{-7} axis by sign test ($6/18$ books, $p = 0.24$), while a bootstrap interval on the mean difference excludes zero ($[-0.024, -0.002]$); both are reported. The reduction's share of that sensitivity is reported below, under "Checks on the source of the stability gap, in full". Whether movement of this kind is specific to the grouping rule under test was measured separately, on the development sample and outside the registered evaluation: at the same 10^{-7} scale, on 19 development books at two perturbation seeds, we refit five grouping rules and scored each against its own untreated fit. That sweep is a separate instrument from Table 4 rather than a wider version of it: it runs on 19 books where Table 4 runs on 18, and it hardens overlapping clusters last-cluster-wins where the rows of Table 4 compare MAP labelings, so an adjusted Rand index from one is not comparable with one from the other. Inside the sweep, the controls separate the rules. A rule that never reads the embeddings, fixed-block contiguous grouping, is unmoved when they are perturbed, at ARI exactly 1.0000 on all 19 books. This shows the instrument can register stability where it is present. Holding the algorithm at Leiden and changing only whether a dimensionality reduction precedes it, the mean ARI falls from 0.9030 to 0.6070, a loss of 32.78% of the stability measured without the reduction, with the reduction-free variant the more stable of that pair on 18 of the 19 books; the DP arm's rule is in turn above the BIC-selected Gaussian mixture on 18 of the 19 books (two-sided

[3] On 21 of the 60 books the observed component count sits at the truncation ceiling of 50, which makes the lower-tail direction of this cell uninformative by construction: no replicate can exceed a saturated observation. On 6 of those every replicate is also at the ceiling, so the cell is degenerate in both directions and the reading on those books rests on the check's 2 other discrepancy statistics, the largest component share and the Gini coefficient of the component sizes. On the remaining 15 the upper direction is informative (between 134 and 199 of the 200 replicates reach the ceiling), and it is the only direction in which the sentence above can be tested on a saturated book; the smallest upper-tail p among them is 0.672. The registered protocol states the stronger form, that all 21 cells are degenerate in both directions, and requires this note wherever this posterior-predictive cell is quoted; we report the measured split rather than the registered characterization. The registered form understated what the check can see: it set aside all 21 saturated cells in the upper direction, whereas only 6 are uninformative there, so the upper-tail count above rests on 54 books rather than the 39 the registration allowed, while the lower-tail count rests on 39 books under either reading. The replicates are drawn from each book's own fitted variational weights under a truncated stick-breaking scheme whose final stick is held at the whole remaining mass, and then as multinomial counts; they are a predictive spread over component sizes, not the assignment spread of §6. The caution that attaches to any quantity read off a fitted posterior applies to them all the same: on the development sample, MAP reassignment under sampled variational parameters moves one fitted hierarchy's partition far less than row-permuted refits of the same model do (§6), so a spread taken from a single fit describes certainty within that fit rather than stability across refits. Those refits showed no clear difference in plug-in log-likelihood, which is not evidence that the data support them equally. That diagnostic was measured on the development books, not on the books scored here.

sign test, exact $p = 40/524{,}288$), which is the paired form this comparison takes, both rules being measured on the same books. Two qualifications travel with that sweep, and they point in opposite directions. Its extreme values concentrate on a single book, the largest in the sample; dropping that book makes the Leiden pair unanimous (18 of 18 books) and widens its loss to 34.19%, while the DP-against-mixture margin weakens to 17 of 18 books. And the run carries a control it did not pass: on that same largest book the BIC-selected mixture and the DP arm's rule did not reproduce their own labeling under an untreated refit, at ARI 0.99988 and 0.99947, a departure orders of magnitude smaller than the movements read above but a failed control all the same. It is also the only quantity above that does not repeat: re-fitting the whole sweep returns every other figure above exactly as printed and moves these two in the fourth or fifth decimal. The sweep is therefore reported here with that book kept in. The scope is 19 development books from a single dataset, at one embedding model, one reduction setting and one perturbation scale, over two perturbation seeds, and the causal isolation is within Leiden alone; we do not extend it to RAPTOR-style indexes in general. The axis establishes only that the fit moves under changes carrying no information; §6 replaces it with the row-permuted refits, which Figure 4 draws book by book against the within-fit axis as a post hoc, exploratory read-out that spends no error budget.

The labeling-asymmetry control. Row *B* of Table 4 (categorical sampling in place of MAP) explains just under a seventh of the gap between rows B′ and A of §6 ($+0.0412$ of $+0.2996$), and the perturbation-axis ARI barely differs between refits that did and did not change the component count.

Rows B′ and P book by book. Figure 4 (§6) draws rows *B′* and *P* of Table 4 for each of the 18 development books: the filled mark is that book's mean ARI to the reference fit under parameters drawn from q with MAP relabeling, the open mark is its mean ARI after a row-permuted refit with the reduced embeddings held fixed, and the segment joins the pair; books appear in the order the driver reports them. Across the sample the paired difference P minus B' has mean -0.2497, book-resampling bootstrap 95% CI $[-0.2700, -0.2288]$, and P sits above B' on 0 of 18 books (sign test, Holm-adjusted within the driver's four-contrast family, $p = 3.1 \times 10^{-5}$). No per-point interval is drawn: the driver reports no interval, and each mark is the mean of the draws it does report. The figure is post hoc and exploratory: the registered contrast set carries no row for it, it spends no error budget, and its scope stops at the development sample.

The parameter draw and the tempering, exactly. The summary in §6 does not say how a draw is taken, so we state it here. The active set is not re-derived inside a draw: it is the reference fit's, a component counting as active when it carries the argmax responsibility of at least one node (Appendix K). *(i)* A vectorised Beta over *every* truncation slot gives the stick-breaking weights; those weights are then restricted to the active components and renormalised over them, so no inactive truncation slot, the unused tail among them, carries mass into the draw. *(ii)* For each active component in ascending index order, a precision matrix Λ_k is drawn from the fit's Wishart factor at ν_k degrees of freedom, with the scale matrix taken as the inverse of $\nu_k C_k$; the mean is then drawn from a normal centred at m_k whose covariance is the inverse of $\beta_k \Lambda_k$, *conditional on that draw*, so the mean's spread follows the covariance actually sampled rather than a plug-in. *(iii)* Node assignments are sampled categorically from the resulting log-scores; the MAP labeling that axis B′ scores is the argmax of the *same* draw's log-scores, with step *(iii)*'s uniform vector still consumed, so the two labelings differ by that categorical layer alone and not by the random stream. The tempering substitutes N_k/λ for every effective count in the conjugate updates, writing λ for the temperature, a scalar distinct from both the truncation T and the precision Λ_k above: the scatter, the per-component counts, and the tail counts entering the stick-breaking Beta are divided by λ, while the Wishart cross term is recomputed at the reduced count because its coefficient is not linear in N_k; the prior terms are untouched. With the fitted responsibilities held fixed, these are conjugate updates of fractional sufficient statistics; they neither re-optimize the responsibilities nor recover the full tempered mixture posterior, and no refit is run; the tempered object is a posterior for a down-weighted likelihood, and λ is reported as a measurement, never adopted as a setting for any arm.

The within-fit against the perturbation axis, in full. The within-fit axis barely moves the partition and the perturbation axis moves it far more (rows B′ and A of Table 4): the gap is $+0.2996$ on average (per-book range $+0.2287$ to $+0.3760$), positive in 18 of 18 books (sign test $p = 7.6 \times 10^{-6}$;

book-resampling bootstrap 95% CI $[+0.2823, +0.3161]$). The perturbation axis is not the yardstick §6 argues from.

Checks on the source of the stability gap, in full. The rule for choosing among the three explanations of §6 (artifact, approximation, model) was fixed before any discriminating measurement. Permuted refits with the reduced embeddings held fixed show no clear mean difference in plug-in mixture log-likelihood from the untreated fit ($+0.011$ nats per row, permuted minus untreated; bootstrap 95% CI $[-0.002, +0.026]$; 7/18 books worse, sign test $p = 0.48$). That interval covers zero but is not an equivalence test, and it compares a plug-in likelihood rather than the variational objective. Part of that permutation sensitivity is contributed by the reduction rather than the mixture: permuted refits that also re-derive the reduction sit 0.0632 lower in mean ARI than the row-permuted refits used in §6, which hold the reduction fixed. A tempering diagnostic agrees under its own likelihood scaling: the weight must be cut 16–32-fold before draws from the variational posterior move as much as those refits do.

Draw-count sensitivity (post hoc). The within-fit axes B and B′ rest on two parameter draws per book. Re-drawing with $M = 2, 8$, and 32 draws per book on the same 18 sidecars (the $M = 2$ stream reproduces the per-book values above exactly; $M > 2$ use independent generator streams) leaves the 18-book mean of B′ at 0.9668, 0.9662, and 0.9670 (book-bootstrap 95% intervals $[0.9642, 0.9694]$, $[0.9635, 0.9689]$, and $[0.9645, 0.9694]$) and of B at 0.9256, 0.9247, and 0.9260; the mean within-book standard deviation of B′ across draws is at most 0.0053. Nothing is refitted, so axes A and P are unchanged.

Post-hoc split by source overlap. Because five of the 18 development books are the same books as NovelQA titles (Appendix R), we re-read the per-book statistics quoted in §6 and above, split by that membership (5 versus 13 books; 11 numeric statistics, each with a two-sided exact permutation test of the difference in group means over all 8,568 splits, and 3 categorical labels; no multiplicity correction). This is a post-hoc analysis, not a registered one. The headline permutation-ARI range is attained within the 13 non-overlapping books alone (0.640–0.791; group means 0.717 and 0.717, $p = 0.994$), the margins of the highest-mass candidate over the variational fit and the tempering factors are not separated ($p = 0.56$ and $p = 1.00$), and the qualitative conclusions hold within each group: 5 of the 5 overlapping books are concentrated and 5 misplaced, against 12 and 12 of the 13 others (the non-overlapping group includes the inconclusive-band book), and the narrowness is present in the non-overlapping books alone (permutation ARI 0.717 against draws from the fitted variational posterior at 0.965). Read as a sensitivity rather than as a group comparison, the separation §6 rests on does not depend on the shared titles: dropping the 5 overlapping books moves row B′ from 0.9668 to 0.965 and row P from 0.7171 to 0.717, so the distance between them is what survives, not either level on its own. Three of the 11 statistics reach or approach the nominal level; with no correction over that many they are read whole, and no one of them is a finding: the exact-mass gap to the nearest far-away partition is larger on the overlapping books (163 against 73 nats, $p = 0.005$), draws from their fitted variational posterior are marginally tighter (0.971 against 0.965, $p = 0.033$), and their perturbation-axis ARI is higher (0.70 against 0.66, $p = 0.053$). The overlapping books are also longer (2995 against 2508 leaves on average), a confound we do not disentangle. If anything, the shared books are the more concentrated ones.

P THE BOOKKEEPING IDENTITY: STATEMENT AND PROOF SKETCH

A fixed frame for reading the comparisons. Without a fixed frame, §4's many pairwise comparisons could be read opportunistically. Three audit labels (reader residual, displacement, evidence absence) apply one per incorrectly answered question and partition the error event. The framing is in our claim register, with its definitions in the paragraphs below. The decomposition is bookkeeping under a shared-operator design, not a causal theorem: it records what co-varies, not which class carries a difference; we execute it post hoc on the 49.2% of evidence-bearing questions whose gold spans can be located in leaf text (Appendix X), no registered claim rests on those shares, and we make no causal claim (Lin & Liu, 2026).

The claims that rest on the definitions below are entered, with their status, in the claim register of Appendix Q. The causal status of this decomposition is stated in the opening paragraph of this appendix.

Setting. Fix a query distribution Q, a budget B, and operators (S, ρ_B) shared across arms and satisfying A1–A4 (stated below); index-construction arms differ in the partition operator π and in what travels with it (Table 30). For each arm, every incorrectly answered question carries exactly one audit label under the disjoint definitions that follow (the opening paragraph of this appendix names them): *reader residual* (the context covered the gold evidence and the reader still failed), then *displacement* (not the first, and coverage would have held with summary nodes excluded from budget competition), then *evidence absence* (the remainder). The three predicates are mutually exclusive as written (displacement excludes the covered case explicitly and evidence absence is the remainder), so the listing order records the definitions rather than resolving ties between overlapping indicators. The labels are defined that way whether or not the audit is executed, and the proposition below concerns the defined quantities; Appendix X reports what the assignment yields when it is run post hoc on the questions whose gold spans can be located.

Bookkeeping identity 1 (contrast-to-component map; proof sketch and full map below) *Under A1–A4, for two index-construction arms in which every operator other than π really is held fixed (among the semantic-family contrasts of Table 30, the DP-versus-GMM+BIC row), the gain in expected correctness equals the reduction in total risk, term by term over the three classes, and every term varies only through π, since π is the only operator that differs. Each contrast of Table 30 then records* which components change together. *It does not narrow which classes can carry the difference: holding an operator fixed does not hold its risk term fixed, because that term is a functional of the context the partition produces, so a fixed reader can still contribute a different reader residual under a different partition.*

That π is the only operator that differs holds exactly for the DP-vs-GMM+BIC contrast, which shares the two-stage representation pipeline; in rows where the map below records that representation or fit scaffold varies as well, the difference is a joint function of the recorded set of changed components, and no term is a function of π alone.

Bookkeeping identity 1 (restated)**.** *(i) For each arm, $R(\pi) = R_{\mathrm{read}}(\pi) + R_{\mathrm{disp}}(\pi) + R_{\mathrm{abs}}(\pi)$. (ii) For any two arms, $\mathbb{E}[Y(\pi')] - \mathbb{E}[Y(\pi)] = \sum_c \big[R_c(\pi) - R_c(\pi')\big]$, the sum running over the three labels. (iii) Given the shared operators, both sides are functionals of the partition operators alone.*

Proof sketch. (i) The disjoint definitions assign exactly one label to each error event, so the indicator of an incorrect answer decomposes as a sum of three disjoint indicators; take expectations over Q. (ii) Subtract the two decompositions and use $\mathbb{E}[Y] = 1 - R$. (iii) Under A1–A3 the maps from π to tree, summary set, and context are fixed functionals of π; under A4 the answer distribution depends on the context alone, so no term depends on the arm label except through π. No step uses independence assumptions or a generative model for the corpus. The statement is bookkeeping over a shared pipeline, which is exactly its value, and its limit. ■

Remark. The audit labels and the mechanism vocabulary of the rest of the paper align only descriptively: displacement is shared between the two; the evidence-absence class is a compound of the retrieval, partition, and summarization mechanisms; and reader residual has no mechanism counterpart at all: an informal account assumes a competent reader, while an exhaustive partition of the error event needs an explicit category for the context being sufficient and the answer still wrong. The class is therefore two different things under the two coverage definitions, and we name them apart: under (A) it is the *reader residual* proper, and under (B) a *coverage-conditional residual*, wider than reader failure, because ancestry covering the localized evidence does not establish that the evidence survived summarization, so it can also be a summary that never carried the fact. The labels partition the error event operationally; that partition does not license the semantic reading. The entanglement paragraph below states why the recorded variables do not separate these explanations under the present audit. Contrasts record which components change together; they neither narrow which classes can carry a difference nor label records.

Assumption register.

(A1) *Summarizer invariance*: S is arm-invariant: same model, same prompt, one shared deterministic cache.

(A2) *Retrieval invariance*: ρ_B is arm-invariant at the same B (iso-budget); no index-construction arm gets a different selector.

(A3) *Shared leaves*: every arm starts from the identical leaf layer.

(A4) *Exclusion*: conditional on the query and the context C, Y does not depend on the arm; the arm label reaches the reader through C and through nothing else.

A1–A3 are implementation facts checkable from the harness in the release (shared cache keys, one chunker, one retriever); A4 is the design assumption that the shared-reader, greedy-decoding protocol exists to approximate.

A corollary of A4, and what it is worth. A4 alone yields a bound that needs none of the labels. Write μ_π for the joint law of $(Q, C_\pi(Q))$. Under A4 the reader's conditional accuracy $g(q, C) = \mathbb{E}[Y \mid q, C] \in [0, 1]$ is the same function for every arm, so for any two index-construction arms

$$\big|\mathbb{E}[Y(\pi')] - \mathbb{E}[Y(\pi)]\big| = \big|\mathbb{E}_{\mu_{\pi'}}[g] - \mathbb{E}_{\mu_\pi}[g]\big| \le \mathrm{TV}(\mu_\pi, \mu_{\pi'}) = \Pr_Q\big(C_\pi(Q) \neq C_{\pi'}(Q)\big),$$

the last step an equality rather than an inequality because both arms share the query marginal and build the context deterministically from it, so pairing on Q is an optimal coupling. If g depends on the context only through a statistic ϕ, the same argument applied to ϕ gives the tighter $|\Delta| \le \Pr_Q\big(\phi(C_\pi(Q)) \neq \phi(C_{\pi'}(Q))\big)$.

Measured, the first form is vacuous. Across the 6 arm pairs the two arms hand the reader a different context on 69.4–92.7% of questions (post hoc and exploratory; book mean over 60 books, 1526 questions, budget 2000 tokens, primary backbone only). On the audited subsample of App. X, where the accuracy difference can be recomputed on exactly those questions, that discordance is, for every pair, at least 27 times the difference measured for the same pair. The bound therefore excludes nothing we could have observed, and quoting it without that factor would dress a tautology as a result. What it does rule out is the opposite reading, that the arms score alike because they retrieve alike. They do not retrieve alike.

The refinement is available under an additional sufficiency assumption. Take ϕ to be gold-evidence coverage, the statistic the three labels are built from, and measure on the 748 questions of App. X where coverage is computable (post hoc, exploratory, and on that appendix's measurably biased subsample). The bound falls to 0.9–6.1% under coverage definition (A) and 9.2–17.7% under definition (B), against raw context discordance of 71.9–92.7% on those same questions: the arms hand over different text on most questions and differ in evidence coverage on far fewer. "Far fewer" is not "never": under (B) it still reaches 17.7% of questions. We do not assert the antecedent (g plausibly depends on more than coverage) and state it so the reader can see what it would buy. It is falsifiable rather than assumed: an observed difference above the bound would refute it, and on the audited questions the observed differences stayed within the bound in 6 of 6 pairs under (A) and 6 of 6 under (B).

What neither form says. A loose bound together with a small observed difference does not establish that the reader is insensitive to the context question by question. Two situations are observationally identical here: g barely moves for almost every query, or it moves a great deal with mixed signs that cancel in the book mean. Both bounds are properties of the contexts, and neither separates them.

Scope of the frame. The frame covers the index-construction arms, which differ in π and in what travels with it (Table 30); the confidence-reranked consumer (§3.3) varies the retrieval operator by design, the document-order control varies context assembly, and the oracle-evidence reference bypasses retrieval, so each of those sits outside the frame and is read only through its own registered contrast. The bound's dependence on π alone is therefore a property of contrasts in which every other relevant operator really is held fixed; a row that moves more than the grouping estimates a joint difference between bundles of operators, not the grouping's contribution inside that bundle.

Why the evidence-absence class does not decompose. The third class is itself a compound of three mechanisms: the selector ranked poorly over the nodes that existed (retrieval), the grouping tore related material apart so that no cluster ever contained the evidence together (partition), or the

Table 30: The contrast-to-component map of Bookkeeping identity 1: what each contrast (registered unless marked †) holds fixed beyond A1–A4, and the comparison its difference reports; no row identifies a single mechanism. Arm vocabulary follows §3.2.

Contrast	Held fixed beyond A1–A4	What the difference compares
contiguous vs. GMM+BIC†	node type and budget; grouping semantics removed	semantic grouping against a fixed non-semantic grouping reference; summary and budget differences travel with it
DP vs. Leiden (first conjunct)	nothing further: the full partition-operator swap	the total difference against the published graph-community index, all classes jointly (grouping operator, representation, and fit scaffold differ together)
DP vs. GMM+BIC (second conjunct)	the two-stage representation pipeline shared; the clustering rule swapped	the same overall performance comparison under the operator swap, against the original mixture index; the conjunctive primary passes only if both conjuncts do
UMAP-matched Leiden variants vs. Leiden (two-stage variant†)	clustering algorithm and all downstream operators; the representation pipeline is what varies here (matched to the mixture arms, not held fixed between these two)	the sensitivity of the Leiden arm to its representation pipeline; not a causal decomposition of the DP–Leiden gap
DP vs. UMAP-matched Leiden	the global UMAP stage the mixture arms use	the DP arm against Leiden grouping; local reduction and fit scaffold differ with it
DP vs. two-stage-UMAP-matched Leiden	the two-stage UMAP skeleton; each rule's local reduction runs on its own groups	the DP arm against Leiden grouping under the matched skeleton
DP vs. fixed-block contiguous	node type and budget; grouping semantics removed	the DP arm against a fixed non-semantic grouping; summary and budget differences travel with it
cluster-count-matched contiguous vs. DP	per-layer cluster counts matched, semantics removed	the DP arm against a contiguous tree with the same per-layer cluster counts; semantics removed, summary text not matched
cluster-count-matched GMM vs. DP	per-layer cluster counts matched, semantic family kept; fit scaffold not matched (one flat GMM per layer vs. the two-stage global+local fit)	two constructions under the same per-layer cluster counts; fit-scaffold differences remain
cluster-count-matched GMM vs. GMM+BIC	one family, per-layer cluster counts freed from the BIC search; fit scaffold not matched (one flat GMM per layer vs. the GMM+BIC arm's two-stage global+local)	a comparison in which cluster-count selection and fit scaffold change together
chapter vs. contiguous	contiguity and block-size parameter kept, split equally within each chapter; upper layers identical	whether respecting author-written divisions at layer 0 changes anything, on the parsed subset with the registered three-way reporting (values side by side; no difference or interval is registered)
budget curve	π fixed per arm; the context cap B varies and retrieval depth rises from top-k 20 to 40 at the largest cap (the main run's 2,000 plus 1,000/4,000/7,400; the raise was intended to let the budget rather than the retrieval width bind; measured, top-k still binds at the upper points, Appendix G)	the total downstream response to the budget per arm; the reader's behaviour changes with the context it reads, so this is not the B-path alone

cluster was right and the summary text dropped the content (summarization). The variables recorded here do not separate this compound across arms, and the obstruction is in this design rather than in the effort spent on it: changing π changes both what every summary sees and which summaries exist, so "the grouping was right but the summary failed" is not, under this audit, a judgement that transfers across arms.

Entanglement, continued. Changing the partition also changes how many summary nodes compete for the budget, which is why displacement has its own instrument (the budget curve) rather than a subtraction. The cluster-count-matched rows of Table 30 exist for the same reason from the other side: they pin the one structural quantity (per-layer cluster count) that would otherwise be a standing alternative explanation, up to the fit-scaffold difference disclosed in §4: the cluster-count-matched GMM is a flat fit per layer, not the two-stage global+local fit of the DP and GMM+BIC arms. A broad survey makes the case for causal methods in LLM development and evaluation (Frauen et al., 2026), beside the identification disclosure that opens this appendix; A1–A4 are enumerated above for the reader's benefit rather than to discharge an identification burden we have not taken on, and no section varies an assumption and reports how the reading changes.

Setup. A document is chunked into leaves $x_{1:n}$, shared by every arm. A partition operator π groups the current level's nodes into clusters, possibly overlapping, and recursing yields the tree $T(\pi)$. "Partition" names the operator here, not the shape of its output: the soft-threshold rule of §3.3 may place one node in several clusters, so π is in general a cover. Every adjusted Rand index in this paper scores a set partition; where π yields a cover, what it scores is a *hardening* of π. The hardenings used include argmax responsibility for a fit, categorical sampling (row B of Table 4), last-cluster-wins in the development sweep of Appendix O, and, for a built overlapping layer, first-cluster-wins, beside which some comparisons also report agreement restricted to singly assigned nodes; on the only cell where a refit's first-cluster-wins labels are scored against the shipped layer's own last-cluster-wins hardening (Appendix V), the two agree only at ARI 0.2732. A stability statement about a global fit's hardening is therefore not automatically a statement about the constructed hierarchy, and we do not carry one into the other. A summarization operator S writes one node per cluster; a retrieval operator ρ_B maps a query q to an ordered context sequence C of nodes drawn from $\mathrm{nodes}(T(\pi))$ within token budget B; a reader produces a correctness indicator $Y(q, C) \in \{0, 1\}$ on the multiple-choice item. The risk of a partition operator is $R(\pi) = 1 - \mathbb{E}_q\big[Y(q, \rho_B(q, T(\pi)))\big]$.

The audit predicates, in full. The three labels named in the opening paragraph of this appendix are assigned by the disjoint definitions given earlier in this appendix, listed in the same order. The replay that defines *displacement* is well-posed because the selector scores each node against the query independently and fills B in rank order (§3.2, Figure 1; Appendix K), so excluding summaries leaves the surviving order unchanged. *Evidence absence* is the case where the selected context does not cover the evidence and removing summaries from the budget competition would not have repaired it.

The contrast-to-component map. Table 30 gives the map named in Bookkeeping identity 1, contrast by contrast.

Q CLAIM REGISTER

Each claim in the register below is entered in a claim ledger that records the claim's text, its type, the evidence that would adjudicate it, and the adjudication once made; read-outs outside the register, including the registered moderator read-out of Appendix E, the routing result of Appendix R and those of the second registration, are adjudicated where they are reported. We reproduce that register here so that the map from claim to verdict is auditable rather than reconstructed by the reader. The register holds 7 claims, and all of them are carried in this manuscript; the fourth to seventh entered at later revisions than the first three, and their entries below say so. Table 31 places them on one time axis.

Table 31: The register on one time axis. *Fixed* says when the statement's rule was settled and against what; *Status* says what it is after adjudication. The regimes below are not interchangeable: rules inside the frozen protocol are *prospectively registered* (C-1, C-2 and C-3); a rule settled outside it after an outcome had been seen is an *outcome-informed follow-up* (C-5); and material written after the confirmatory verdict was in is *post-hoc exploratory* (the audit that C-4's framing defines). Neither the protocol hash nor the answer-key seal is anchored by a third-party timestamp; the ordering rests on internal version-control and run ledgers, as Appendix J states.

	Fixed	Status
C-1	In the frozen protocol, before the answer key was unsealed; the only claim that spends the first registration's alpha	Adjudicated §4, not supported; the registered verdict is powered abandonment
C-2	In the frozen protocol; descriptive, spends no alpha	Adjudicated and reported for the primary backbone, in part for the robustness backbone, subset citation forbidden by the registration
C-3	In the frozen protocol, carrying its own deletion rule	The kill-switch fired (§6); carried as a registered negative, not as support
C-4	Framing entered in this register at a later revision than C-1, C-2 and C-3, in neither preregistration; it gates nothing	The per-record audit it defines ran only after the confirmatory verdict, on a subsample that differs from the full band (Appendix X), and is exploratory
C-5	Pre-specified outside the primary registration, and registered in the second as hypothesis H7 (test H7a not yet read out); every rule of the collapsed-mass diagnostic fixed before that diagnostic scored any book	Reported (§6); compares two sources of assignment variation and estimates neither calibration nor posterior width
C-6	Written after the earlier attempt was halted at its registered sanity gate	Disclosure of research history; adjudicates nothing and appears in no decision table
C-7	Written after a generative-AI review pass over the manuscript (AI use statement); the second registration's words point both ways on the quantity, which we treat as unregistered	Exploratory seed-averaged paired contrasts (Appendix C.1); spends no alpha, replaces no registered verdict and appears in no decision table

C-1: confirmatory; the only claim that spends the first registration's alpha. Quoted rather than retyped from the protocol registered under SHA-256 prefix `8e40dfc4c5e5` (§J); during review the generator reads that document's English rendering, in which this sentence is byte-identical:[4]

> *VB-DPMM tree construction, under the unchanged flat collapsed-tree retriever and QuALITY-frozen hyperparameters (single seed 224), achieves higher NovelQA multiple-choice accuracy than a Leiden graph-community tree (Liu et al. 2026 design) on the full public-domain NovelQA set (60 books, 1526 questions), under the PRIMARY backbone Llama-3-8B-Instruct, with a pre-registered Qwen2.5-7B-Instruct replication reported as backbone-generalization evidence (decision 0057). Framed as a hierarchical retrieval-index construction method (decision 0056).*

That sentence records what the comparison was expected to show before the answer key was unsealed. The result did not support it, and the paper built around it is an audit of the assumptions the sentence takes for granted rather than a demonstration of the advantage it names; the register is reproduced unedited so that the difference stays legible. The sentence above names the Leiden comparator alone; the registered predicate that adjudicates it is conjunctive, the conjunction having replaced a single Leiden comparison at the re-freeze (§3.4, with the change table of Appendix J). The predicate governs, and the outcome does not turn on which of the two is read: the Leiden comparison on its own does not resolve either, at $+0.0096$ ($[-0.0114, +0.0306]$). *Adjudicated* in §4; *not supported*. The conjunctive predicate resolves in neither direction: DP minus Leiden is $+0.0096$ ($[-0.0114, +0.0306]$) and DP minus GMM+BIC is $+0.0164$ ($[-0.0050, +0.0378]$), while the DP index stores $5.0\times$ the Leiden index's summary tokens and takes 28.05 h to build against the Leiden index's 6.83 h (Appendix U). The registered verdict is powered abandonment: no evidence of an effect of 3.4 pp or more, the DP-over-Leiden direction discontinued, and the null not thereby proven. Neither interval is contained in the registration's ± 1.0 pp practical-significance band (a

[4] "Decision 0056"/"0057" are identifiers in the project's internal decision log, which is not part of the supplementary material.

conventional reporting threshold, not a derived value), so the within-band descriptive label registered for this contrast does not apply either; no non-inferiority or equivalence margin was registered for it. The equivalence read-out of §4 borrows the power target δ^* for that role after the outcome was known, which is why it is labelled post hoc there and spends no α. The registration does fix a one-sided upper limit, but for the separate displacement question of §4, where it is reported.

C-2: descriptive; spends no alpha. How the index-construction rule shapes retrieval composition (leaf versus summary share of a fixed retrieved budget), gold-evidence recall inside that budget, and index budget (effective cluster count, summary-node count, summary tokens). Registered together with a standing language constraint: reported as estimates with intervals, never in hypothesis-test vocabulary. The registration also forbids post-hoc excavation and promotion of new mechanism metrics outside it; the post-hoc read-outs of Tables 36 and 41 are labelled exploratory and settle no registered comparison. *Adjudicated* in §4 and Appendices C, F, G, H and I, reported for the primary backbone in full for the arms of Tables 7 and 17, while for the other tree arms of Table 6, the fixed-block contiguous control among them, the registered analysis computes only the index-budget cells; for the backbone robustness check only some of the cells reported for the primary backbone are printed, and that run's released analysis transcript carries them. The composition and recall cells of Table 7 and Table 17 are printed as point values without the intervals that constraint asks for, a departure from the registration.

C-3: registered with a kill-switch. Whether the stored assignment signal is informative enough that confidence-weighted reranking improves on the index it reranks. The registration attaches its own deletion rule: if the reranked arm does not exceed that tree, the claim is withdrawn and reported as a negative ablation instead. *Adjudicated* in §6; the kill-switch fired. The reranked-minus-DP difference is -0.0181 ($[-0.0338, -0.0023]$), with the reranker changing the retrieved node set on 79.0% of queries and the final answer on 11.7% (§6), and the result appears in this paper as that registered negative: not as evidence for the proposed improvement, and retained as an empirical result of the audit.

C-4: framing; carried as Appendix P. A decomposition of the index's error event into named audit labels, together with a statement (Bookkeeping identity 1) of which components each registered arm contrast changes together. It is a framing rather than a registered measurement: it spends no alpha and gates nothing. The per-record audit it defines *is* executed, but only after the confirmatory verdict and only on the questions whose gold spans can be located in leaf text; Appendix X reports those shares together with their exploratory status and the ways that subsample differs from the full band, and Appendix P states the same limits. Earlier drafts of this manuscript listed C-4 without developing it; it entered at a later revision, and the internal decision log, not part of the supplementary material, records both states and the change.

C-5: descriptive diagnostic; spends no alpha. The within-fit versus refit assignment-stability diagnostic of §6: on the 18 development books, MAP reassignment under parameters sampled from the fitted variational posterior (two draws per book, active components fixed to the reference fit's) stays far closer to the reference labeling than refits after row permutations of the same reduced embeddings do, refits for which no clear mean difference in plug-in mixture log-likelihood was detected (§6, Table 4); it compares two sources of assignment variation and does not estimate the calibration of the stored responsibilities or the width of the full partition posterior. The highest-mass candidate partition is not the fit's own on any book. The candidate-wise exact masses are point masses within a finite catalogue: they do not measure total posterior mass, so they do not separate model concentration from approximation error, and the register claims no such separation. It is an outcome-informed follow-up rather than a prospectively registered test, and the register carries it as a diagnostic, not as a hypothesis: it spends none of the first registration's alpha, and the second registration tests the same statement as its hypothesis H7, whose confirmatory test H7a is not yet read out; the NovelQA registration contains no decision table for it; its own decision table was fixed after the narrowness had been observed on these books and before the discriminating measurements were run (§6); and its scope stops at the development sample, five of whose books are also NovelQA titles (Appendix R), so it licenses no statement about the confirmatory corpus. The two samples carry disjoint questions and labels and are analyzed separately, but their source documents partly coincide, so statistical independence between this diagnostic and the confirmatory results is not established

and we do not read them as independent evidence. It entered the register at a later revision, after the diagnostic had been delivered; the internal decision log (not part of the supplementary material) records the entry.

C-6: disclosure of research history; spends no alpha. The halted QuALITY confirmatory attempt of Appendix J.1: that the confirmatory experiment in this paper was preceded by a pre-registered confirmatory attempt of ours, that this attempt was unblinded and halted at its registered sanity gate with its primary and secondary hypotheses unadjudicated, and that we report it. It is a statement about how this study came to be run, not a measurement: it spends no alpha, adjudicates nothing, and appears in no decision table. Its register entry carries the limits the disclosure must keep if it is to be worth making. The earlier run is a confirmatory result, so the accuracies it produced are not described anywhere in this paper as development diagnostics. The halt is not presented as fully scripted: the registered rule routed a sanity failure to a suspected pipeline defect and required a recheck, the recheck found no defect, and the branch that then applied was undefined, so the appendix states that gap wherever it leans on the halt. The change of dataset is not presented as outcome-independent, because the successor registration's own change table records it as an outcome-dependent decision and withdraws that word as an overclaim. What carried over unchanged is limited to what a configuration hash pins, and what did not carry over is named alongside it. Nor is the earlier outcome allowed to bear on this paper's verdicts: it licenses no statement about the confirmatory corpus, and where the appendix gives the pre-data expectation behind the change of dataset it also gives the result that did not meet it. It entered the register at the latest revision, discharging a disclosure obligation this project registered for itself before the confirmatory run began; the internal decision log (not part of the supplementary material) records the entry.

C-7: exploratory interval; spends no alpha. The seed-averaged paired arm contrasts of Appendix C.1: each census book's accuracy averaged over the 5 construction seeds, the paired difference between arms over the 54 books, and the book-level t interval the primary analysis uses, for the DP-mixture arm against its three registered comparators. The second registration declares a seed-averaged arm difference unregistered and withdraws the arm effect from its model path on that ground, while another of its lines asks for the first registration's bootstrap, applied to the seed mean, to be reported as well; its words point both ways, and we treat the quantity as unregistered and exploratory: it spends no alpha, enters no decision table, and neither replaces nor sharpens the registered verdicts of that appendix. It is reported because the released shards determine it and withholding a descriptive interval they fix would be selective reporting; its values enter the manuscript only through a generator whose controls pin the per-book seed spreads to the adjudicated artefact. It entered the register at the latest revision, after a generative-AI review pass over the manuscript (AI use statement) asked for the quantity.

R Posterior responsibility versus cosine geometry: full study

Summary. A natural hope is that build-time responsibilities carry an uncertainty signal that predicts geometry-defined routing misses. Tested observationally on the 69 InfiniteBench En.QA documents (Zhang et al., 2024) (source of the 18-book sample, analysed separately from the registered NovelQA evaluation, with documented source-text overlap), posterior-derived ambiguity predicts such misses above chance (document-mean AUC 0.5882), but a proxy scoring the same fitted partition by cosine geometry alone predicts them better (0.6535). On a different outcome, second-choice hit rate among missed pairs, the document-paired difference also favours geometry (-0.0928, 95% CI $[-0.1095, -0.0772]$); that number is a hit-rate difference, not a difference of the two AUCs. This is a negative development result for posterior responsibility as a stand-alone signal on this routing proxy, outside the registered evaluation; it evaluates neither gold-evidence relevance nor a posterior-plus-geometry combination. The sweep over space, dimensionality and distribution family, its failing negative control, the definitions and the three untested gaps that bound the claim are set out in the remainder of this appendix.

A natural hope for a Bayesian nonparametric index is that the posterior responsibilities produced at build time carry an uncertainty signal that could predict retrieval failures, and eventually license query-time actions. We tested this hope directly, as an observational study on a development set (InfiniteBench En.QA (Zhang et al., 2024)), analyzed separately from the pre-registered NovelQA

evaluation and its registered quantities, on disjoint questions and with partial source-document overlap. That overlap, measured post hoc (8-word shingle containment; same-book threshold 0.5; noise floors: NovelQA-versus-NovelQA 0.059, unmatched En.QA-versus-NovelQA 0.003; an edition-level rewording control, one word in sixty replaced, still scores 0.87): 5 of the 18 development books are the same book as one of the 60 NovelQA books (containment 0.81–0.88, word-length ratio at least 0.98, so whole book against whole book), the other 13 score at most 0.003, the replay-gated nineteenth book is also a NovelQA book, and 21 of all 69 En.QA books match a NovelQA book (5 of them only partially, the En.QA context covering part of the NovelQA text). No question, gold label, or registered quantity is shared, so the independence claimed is of the evaluation, not of the source text. The replay gate that excludes the nineteenth development book rejects a stored fit that does not reproduce from cached inputs in an independent process (every gated field compared: arrays to 10^{-5}, memberships and activity exactly, scalar gate fields to 10^{-9}; the 18 passing books replayed bit-exact on all 1,131 fits, and the excluded book differs by 1.0 in responsibility on one of its 94 fits, so the tolerance separates nothing that a tighter one would not). Posterior-derived ambiguity does predict geometry-defined routing misses (Table 32) above chance (document-mean AUC 0.5882, 95% CI $[0.5681, 0.6098]$). However, a proxy that scores the same fitted partition by cosine geometry alone (the clusters and the miss label come from the same fit; only the ambiguity score is posterior-free) predicts them better (0.6535, $[0.6363, 0.6703]$), and on the routing read-out the posterior trails geometry: the document-paired difference in second-choice hit rate is -0.0928 ($[-0.1095, -0.0772]$). The three numbers are different quantities on different populations; Table 32 fixes them.

Table 32: Study B quantities. Each is computed per document and summarized by the unweighted document mean with a percentile bootstrap over documents (1,000 resamples, seed 224; the von Mises–Fisher run of Table 33 uses 5,000).

Quantity	Definition
Population	69 En.QA documents; per document, every (question, leaf) pair over the 50 leaves most cosine-similar to the question.
Miss label	The leaf's hard parent cluster is not the cluster whose centroid is nearest the question (a routing miss at the first choice).
AUC (posterior / geometry)	Within-document AUROC of the leaf's ambiguity score for predicting the miss label. The responsibility score is the leaf's second-largest responsibility, $\gamma_{(2)}$; the geometric score is the second-largest minus the largest leaf–centroid cosine similarity, $s_{(2)} - s_{(1)}$, on the same embeddings. Each is reported as its document mean; the two scores are never combined into one predictor.
Second-choice hit rate	Among missed pairs, the share whose second-choice cluster is the question's nearest cluster. The second choice is a target, distinct from the scores above: under the posterior it is the cluster with the second-largest responsibility; under the geometric proxy it is the centroid nearest the *leaf's* embedding among clusters other than its hard parent.
Difference	Per document, posterior hit rate minus geometry hit rate; its document mean is -0.0928. It is not a difference of the two AUCs.

The sweep is exploratory: its rows carry weaker controls than the registered path (whose protocol is not among the released registrations) and are read as a diagnostic (Controls, below), and configurations whose negative control fails do not corroborate the primary finding. Sweeping the space, dimensionality, and distribution family: in low dimensions the posterior signal exists but still trails geometry; at ≥ 200 dimensions no document meets the predefined scoring rule under a Gaussian family: on $69/69$ documents the hard-argmax assignments occupy fewer than three components (Table 33 counts the one- and two-component cases separately), so those documents are not scored under the minimum-occupied-component rule rather than lacking a variational posterior; and a von Mises–Fisher family at $d = 768$ keeps every document scorable (median κ 950, range 683–1206) but its responsibilities agree closely with the geometric ordering: agreement between posterior-based and geometry-based orderings reaches 0.7730 $[0.7662, 0.7798]$ (vs. 0.4317 for the Gaussian), and in that family the document-paired difference in second-choice hit rate (posterior minus geometry) has a bootstrap upper bound of $+0.0013$, about $+0.1$ pp, so even where the two orderings nearly coincide the posterior shows no clear advantage over geometry on the reported read-outs. In none of the distribution families or spaces we tested did posterior responsibility show a clear advantage over

cosine geometry on either read-out of Table 32; the number of documents retained for this routing analysis differs by configuration (Table 33), so the rows of that table are not a like-for-like sweep.

Table 33: Documents retained for the routing analysis in the Study B sweep, of 69 En.QA documents. Scored documents satisfy the analysis rule $K_{\text{used}} \geq 3$, where K_{used} is the number of components occupied by the hard argmax assignments; excluded documents have K_{used} equal to one or two and are counted separately. Exclusion under this rule is not the absence of a variational posterior: with two occupied components both signals can only name the one other component, so the second-choice read-out is degenerate, and the predefined rule requires at least three. Every mean is over the retained documents only, so the configurations differ in population as well as in space. The third K_{used} count is the scored n each row's means are taken over. Second-choice difference is posterior minus geometry hit rate (document mean, percentile-bootstrap 95% interval); AUCs are document means. The last column is *this sweep's* shuffled-posterior negative control, whose null lift is zero: † marks a configuration whose 95% interval excludes that null, and "—" a configuration with no scorable document, where the control was not measured at all rather than passed. Their signs and sizes are read below; the registered path's own control, which covers its null, is a different measurement and is reported there too. Every interval from the Gaussian runs, here and in the summary paragraph that opens this appendix, is drawn from a bootstrap generator keyed by (configuration, statistic) under seed 224 with 1,000 document resamples, so the primary row reproduces the interval quoted there exactly; the von Mises–Fisher row's interval comes from a separate script, which takes the same keyed rule and seed but 5,000 resamples; the scripts behind this table are not in the released code, though the files they wrote are.

Space	Family, covariance	Docs by K_{used} 1 / 2 / ≥ 3	Second-choice difference posterior − geometry [95% CI]	AUC posterior / geometry	Negative control lift
UMAP-10 (primary)	Gaussian, full	0 / 0 / 69	-0.0928 $[-0.1095, -0.0772]$	0.5882 / 0.6535	$-0.0036^{\dagger}$
PCA-10	Gaussian, full	0 / 3 / 66	-0.0764 $[-0.0935, -0.0586]$	0.5627 / 0.6191	$-0.0089^{\dagger}$
PCA-10	Gaussian, diagonal	0 / 0 / 69	-0.0644 $[-0.0887, -0.0376]$	0.6315 / 0.6546	$+0.0002$
PCA-50	Gaussian, diagonal	37 / 27 / 5	-0.2038 $[-0.4431, -0.0303]$	0.5974 / 0.7970	$+0.2843^{\dagger}$
PCA-200	Gaussian, diagonal	47 / 22 / 0	—	—	—
raw 768-d	Gaussian, diagonal	54 / 15 / 0	—	—	—
raw 768-d	Gaussian, spherical	60 / 9 / 0	—	—	—
raw 768-d	von Mises–Fisher	0 / 0 / 69	-0.0052 $[-0.0111, +0.0013]$	0.6255 / 0.6272	$-0.0029^{\dagger}$

A post-hoc sensitivity run that relaxes the rule to $K_{\text{used}} \geq 2$ (not the pre-specified rule; the run writes to its own directory, whose summary records `registered_rule: false`) changes no sign. In the scored two-component cases of this sweep (every one with $K_{\text{eff}} = K_{\text{used}} = 2$), a document is scored only when some top-M retrieval event misses the query's cluster, and its second choice is then the only other component, so each enters with a second-choice hit rate of 1.0 under both signals and a difference of exactly zero. The three two-component PCA-10 full-covariance documents move that row's second-choice difference from -0.0764 to -0.0731 (95% CI $[-0.0895, -0.0556]$). Of the 27 two-component PCA-50 documents, 25 have no missed event and two enter, moving that row from -0.2038 ($n = 5$) to -0.1456 ($n = 7$; $[-0.3381, -0.0216]$). In the three configurations without a three-component document, exactly one two-component document is scorable, the same document in all three, with a posterior AUC of 0.5 against 0.6388 for geometry. The registered rule therefore excludes documents on which the second-choice read-out carries no information.

Controls: on the registered path the blocking controls pass – the shuffled-posterior negative control covers its null on both tests – and a recomputation of the RQ-B1 primary contrast from that path's recorded responsibilities reproduces it with maximum per-document deviation 0.000000. The sweep is weaker, and we report it as a diagnostic rather than a controlled measurement: the same negative control's interval excludes zero on 4 of the 5 scored configurations. On 3 of them the departure is small and *negative* (-0.0089 to -0.0029, at most about 6.4% of that configuration's own lift over chance on the second-choice test, the statistic whose null the control shares), the opposite sign from the leakage this control exists to catch; on PCA-50 it is $+0.2843$, 105% of that row's lift over chance (not of its posterior-minus-geometry difference), and the conclusion here does not rest on that row. Three gaps remain untested and bound the claim: low-dimensional vMF, a vMF–DP variant with a posterior over the occupied component count, and feature operators over the full responsibility vector γ. We therefore treat the result as a negative development result for posterior responsibility as a stand-alone signal on this geometry-defined routing proxy (it is why this paper develops no additional query-time mechanism on it), while noting that it is a dev-set observation, separate from

the registered evaluation reported above; that the target is itself defined by the cosine geometry one of the two signals is built from, so the comparison is not neutral between them; and that it evaluates neither retrieval failure, nor gold-evidence relevance, nor the quality of the stored responsibilities as uncertainty, nor a combination of posterior and geometric features.

S LIMITATIONS IN FULL

The post-hoc equivalence read-out and the abandonment label, in full. The DP−Leiden difference of §4 corresponds to Cohen's $h = +0.0193$. Read after the fact as an equivalence question at that same target (two one-sided tests at $\alpha = 0.05$, equivalently a 90% interval inside ± 0.034), the DP−Leiden difference clears it ($[-0.0080, +0.0271]$, $p = 0.012$) while the DP−GMM difference does not ($[-0.0015, +0.0342]$, $p = 0.052$, the upper limit 0.023 pp outside); δ^* is a power target and no equivalence margin was registered, so this read-out spends no α. The label follows the registration's threshold rule on the realized between-book dispersion s (Appendix J), not achieved 0.80 power, and the realized s sits near that threshold, so we lean on the intervals rather than the label (statistic, threshold, and the margin against the statistic's own estimation noise: below).

The five boundaries of §7, in the form their measurements require. *Reader.* The confirmatory run uses a single 8B open-weights reader; a same-protocol backbone robustness check on a second 7B reader resolved the same way, so the registered cross-backbone cell is primary FAIL × replication FAIL, which bounds how far the result is reader-specific but does not extend it upward. That check is itself a powered abandonment at its own α (no evidence of an effect at the registered design target), with conjuncts $+0.0045$ ($[-0.0168, +0.0259]$) over Leiden and $+0.0036$ ($[-0.0146, +0.0217]$) over BIC-selected GMM, never pooled. Both co-reports of §4 follow it there. All three of its per-arm floor criteria fail, at one-sided lower bounds -0.0330, -0.0320 and -0.0312, with -0.0250 for the UMAP-matched arm; and on one-sided upper limits that backbone shows all three advantages over flat smaller than $+0.02$ ($+0.0001$, -0.0028 and $+0.0055$), including the DP arm, where the primary backbone could not exclude a larger gap. Contamination is diagnosed by the closed-book floor rather than excluded by dates for this reader too: that model's vendor states no pre-training or knowledge cutoff for it (none appears on its published model cards, in the technical report we cite, or in its release announcement), so the only time information we can record is an upper bound, its 2024-09-19 release, which postdates the benchmark's public release; we cannot exclude item exposure, so its closed-book floor (0.4107) is reported as a diagnostic and the check is read conservatively: exposure would dilute the information its nominal intervals carry, and the sign of any interaction with the hierarchical arms is not identified by this design. The two backbones are asymmetric in what can be said here: the primary reader's model card states a base-pretraining cutoff ("The pretraining data has a cutoff of March 2023 for the 8B"), while no such statement exists for this one; the confirmatory decisions rest on the reader whose cutoff is published. Stronger readers (the regime where flat document-order baselines have been reported to beat hierarchical indexes (Laitenberger et al., 2025)) are outside our registered hypotheses; the budget sweep probes context allocation, not reader capacity. A post-hoc, exploratory read of the frozen indexes by a larger same-generation reader is reported in Appendix W; it carries no registered status, no error rate, and is not pooled with the confirmatory records. The reader's base-pretraining cutoff predates the benchmark's public release, but the vendor does not disclose the instruction-tuning data window, so the cutoff argument covers base pretraining only; residual instruct-stage exposure is diagnosed by the closed-book floor, not excluded by dates, and that floor is not an upper bound: it measures what the reader answers with no document, not what exposure adds once a document is supplied. In the registered comparisons, summaries are written by the same model that answers, so a summary can inject parametric knowledge absent from the leaf text, an interaction that blurs the tree-versus-flat comparison and is disclosed rather than controlled. The validity floor SC0 (flat dense minus closed book) compares an open-book prompt that says "Read the context and answer the multiple-choice question" with a closed-book prompt that says "Answer the multiple-choice question using your own knowledge", so its difference carries an instruction-wording component besides context presence; the registration records this confound, and we read SC0 as evidence that retrieval does work, not as a pure context-presence effect. *Seeds and shape.* Tree construction runs under a fixed seed with a single-threaded reduction, and the confirmatory accuracy numbers use the single registered seed; construction is still not fully reproducible under that seed, and measured exceptions are disclosed below rather than averaged over. A post-run sensitivity study on six confirmatory books (primary backbone only) shows, however, that the fitted partition is not stable across seeds: refitting the whole stack (dimensionality reduction included)

under three seeds moves the layer-1 partition by pairwise ARI 0.19–0.59 (mean 0.44), worst on the longest book (the regime this study targets), while the fitted component count moves far less (at its truncation ceiling of 50 on three of the six books at every seed, and between 31 and 49 on the other three). Two caveats bound the ARI reading. These are first-cluster-wins hard labels, which overstate disagreement where leaves are multiply assigned, and the multiply-assigned share itself grows with book size (from 7% to 59% across books and seeds), so the size gradient is confounded with this labeling artifact, and we do not rank books or conditions on it; the lowest agreement and the highest multi-assignment simply co-occur on the largest book. On the covariance axis, where a common-support variant was also computed (leaves singly assigned under every shape), agreement there rises to 0.56–0.77; with single-assignment labels alone it is 0.55–0.71. The axis rests on 1,044 fits over 6 books and 4 covariance shapes, of which 0 failed to converge, and the multiply-assigned leaves summed over the books number 21,842 (full), 25,497 (tied), 21,142 (diagonal) and 22,097 (spherical). The registration asks for all of these together. What the study does measure is how far accuracy moves when the grouping changes by more than a seed does: the retrieval arms of Table 1 above its last two rows, from flat dense, which groups nothing, through the fitted trees and the reranked index, lie between 0.5589 and 0.5770, a span smaller than the registered target, while correctness flips on 16.8% of questions between the DP index and flat dense (Appendix V). That calibrates the resolution at which a seed axis would have to act on the development-track diagnostic below, which propagates no axis to accuracy and predates the second registration's rebuilds (§5); it does not bound it. Varying only that covariance shape moves the partition by a comparable amount (ARI 0.24–0.56 against the full-covariance fit, and on one book a spherical shape drops the component count from 50 to 17). A registered assignment-rule ablation found the four multi-assignment thresholding rules effectively coincident on QuALITY's small trees, the registered one-line robustness note for the assignment rule. Neither axis was propagated to accuracy (summaries were not regenerated), so these are statements about the fitted object, not about the reported accuracy numbers; but they mean the specific partition our index commits to is one draw from a wide optimization/initialization stability ensemble, in the same qualitative direction as (though on different layers and data than) the posterior diagnosis of §6. On one of the 60 books, the *GMM-clustered baseline* tree's first-layer cluster count differs between the development-time and the production-time build; on the same book the DP-mixture tree's first-layer count is unchanged with a small membership reshuffle, and we treat count equality as no proof of reproducibility. Separately, on one perturbed development input, the global DP stage produced two distinct labelings within a single process in one of 54 measurement cells. The local reduction applied within each global cluster is not reliably reproducible either: on one of 50 real layer-0 subsets it reduced differently across processes, and on a further subset, from a development book, repeating the same call within one process returned different embeddings; the replay mismatches of Appendix K are not separated from this cause. *Domain.* The primary benchmark is fiction; the transfer check is reported cell by cell in Appendix M, unpooled and without a label, with its scope, its exclusion rule and the DetectiveQA protocol departure stated there. Between the short-document regime of prior work and our book-scale regime lies a length band our confirmatory benchmark does not cover; we make claims about neither regime's interior. *Posterior scope.* The stored confidence reflects the global clustering stage only (§3.3; Appendix K); the confidence-weighted product can reverse the intended confidence preference for negative-similarity candidates, which the deployed pipeline drops after the top-k slice (Appendix K, "The deployed ranking order"), a scoring boundary we state rather than resolve. And the stability diagnostic licenses exactly one conclusion, that MAP reassignment under sampled variational parameters moves the partition less than row-permuted refits do (§6), not the broader claim that the posterior fails to capture real-world uncertainty. We do not use the machine-precision axis as that yardstick: a zero-information row permutation is not separable from it by sign test (a bootstrap interval on the mean difference does exclude zero; both are reported in Appendix O), so what it measures is not specific to numerical representation (§6; Appendix O).

Retrieval stack. The frozen comparison included no sparse–dense fusion arm; the registration history records its exclusion as kept for the small-open-models scope wording and the cap, and the second registration later added one (Appendix L.1). Fusion answers a *retrieval-stack* question rather than the index-construction question this paper registers, would change the retriever for one arm, and the held-fixed retriever makes the index contrast attributable here. The lexical axis is bounded instead by a BM25 arm (Robertson & Zaragoza, 2009) over the same chunks, the same reader and the same token budget, carrying a registered duty to report in the foreground if it beats the dense floor. Any arm added after the freeze is descriptive and spends no α by construction (Appendix J). The released

harness carries a hybrid retriever class inherited from the codebase this work forks; it is constructed only in unit tests and no registered arm is wired to it.

Effect size and option position. The primary comparison was sized for a 3.4 pp difference on 60 books; effects of the size RAPTOR's own ablation reports for the clustering step (under one accuracy point) are not resolvable at that sample, and the abandonment verdict alone says nothing about them; the reported intervals do bound the difference from above, and the post-hoc equivalence read-out of §4 rules out an effect of the registered size against Leiden, though not against the mixture. A registered post-run check of option-position sensitivity, which re-scores the multiple-choice questions under a shuffled option order, was run after the primary analysis; both arms it covers lose accuracy under the shuffle (Appendix T). Every accuracy reported outside that appendix is on the benchmark's original option order.

What the powered-abandonment label rests on. That label is decided by a single realized dispersion, the between-book spread of the DP-minus-Leiden deltas, compared against a threshold frozen before the run. The threshold was deflated for the conjunction using a ratio fixed at design time from a *projected* spread ratio between the two conjuncts, not from the spread the second conjunct realized. Where that projection is wrong, the label can be reached on the first conjunct's spread while the second conjunct is noisier than the design assumed. A registered sensitivity analysis sweeps that projected ratio and moves the threshold by up to about 14.6%; the realized dispersion sits 2.8% inside the threshold (realized $s = 0.0813$ against the frozen $s^* = 0.0836$), well within that movement and well inside the $\sim 9.2\%$ estimation noise the registration records for that statistic. Scoring the same label on the second conjunct's own realized spread ($s = 0.0828$, 1.0% inside the threshold) returns the same label, so the case the projection leaves open did not arise here. Nothing in the confirmatory verdict depends on the label: it splits two cells that are both failures of the primary hypothesis, and the intervals, not the label, are what we read.

What the release does not let you re-derive. The release is the anonymized code mirror, the ledgers its manifest selects, and the files its extra manifest stages from outside the code repository. The evaluation arms of Table 1 and the registered contrasts over them are reaggregated from those by one released command; every other accuracy statistic is the same aggregation of the same released ledgers, condition by condition, except the reader probe named below and the second registration's reader run reported with it, and the index rebuild named after it, whose accuracy figures and intervals are copied from the analysis summaries the release ships, and the halted QuALITY attempt, whose accuracies are quoted from the registration that records it (Appendix J). Ten of the sixty-two quantity generators in the paper repository stop on a file the release does not carry: eight on the manuscript source, one on a macro file whose writer needs the cross-backbone combine transcript, which we do not release, and one on a macro file that the generator of the third category below stops before writing. One stops on a field the scrub removes from records that ship (the third category below), and what is left of this boundary is our own governance ledgers and the benchmarks' own snapshots, counted and described below. The reader-probe generator of Appendix W, which is post-hoc and exploratory, and on which no registered comparison, interval or verdict rests, is one of the eight; its run manifest, analysis summaries and spend ledgers now ship, while its per-query transcripts stay out of the release; no generator of a printed number opens them, and the files derived from them that those generators read ship. The generator of the post-hoc index rebuild of Appendix V.2 is not among the ten: its analysis summary, set timeline and spend ledgers ship, and it reads nothing else, while the rebuild's per-query records stay out of the release. The rest of this boundary closed earlier, and the same way. The development posterior-width probes and collapsed-mass diagnostic, the development overlap and overlap-split statistics, the truncation-binding and semantic-perturbation sweeps, the chapter-marker parse, the length-tercile, posterior-confidence and prewarm-recovery sidecars, the Study B responsibility-versus-cosine and representation-space runs, and the convergence-replay and build-recovery sidecars kept beside the manuscript rather than in the code repository now ship, so no generator stops on them any more, though some that read them stop on another file named in this paragraph; among those is the one behind the row-permuted refit mean quoted in the abstract, which opens an internal ledger for a check. The registration document in its original language ships as an anonymized copy during double-blind review, and the generators that read it on the release read its English rendering, never that copy; the rendering ships too (Appendix J). The printed claim register's size closed the same way: the identifiers it counts now ship as a list of their own, so the size re-derives, while the working ledger those identifiers are kept in stays out of the release and is read only to cross-check them. Four further generators stop on a filesystem restricted to the

release, because each opens an internal ledger of ours and those internal ledgers are not released; five of the eight that stop on the manuscript source open one too, and all nine are recorded as stopped rather than as complete, so none of them substitutes a degraded number for the one it could not read. For eight of the nine the ledger carries only a check that the numbers still match the entry that registered them, so what cannot be re-run is that check, while every digit they print stays recomputable from, or readable as recorded in, a released artefact; that the ledger read is a check and not a value source is stated per generator by hand, because the gate at the end of this section cannot tell the two apart. One of those eight reads the benchmark's own book metadata rather than any file of ours, so recomputable there means from the pinned public snapshot (Appendix B), which a reviewer fetches upstream rather than from us. Nine more generators read a benchmark snapshot without opening a ledger. Eight read that same snapshot: seven take only the question-to-book map from it, which the released question manifest also carries, and another counts each question's options; the remaining one reads a transfer benchmark's snapshot for a field no file of ours carries, the paragraph at which that benchmark's answer appears. So three generators in all print a number that resolves upstream rather than against a file of ours. For the ninth ledger-opening generator, the re-freeze line and edit counts reported in Appendix J, the ledger is where the generator takes them. Those digits are not recordless: the frozen protocol and the applier that edited it both still exist, and the generator reads both as cross-checks. They are kept in the surrounding research workspace rather than in the code repository, so they reach a reviewer no more than the two above do. Digests need one caveat: a digest that a released file records for another file of ours need not match the copy we ship. The scrub removes the benchmark's question text and answer-key field from the staged ledgers, and the anonymization rewrites identifying strings in the files it copies, so such a digest can be that of a pre-scrub or pre-anonymization original, or of an earlier version of the file. The second registration's pin of its spend-guard module is such a digest: it matches no committed version of that file, and the shipped module's spending cap differs from the constant the registration quotes from it, while equalling the cap the registration itself states. Digests over third-party benchmark files are untouched, and the digest list the release ships over its own bytes verifies. A third category holds one generator. It is for a generator that completes every read against a filesystem restricted to the release and still cannot reproduce what it prints, because a field the records once carried has been taken out of the copies we ship. The generator of Table 46 is not in it: that table prints each monitored row's identifiers, letter counts and first-versus-last flag and no completion text, so it rebuilds from the released copies although the scrub takes the completions out of them. The one generator this category holds prints the development-track rows of the second registration. Its spread columns come back from the released ledgers unchanged, because a spread over build seeds is a function of per-question correctness and that ships. Its answer-flip columns do not: a flip is an inequality between the option the reader chose at a build seed and the option it chose at another, the chosen option is among the withheld fields, and with it gone every pair is uncomparable, not unequal, so it stops rather than printing a flip rate counted over no comparable pairs at all. The budget-curve table's answer-change counts were in this category too, until the release began shipping, for each curve point, one boolean per question saying whether the answer differs from the registered run's. That boolean names no option, and it was the only input those counts needed that the release did not carry, so they now come back from it like the binding counts and token quantiles beside them. This boundary is measured rather than asserted, the snapshot reads and that per-generator classification above excepted: a gate runs every generator against a filesystem restricted to the release and to the macro files its generators rebuild from it, leaving the benchmark snapshots open, with the same fields taken out of the ledgers that we take out of the released copies, and fails if any of three sets changes in either direction, the generators that cannot complete, the generators that need an internal ledger, and the generators that complete and then report, by their exit status, that what they rebuilt no longer matches what they print. A generator that printed a degraded value without reporting it would not be in that third set, and no sweep of this shape could see it. Table 20 is outside that gate and does not re-derive from the release: it diffs revisions in the code repository's commit history, and the anonymized mirror keeps none of that history.

T REGISTERED OPTION-POSITION PROBE

The registration provides for a post-run check of option-position sensitivity: once the primary analysis is closed, the multiple-choice questions are re-scored with the options presented in a different order, which exposes a reader that keys on where an option sits rather than on what it says. The

check is registered as descriptive. It tests no hypothesis, spends no error budget, and its result enters no confirmatory verdict.

The probe re-ran two arms, the DP arm and the Leiden arm, over all 1,526 questions on the primary backbone, under a per-question derangement of the option order (seed 224): every option moves, so none keeps its original slot. The permutation is a function of the question identifier and the seed, not of which option is correct, so the presented-position distribution is the same for both arms; the gold answer occupies each of the four presented positions between 23.92% and 25.75% of the time.

Both arms lose accuracy under the shuffle. These accuracies are question-weighted, not the book means the registered comparisons use: the unshuffled values are the question-mean column of Table 1, so they are not comparable cell-for-cell with the book-mean cells of Table 15. The DP arm falls from 0.5675 to 0.5452, a drop of 2.23 points; the Leiden arm falls from 0.5623 to 0.5537, a drop of 0.85 points. Question by question, the reader's choice changes on 39.78% of questions in the DP arm and on 37.16% in the Leiden arm. The two conditions join on question identity with no question dropped on either side, and neither side carries a generation error. The released copies of both conditions withhold the reader's choice, as every released record does, so on the release both shift rates are read from the probe's shipped aggregate, which the generator accepts only after it agrees with the released records on every field that needs no choice; on the working copy they are recounted from the records and must equal that aggregate. The accuracies, the join and the error counts recount from the records on either copy.

Three limits travel with these numbers. A derangement moves every option, so these numbers are a sensitivity check under one fixed reordering rather than a bound on milder ones, and a small drop would not license a claim that no position effect exists. The two conditions are read together as a format-sensitivity statement about the reader; the difference between the two drops is not a registered comparison and we do not read one arm as more position-sensitive than the other from it. And the 10 two-option questions in the band admit only one alternative ordering, so for them this probe measures sensitivity to a single swap.

Every accuracy reported elsewhere in this paper is on the benchmark's original option order.

U PER-ARM RESOURCES

Figure 3 (§4) places construction cost beside question-answering accuracy for every arm that builds a summary tree of its own, with the flat dense reference at zero footprint. Its vertical axis is the 60-book book-mean accuracy of Table 21 (the chapter arm from Table 13) and does *not* start at zero; the window spans the arms actually observed. Panel (a) plots the stored summary-token footprint of Table 6(a); the flat dense arm builds no summary layer, so it has none by construction and sits on the axis (Table 34 prints that zero too). Panel (b) plots the index build wall-clock of Table 34, *primary backbone only*, the Leiden point including a tree the main run reused from the registered pre-freeze feasibility check, timed in that check (Appendix J); build wall-clock is a shared-machine condition and is not compared across backbones, so the robustness backbone's hours are not drawn, and the flat arm, whose index-build time was not measured, is omitted from (b) only. Neither axis alone is the index's cost, and the axes do not agree: against the Leiden index, which commits to a fifth as many summary nodes, the DP index stores 5.0× the summary tokens and takes 28.05 hours to build against 6.83, while occupying 789 MB on disk against 700 MB. Filled marks are the semantic family, the DP arm in colour; open circles are the non-semantic grouping controls (fixed-block contiguous, contiguous at DP capacity, chapter boundaries); the open square is the flat reference. What the picture is not: there are no error bars, because per-arm accuracy intervals are not registered, only paired contrasts are, so the separation or overlap of points is descriptive and settles no statistical comparison. The registered contrasts are the only warrant, and the DP arm's differences span zero against Leiden, against GMM+BIC, and against the fixed-contiguous control (-0.0005, $[-0.0255, +0.0245]$). The chapter arm sits highest in both panels, and in panel (b) at the shortest recorded build time of any tree arm (5.53 h, Table 34; on the books where it falls back to the contiguous construction its recorded builds are far shorter than the contiguous arm's, per the scope notes below), but chapter against DP is *not* a registered contrast: no interval is reported for it, the design was not powered for it, and its position is a point estimate. Statistical ordering is read from the paired intervals of Figure 2, not from these points. Two readings are not warranted: the chapter-boundary arm falls back to the plain contiguous construction on 24 of the 60 books (40.0%,

no chapter structure surviving both gates; Appendix E), so its distance from the contiguous point is attenuated by construction; and the horizontal spread is a property of the stored index under the registered component cap, not evidence about what a different cap would store.

Table 34 reports the cost axes for every arm of the first registration at its registered budget, with the closed-book and oracle-evidence references, on both backbones: total build wall-clock over the 60 books, serialized index size, total summary tokens, and end-to-end answer latency (retrieval + generation, per query). Scope and honesty notes, in order of importance. *Recovered build cells:* in the backbone robustness check, 4 arms (GMM+BIC; Leiden; Leiden, UMAP-matched; Fixed-block contiguous) were summarized from an LLM cache a separate warm session had already filled, so the time recorded inside the run measures a rebuild with generation served from cache, not a build. Those cells report the build wall-clock of the same trees as timed in that warm session; every other cell is timed inside its own run, except that the primary-backbone Leiden cell includes a tree the main run reused from the registered pre-freeze feasibility check, timed in that check. Where only a subset of an arm's books met the warm cache (Fixed-block contiguous (3 of 60 books)), the remaining books were built for real, and a control checks that the substituted set is exactly the set of books the recorded timings mark as cache-served. Nothing else in the table is substituted, and no accuracy figure is affected: the substitution matches trees by configuration hash and per-book build record, so the cells describe the same trees the run used. The chapter-boundary cells are not substituted either: on the 24 books where that control falls back to the contiguous construction, its recorded builds total 0.08 h and 0.07 h (primary and robustness backbone) against 3.82 h and 3.51 h for the same books in the contiguous arm. Both contiguous-arm totals are the recorded ones: on the robustness backbone the total includes one of these books at its cache-served recorded time. *Latency:* percentiles are over cache misses only (the method's cost, not the cache's), so the per-condition cache-hit rate travels with them in the table; they were measured while the backbone robustness check run shared the machine, so they are reference values, not a controlled measurement. *Missing cells:* a dash means no value (no build, a reused index, or not recorded), not zero: the reranking arm consumes the DP arm's tree and records no build of its own, and serialized sizes were not recorded for the derived rebuild arms. The summary-token multiplier quoted in the body and the build wall-clock multiplier below measure different axes of the same gap: $5.0\times$ is the summary-token ratio (DP over Leiden), $4.1\times$ the build wall-clock ratio (28.05 h over 6.83 h). Both denominators are the Leiden index, which commits to 4,855 summary nodes against the DP index's 24,882 (Table 6). Against the mixture the DP arm replaces, the same two axes read 6,205,729 over 4,822,747 and 28.05 h over 23.66 h. Matching the cluster count closes most of the token gap and reverses the wall-clock one. At the DP arm's own 24,882 summary nodes, the *GMM+BIC at DP capacity* row writes 5,884,945 summary tokens in 28.94 h and the *Contiguous at DP capacity* row (the count-matched twin of the fixed-contiguous non-semantic control, not that control itself) writes 6,178,512 in 30.42 h, against the DP arm's 6,205,729 in 28.05 h. Holding the count fixed therefore moves both axes by a small fraction of what changing the count moves them, and on wall-clock it moves them the other way: the multiplier tracks how many groups an index commits to far more closely than which rule chose them. We read it as a property of this comparison rather than of the mixture model. Both multipliers are ratios of totals over the 60 books every arm built, so they weight the books that took longest. Read per book on that wall-clock axis they move, and not in the same direction: the DP-over-Leiden multiplier rises to $5.4\times$ as a ratio of book medians and $5.1\times$ as the median of the per-book ratios, while the DP-over-GMM+BIC multiplier falls from $1.19\times$ to $1.13\times$ and $1.02\times$. The arms were built in one fixed order on a shared host (Leiden, then the DP mixture, then the UMAP-matched Leiden arm, then GMM+BIC), so a machine-level slowdown during the run is absorbed by the arms built after it. The arm effect and the time effect are therefore not separable in this design. We report both bases rather than choose one, and read none of these multipliers as a timing of the clustering rules alone. On the same wall-clock axis the backbone robustness check gives $4.6\times$ (25.33 h over 5.50 h; its numerator is timed inside that run and its denominator in the warm session, per the recovered cells above): the build-cost gap was observed under both writer/reader configurations. Observed p95 answer times differ across tree arms under these shared-host runs and are not controlled latency benchmarks; the flat floor is faster at the tail (p95 0.64 s vs. 1.21 s for the slowest tree arm).

Table 34: Per-arm resources, both backbones. Build hours are wall-clock totals over 60 books; latency is end-to-end (retrieval + generation) per query, percentiles over cache misses only, with the per-condition cache-hit rate alongside. "0" is a structural zero: the condition stores no summary layer of its own. "—" never means zero. In the summary-token column it marks the reranking rows, which retrieve from the DP arm's index and so have no footprint of their own; in the build and index columns it marks a condition that built no tree, an index reused from another arm rather than built, or an index whose size was not logged. In the lower panel the build hours for GMM+BIC; Leiden; Leiden, UMAP-matched; Fixed-block contiguous are timed in the separate warm session that had already filled the summary cache, not inside the run; in Fixed-block contiguous (3 of 60 books) only that subset met the cache and the rest of the arm is timed in the run. See the scope notes above.

Condition	Build (h)	Index (MB)	Summary tok (M)	p50 (s)	p95 (s)	Hit (%)
Primary backbone						
Flat dense	—	—	0	0.55	0.64	0.2
Flat BM25	—	—	0	0.56	0.68	0.0
GMM+BIC	23.66	764	4.82	0.61	1.21	13.4
Leiden	6.83	700	1.24	0.61	1.12	37.2
Leiden, UMAP-matched	6.25	703	1.40	0.61	1.15	35.8
Leiden, two-stage-UMAP matched	17.02	—	3.53	0.61	1.12	20.2
DP mixture	28.05	789	6.21	0.61	1.13	11.7
DP + confidence reranking	—	—	—	0.56	0.66	12.0
Fixed-block contiguous	8.08	706	1.60	0.59	1.13	34.0
Contiguous at DP capacity	30.42	—	6.18	0.61	1.14	13.6
GMM+BIC at DP capacity	28.94	—	5.88	0.61	1.18	11.9
Chapter-boundary control	5.53	—	1.85	0.58	1.08	53.8
Document-order repack	—	—	0	0.53	0.62	0.0
Closed book	—	—	0	0.38	0.48	0.0
Oracle evidence	—	—	0	0.38	0.48	0.0
Backbone robustness check						
Flat dense	—	—	0	0.53	0.67	0.1
Flat BM25	—	—	0	0.52	0.65	0.0
GMM+BIC	24.08	765	4.99	0.61	1.23	16.7
Leiden	5.50	700	1.24	0.54	1.09	42.0
Leiden, UMAP-matched	6.29	703	1.40	0.61	1.18	38.8
Leiden, two-stage-UMAP matched	15.56	—	3.55	0.61	1.10	24.9
DP mixture	25.33	790	6.31	0.57	0.84	15.5
DP + confidence reranking	—	—	—	0.61	0.72	11.2
Fixed-block contiguous	7.43	706	1.60	0.61	1.13	37.0
Contiguous at DP capacity	26.93	—	6.33	0.61	1.13	14.7
GMM+BIC at DP capacity	26.51	—	6.23	0.58	1.12	15.1
Chapter-boundary control	5.09	—	1.85	0.61	1.09	55.4
Document-order repack	—	—	0	0.61	0.72	0.0
Closed book	—	—	0	0.38	0.49	0.0
Oracle evidence	—	—	0	0.41	0.52	0.0

V Post-hoc exploratory read-outs

Computed after the confirmatory verdict and carrying no registered status: nine read-outs in the paragraphs below and a tenth in §V.1, which sharpen the anatomy of §4 and the diagnostics of §6 without touching any gate. Other post-hoc read-outs, of the outcome records or of other objects, appear elsewhere in this paper, and a post-hoc rebuild of the DP arm's trees from collapsed-Gibbs partitions closes this appendix (§V.2).

Re-reads of the registered contrasts without the normality assumption. The registered intervals are cluster-t intervals over 60 book-level deltas, and they assume the book means are approximately normal. Two summaries of *the same* 60 deltas that do not rest on that normality assumption (each carries its own: the bootstrap's resampling model, and exchangeability under sign flips) are declared here: a percentile bootstrap (10,000 resamples, level 0.95, seed 20260901, primary backbone only) and a sign-flip permutation p over the same resample count. Nothing is re-measured, so every point estimate is identical to the registered one: these are three sensitivity analyses of the same outcomes, not three independent tests, and no contrast gains evidence by appearing in all three.

On the primary contrast the bootstrap interval is $[-0.0107, +0.0299]$ with $p = 0.37326$, against the registered $[-0.0114, +0.0306]$; for SC4, $[-0.0040, +0.0364]$ with $p = 0.13239$; for the descriptive reranking comparison SC3 (§6), $[-0.0335, -0.0028]$ with $p = 0.02790$; for the anti-trivial floor SC0, $[+0.1460, +\infty]$ with $p = 0.00010$. Across all 11 contrasts 0 verdicts move: on the 10 contrasts checked against their registered interval, every zero-exclusion agrees. The permutation p has a floor of 0.00010 at this resample count, so a p at the floor means no permutation was more extreme rather than an inequality. What this licenses is narrow: the registered verdicts do not rest on the normality assumption. It is not further evidence for any of them.

Question-level churn. The near-zero book-level deltas between the paired arms below are not near-zero at the question level. Swapping the index changes the outcome on more than a hundred questions in every pair: DP vs. Leiden fixes 124 questions and breaks 116 (churn 15.7% of 1,526); DP vs. flat dense 133/124 (16.8%); DP vs. GMM+BIC 113/91 (13.4%); DP vs. fixed-block contiguous control 107/115 (14.5%); Leiden vs. flat dense 77/76 (10.0%); DP vs. UMAP-matched Leiden 121/104 (14.7%). The trees change *which* questions are answered correctly far more than *how many*; the fixed and broken sets nearly cancel. No test is attached to these counts (within-book correlation makes a question-level test anti-conservative); they are declared as counts. Figure 7 draws every pair.

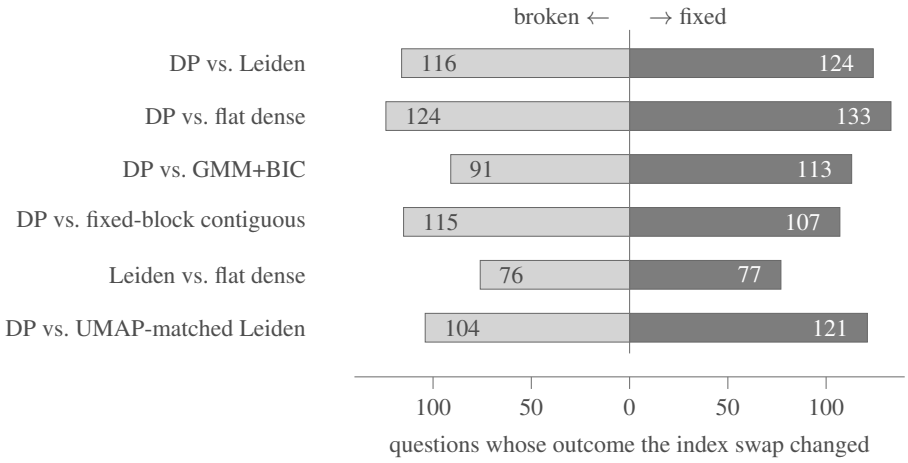

Figure 7: Question-level fix and break counts for every index pair the churn read-out of Appendix V reports, drawn as opposed bars in the order that read-out reports them. *What it shows.* Swapping the index moves the outcome on more than a hundred questions in every pair while the opposed directions very nearly cancel, so the near-zero registered book-level deltas between those arms sit on top of a large amount of question-level movement: the index changes *which* questions are answered correctly far more than *how many*. *What it is not.* These are counts over the pooled question set of 1,526 questions, whereas the registered deltas are equal-book means over books, so the aggregations weight books differently and carry different units; a small question-level net and a book-level delta of another size are not in conflict, and neither is evidence about the other. No test is attached and no interval is drawn: within-book correlation makes a question-level test anti-conservative, so no gap between bars here is a significance claim, and the bars are not a vote. The pairs drawn are the reported set, not a selection. Every count is a quantity of that read-out; the figure adds none of its own.

Structure–outcome correlations, all cells. Over the declared fixed set of two layer-0 structure diagnostics (K_{eff} and mean maximum responsibility) crossed with three per-book deltas (DP−Leiden, DP−flat, DP−contiguous control), no cell's 95% interval excludes zero. The largest magnitude is K_{eff} against the DP−Leiden delta: $r = -0.190$ ($[-0.423, +0.067]$); the remaining five cells are $r = -0.131$, $+0.016$, $+0.102$, $+0.121$, and $+0.033$, each with an interval covering zero. None of the six tested book-level correlations has a 95% interval excluding zero; the test says nothing about other predictors, non-linear relations, or query-conditioned interactions.

Posterior-confidence saturation. The build-time confidence the reranking arm consumes (§6) is one minus the normalized entropy of the node's responsibility row (Equation 1; c_i=1 at a one-hot row), and it is compressed toward one: over the 195,837 layer-0 nodes of the 60 books it has median 0.998, fifth percentile 0.835, smallest value 0.525, and 89.2% of nodes above 0.9 (66.6% above 0.99); the least concentrated book still has median 0.978 and the most concentrated 1.000. So the reranker multiplies in a signal with little spread, though not a constant, and that is a statement about NovelQA, not a dataset-independent property of the object. The audit behind finding (iii) covers assignment confidence across its uses, and this is the range it had to work with; the read-out is post hoc and was computed after the stage-2 verdict, so it qualifies what the audit could resolve and is not offered as a reason for any registered outcome. Table 35 carries the same read-out layer by layer, recovered post hoc from the stored sidecars: each row is that layer's global fit, one row per node, and the local fits that re-cluster the same nodes are excluded. The layer-0 row reproduces the read-out just given field for field, and this licenses the correspondence between a stored fit and a tree layer; it says nothing about coverage, which is partial. Saturation deepens with height rather than easing: the share above 0.99 rises from 66.6% on layer-0 nodes to 85.6% and then 96.0% on the layers above, whose median is 0.99997 already, printed rounded in the table. A confidence is stored on a node only where its layer was clustered into another, so the top layer of every book carries none: the table speaks for 220,433 of the tree's 220,719 nodes, and the remaining 286 are not dropped but enter the reranker at a neutral 1. That is also why the last row is thin. Only 4 of the 60 books build a layer above layer 3, so of that layer's 355 nodes only 80 carry a confidence; the row is reported for completeness, not as a corpus statement. Post hoc and exploratory: no registered contrast covers these rows and they spend no error budget. The book-level structure diagnostic is a different quantity: over the 60 books, layer-0 mean maximum responsibility spans 0.9067–0.9938 with median 0.9782, and it tracks log book length at $r = -0.894$: the diagnostic is effectively a length proxy. The companion diagnostic is partially censored: K_{eff} sits at its truncation bound of 50 in 21 of 60 books (range 31–50, median 49). On the confirmatory corpus, then, layer-0 confidence is strongly concentrated near one. The registered product reranker did not improve accuracy (§6), but these marginal distributions and book-level correlations do not establish that responsibility-derived features carry no useful information for other consumers, and they are not a confirmatory-corpus replication of the development-sample diagnostic of §6.

Table 35: Stored per-node cluster confidence by tree layer, on the 60 confirmatory books of the primary backbone. Each row is that layer's global variational fit, one row per node; the local fits re-cluster nodes the global fit already placed and are excluded, which is why the counts fall with height rather than partitioning a fixed total. The layer-0 row is the read-out of §V recovered from the sidecars and agrees with it field for field. The registration attaches, to statements citing sidecar figures, the fidelity of parent membership rebuilt from the sidecars, measured on one NovelQA book's layer refits: symmetric difference 0 on the top three layers and 0.3665% (258 of 70,389) at layer 0; these rows rebuild no parent membership and place each fit on its layer by the sidecar's own manifest. Post hoc and exploratory, no registered status, no error rate.

Layer	books	nodes	mean	p5	median	min	> 0.99
0	60	195,837	0.972	0.835	0.998	0.525	66.6%
1	60	21,796	0.992	0.949	1.000	0.663	85.6%
2	60	2,720	0.998	0.993	1.000	0.687	96.0%
3	4	80	0.999	0.999	1.000	0.911	98.8%

How far confidence reranking moves the list. The post-hoc intervention rates of §6 (how often the reranker changes the retrieved node set and the final answer) say that the retrieved list changed, not by how much, and "the order changed" is equally true of a single adjacent swap. Read post hoc off the same two sets of per-query records, over the 20-node lists the arms return for each of the 1,526 questions: the two arms share 18.1 nodes per question on average, a shared node moves 1.21 places on average (median 1, ninetieth percentile 3, largest 18), the rank correlation between the two orderings of those shared nodes averages $\tau_a = 0.888$ and never falls below 0.283, the top-ranked node differs on 12.0% of questions, and 89.7% of the DP arm's top 5 survives inside the reranked top 5. Re-ranking each arm over the shared nodes alone, which takes out the shift a dropped node imposes on everything below it, leaves the mean displacement at 0.79 with 58.7% of shared nodes stationary: part of the apparent movement is churn in the retrieved set rather than reordering, and both readings are reported because what reaches the reader is packed from the full list, not from the

shared nodes alone. Post hoc and exploratory: the registered contrast set carries no row for these, and they spend no error budget.

Retrieved-context overlap between arms. Do the arms hand the reader the same evidence? Two definitions are declared together and are not to be cited apart: (A) *leaf-only*, the Jaccard overlap of the leaf chunks retrieved directly, and (B) *leaf-coverage*, the same overlap after every retrieved summary node is expanded to the leaves it summarizes. Over the 1,526 questions and 60 books (primary backbone only), measured on the nodes actually placed in the prompt (a question on which neither arm retrieved a leaf directly is undefined under (A) and dropped; at most 8 per pair), the DP-mixture and Leiden arms overlap (book-mean Jaccard) at 0.7124 under (A) but 0.3754 under (B); DP against flat dense at 0.6395/0.3905; DP against the contiguous control at 0.7479/0.3298. Every pair is in the artifact; across all of them (A) spans 0.6395–0.9278 and (B) 0.3298–0.7044; the closest pair under either definition is the contiguous control against the chapter index (0.9278/0.7044). The definitions disagree because a summary node stands in for many leaves: per question the DP-mixture context covers 112.6 leaves against 12.2 retrieved directly, Leiden 155.4 against 16.7, the contiguous control 192.6, the chapter index 208.7; over these four arms the covered mass is 9.3–14.1 times the directly retrieved mass, all of it invisible to (A). Under (A) the arms' leaf sets are nested prefixes of one shared ranking (0 subset violations in 15,097 comparisons; a comparison is skipped when either arm retrieved no leaf directly): collapsed retrieval scores leaves identically in the five arms measured here (none of them reranks by confidence), and a tree only inserts summaries that push leaves out of the top-k. That nesting is not a finding about evidence; it is the reason (A) is uninformative. Under (B) the nesting fails in 72.9% of comparisons (11,127 of 15,260). The clustering rule therefore changes which source passages the packed context descends from, through what the summaries represent rather than through which leaves are retrieved; the closure identifies source ancestry and does not establish which evidence survived summarization. Why near-identical accuracies follow from substantially different evidence is not in this statistic, and we do not supply a mechanism. The nesting under (A) is a property of collapsed retrieval and does not transfer to tree traversal.

Does the truncation bound bind? The K_{eff} ceiling above (the 21 books pinned at 50) asks whether the registered truncation $T = 50$ caps the fitted partition or merely sits above it. On the longest confirmatory book (34,011 leaves; also the pilots' development book) we refit layer 0 at $T \in \{10, 25, 50, 100\}$ under the same seed and the same reduced array, with no LLM call (primary backbone only). The global fit saturates the truncation it is actually fitted at, up to the registered one: $K_{\text{eff}} = 50$ at $T = 50$, and it first falls short at $T = 100$ ($K_{\text{eff}} = 94$). The registered bound binds; the first tested truncation that is not fully occupied is $T = 100$. Two readings these numbers invite are wrong. At $T = 10$ and $T = 25$ the fit did not converge at its nominal truncation and was refit one rung down the retry ladder (at 5 and 12, where it again saturated): the model did not choose a smaller K, it failed at the larger one. And the layer-0 saturation count is 25 at $T = 50$ and 25 at $T = 100$; the share falls from 0.490 to 0.263 only because a finer global partition spawns more local fits (51 against 95 fits in all), a dilution, not a reduction, so the share is not to be cited alone. Only the global fit is paired across caps; local fits are re-derived from a cap-dependent partition, so the layer-0 aggregates compare populations, not matched fits. A single book and layer 0 only: deeper layers take summaries as input, and a faithful refit there would need LLM calls.

A semantic perturbation beside the numerical one. The perturbation axis of §6 is numerical and was measured on the development books of a different corpus at the sweep's own parameterization. Here the same kind of ladder is run on the longest confirmatory book of the cap sweep above (primary backbone only), at the registered clustering configuration (so its noise numbers are a different parameterization and are not comparable to those of §6), with the reduced array recomputed rather than shared (determinism floor between two identical runs: ARI 0.9987 raw, 0.9940 on singly assigned leaves). Beside a noise ladder at 10^{-9}, 10^{-7} and 10^{-5} we place one semantic change: the registered encoder is swapped for `intfloat/e5-base-v2`. The layer-1 clustering is overlapping (in this cell's base refit 57.5% of the book's 34,011 leaves are multiply assigned, and the frozen tree carries 2.07 assignments per leaf on average), so the floor, ladder and encoder-swap ARIs are each given twice (raw, over first-cluster-wins hardened labels, and restricted to leaves singly assigned in both partitions), and the base refit's first-cluster-wins labels agree with the shipped layer's last-cluster-wins hardening only at ARI 0.2732. The noise ladder is flat and non-monotone, raw 0.1838, 0.2335, 0.2267 (single-assignment 0.5862, 0.5589, 0.6074), with neighbourhoods preserved

(retention ≥ 0.9911): a step at zero, not a dose–response, and it lies in the regime of the seed-only spread on this book (raw ARI 0.1949–0.2186 across 3 seeds on byte-identical input, as read from the seed reference that the released perturbation file carries; the source file that reference names is not released). The encoder swap changes the geometry (retention 0.1366) and lands below every noise rung in both columns: raw 0.0529, and single-assignment 0.2429 on 6,707 leaves against 9,215–9,620 for the noise rungs: different populations, so the single-assignment column is read for direction only, not for size. What this licenses is one sentence: a semantic change is distinguishable from a numerical one on this axis, which is consistent with, and weaker than, the route by which §6 already rejects the arithmetic-artifact reading; the flat ladder and the seed comparison are likewise re-confirmations at the registered parameters, not new findings. It does not license reading the axis as input sensitivity, and a single book, a single alternative encoder and layer 0 bound nothing on their own.

Does a window carrying more summary text answer worse? Every tree arm puts summary nodes into the reader's window and the arms differ systematically in how many, so the question costs nothing to ask of the frozen records: the retrieval step logs each retrieved node's level and the scored record carries the outcome (primary backbone only). The 6 tree arms analyzed here are those of Table 7 plus the fixed-block contiguous control, which that table does not carry; the manuscript's other arms are outside this read-out.

Pooled over questions, accuracy in the highest summary-share bin sits below the zero-share bin in 5 of the 6 arms, but the pattern is weaker than that sounds: only 1 of them declines across every bin, the top bins hold 15–169 questions, and for the global-UMAP-stage-matched Leiden control the top bin sits above the zero bin. The pooled bins are confounded by book, and the confounding is measurable rather than assumed: correlating each book's mean summary share with that book's accuracy, a point per book, gives -0.1959 to -0.4767 across them (Table 36, between-book column).

Within books the association is smaller in every arm, by an order of magnitude or more in all but the GMM+BIC arm. Every arm's point estimate is negative, so the direction does survive, but only the GMM+BIC arm's interval excludes zero (-0.0798, $[-0.1486, -0.0109]$), and one exclusion among 6 95% intervals over the same corpus is unremarkable arithmetic (the intervals are not independent). Books carrying fewer than 10 scored questions, and books in which either the summary share or the outcome does not vary, cannot carry a correlation and are dropped from the 60: across arms that costs 1 book to the first and 0–1 to the second.

Permuting correctness within each book, which destroys any within-book association while leaving the book structure intact, flattens the book mean. Averaged over 200 permutations within an arm, its absolute value runs 0.0224–0.0246 across the 6 arms; the spread is on the signed quantity, whose per-arm standard deviation averages 0.0293. This is a control on the metric, not a claim about the effect. The last column of Table 36 counts, for each arm, the permutations whose book mean was at least as negative as the observed one; the counts are reported because they were computed, and no test is attached to them. We read nothing mechanistic from any of this. An earlier version of this read-out grouped questions by identifier prefix, which manufactured pseudo-books and returned a uniformly negative correlation in every arm, which is why the book mapping is pinned by a control rather than assumed.

Table 36: Summary share against correctness, post hoc and exploratory, primary backbone only. Within-book: the per-book correlation between a question's summary share and its correctness, averaged over books, with a cluster-t interval. Between-book: a point per book, its mean summary share against its accuracy. Permutations: of 200 within-book permutations, how many gave a book mean at least as negative as the observed one, a property of the metric, carrying no test. The $r < 0$ column is counted on the exact rational covariance rather than on the floating-point correlation: a book whose within-book covariance is exactly zero has no sign, but the value a floating-point sum reports for it does, and which side of zero it lands on depends on the summation order. Such books occur in this run.

Arm	Within-book r [95% CI]	Books	Books $r < 0$	Between-book r	Perm.
GMM+BIC	-0.0798 $[-0.1486, -0.0109]$	58	39	-0.3416	1
Leiden	-0.0004 $[-0.0646, +0.0639]$	59	32	-0.3590	96
Leiden, UMAP-matched	-0.0002 $[-0.0631, +0.0627]$	58	30	-0.3572	93
DP mixture	-0.0300 $[-0.0897, +0.0297]$	59	32	-0.4146	25
Confidence-reranked	-0.0204 $[-0.0836, +0.0428]$	58	31	-0.4767	57
Fixed-block contiguous	-0.0132 $[-0.0749, +0.0485]$	58	28	-0.1959	63

V.1 The diagnostic variation axes, re-measured on the confirmatory corpus (post hoc, no registered status)

The diagnostic of Appendix O was run on a development sample, and its scope stops there: no claim in this paper about the confirmatory corpus rests on it. After that result was in hand we ran the same procedure over the confirmatory corpus, and we report the outcome here because a reader who accepts the development reading will ask whether it survives the larger corpus. The extension carries no registered status: it was computed after the confirmatory verdict was known, no registration analyses it (the second registration's deviation record cites it as prior information for its test H7a), and nothing outside this appendix depends on it.

What was measured. Every book of the confirmatory corpus: 60 books, 195,837 leaves, none excluded and none lost mid-run. The axes are the ones Table 4 defines, computed by the same routine: MAP reassignment under parameters sampled from the fitted variational posterior, against a refit after an input perturbation of scale 1.0×10^{-7}. The comparison is paired within a book, and the sign test is over books. The procedure's own positive control, an untreated refit that must reproduce the recorded labeling, did so exactly on all 60 books, the largest of them at 34,011 rows. The comparison uses rows A and B′ of Table 4, and the two axes are not symmetric in what they recompute: row A re-derives the reduction from the perturbed input and then refits, while row B′ relabels on the reduction the recorded fit already used. The gap therefore prices the reduction and the fit together and does not isolate either; it is not the B′-against-P contrast of §6, whose row P holds the reduced embeddings fixed. The recorded fits that both axes are compared with, and the posterior that the within-fit axis samples, come from the confirmatory run's build sidecars. The registration attaches, to statements citing sidecar figures, the fidelity of parent membership rebuilt from the sidecars, measured on one NovelQA book's layer refits: symmetric difference 0 on the top three layers and 0.3665% (258 of 70,389) at layer 0; both axes stay within layer 0's global fit and rebuild no parent membership.

What the confirmatory corpus gives. The within-fit axis stays nearer the recorded labeling than the perturbation axis does, by +0.2859 $[+0.2724, +0.2994]$ (book-resampling bootstrap), on 60/60 books, sign test $p = 1.7 \times 10^{-18}$. On the development sample the same quantity was +0.2996 $[+0.2823, +0.3161]$ on 18/18 books. Those intervals overlap, which is an observation and not a test: we did not test the difference between the samples, and no error rate was set aside for anything in this section. The perturbation axis itself sits at ARI 0.6840 here against 0.6672 in development. The component-count stratification points the same way in both, and we say no more than that: the development run put the gap between draws whose component count moved and draws whose did not at +0.0106, and here it is +0.0206, over the 78 of 120 draws whose count moved. The unit there is the draw and not the book: each book contributes one observation per perturbation draw.

The size covariate, and why we read nothing from it. The correlation between a book's leaf count and its perturbation-axis ARI is +0.3008 on the development sample and -0.3887 ($p = 0.002$)

here, and a sign change of that size invites a reading about larger books. It does not support one, for three reasons we state together because any one of them alone would be easy to discount. The rank form of the same correlation is near zero on both samples: Spearman +0.0939 ($p = 0.711$) in development and +0.0131 ($p = 0.921$) here. The development coefficient itself is +0.3008 at $p = 0.225$, so nothing was established there for this corpus to fail to reproduce. And the coefficient here rests on one book: the corpus runs from 1,011 leaves to 34,011, with 9,510 for the second largest, so that book is the only observation in the upper part of the range and the development sample held nothing comparable to it (its own largest is 6,395 leaves). Removing it alone moves the coefficient to +0.0448 ($p = 0.736$). The paired difference this section reports does not depend on that book either: without it the difference is +0.2828 against +0.2859 with it. We print the Pearson coefficient anyway, beside the three quantities that bound it, because printing it alone is what would mislead.

V.2 The DP arm's trees rebuilt from collapsed-Gibbs partitions (post hoc, no registered status)

Status. This rebuild is **post-hoc and exploratory, outside the registered analysis**: the registered analysis program does not read its records, which are kept apart from the registered run's records, and it was run after the first registration's confirmatory analysis had closed. It carries no registered status and no error rate, no significance level was spent on it, its records are never pooled with the registered records, and nothing in the body rests on it. No new model call was made for this section: every number in it comes from the aggregates the rebuild's own analysis program wrote, from its set timeline and from its spend ledgers.

What was built. The DP arm's trees on the primary backbone were rebuilt from partitions drawn by collapsed Gibbs sampling instead of from its truncated variational fit, and the registered question answering was run again on every rebuilt tree, over the 1,526 questions and 60 books of the registered band. Each rebuild follows a chain of its own, the chains started from different seeds. For each fit, the partition kept is the retained snapshot of that chain with the largest collapsed joint probability; every node's responsibilities are then recomputed under that partition's clusters, and the overlapping cover and the confidence score are derived from them by the arm's own rules. The tree scaffold, the summary prompt, the summarizer, the retrieval and the reader are the registered ones. The rebuild therefore differs from the frozen index in its construction, which is more than its inference algorithm: the sampler's normal–inverse-Wishart base has more degrees of freedom than the variational fit's prior and a small ridge on its scale matrix, the sampler has no truncation, the partition is selected by a different rule, and the cover is derived from different responsibilities. It also differs in epoch: every tree was rebuilt later, its summaries were regenerated by the same summarizer model, and no model output was reused from the frozen run's cache. No contrast here isolates the error of the variational approximation.

A stopping rule the sampling run did not implement. Each book's first partition in each chain, the global fit at the lowest layer, is a re-derived draw of an earlier collapsed-Gibbs sampling run over the same books, checked against that run's record before the rebuild began. The specification of that earlier run fixed a stopping rule before it: time the largest single fit alone at reduced sweeps, extrapolate its full cost, and stop and re-plan if that cost exceeded the power-law prediction by more than a factor of two. The sampling driver never implemented the rule, and the run started without that timing. Extrapolations of that fit's cost from the fits completed while the run was in progress put it at about twice the prediction, at the rule's threshold; the rule's own measurement, a timing of that fit at reduced sweeps, was never made, so whether the rule would have stopped the run is not known. Once completed, the fit cost slightly less than the prediction, which shows that those extrapolations were pessimistic, not that the rule was met. The rule concerned cost alone and read no outcome.

What ran, and what did not. Before the run the design fixed four rebuilt sets: one for each chain; a cluster-count-matched contiguous grouping of the first chain's trees, which keeps that chain's per-layer cluster counts and drops its semantic grouping (the capacity control); and a variational rebuild of the DP arm made in the same epoch as the chains (the epoch control, optional in the specification and chosen before the launch). Of these, two ran, those of the chains; the contiguous twin and the same-epoch variational rebuild did not run: the driver's per-set halt, fixed before the run, stops the process after a set that ends with a build-failed book and starts no further set, and the second chain's set ended with one. Table 37 prints every contrast the specification names, in its order, with the

analysis program's own status word in place of numbers wherever a set it needs did not run. Its last two rows, the direct contrasts against the same-epoch variational rebuild, and the form of the reading that uses them entered the specification after the run had started, while the first chain's set was being built, in a revision that also renamed what a rebuild changes from its inference algorithm to its construction and tightened the wording of the other readings. No record shows a per-query result read before that revision; access that leaves no record, such as a file opened by hand, cannot be excluded.

Table 37: The DP arm's trees rebuilt from collapsed-Gibbs partitions: the contrasts fixed before the run, in their fixed order, and the readings built on them. A row label reads *condition − reference*, and the sign follows that order; *first chain* and *second chain* name the rebuilds from the first and the second Gibbs chain, and *frozen DP index* the registered DP arm. A numeric cell is a book-paired accuracy difference over 59 books with its descriptive 95% cluster-t interval; a cell holding `MISSING` or `NOT RUN`, the analysis program's own status words, belongs to a row that needs a set that did not run. The rows above the rule are the contrasts the design fixed before the run; those below it are the contrasts the readings add, and the last two of these entered the specification after the run had started. Post-hoc and exploratory, outside the registered analysis, no error rate, never pooled with the registered records.

Contrast	Difference [interval]	Against zero
First chain − frozen DP index	−0.023 [−0.044, −0.002]	excludes zero
Second chain − frozen DP index	−0.020 [−0.040, −0.000]	excludes zero
First chain − second chain (placebo axis)	−0.003 [−0.021, +0.015]	covers zero
First chain − its contiguous twin (capacity)	`MISSING`	—
Difference in differences: the capacity row net of the frozen DP index's margin over its own contiguous twin	`MISSING`	—
Same-epoch variational rebuild − frozen DP index (epoch)	`NOT RUN`	—
First chain − same-epoch variational rebuild	`NOT RUN`	—
Second chain − same-epoch variational rebuild	`NOT RUN`	—

How the rows are read. The reading chosen before the run is the one that uses the same-epoch variational rebuild. That rebuild did not run, so neither the direct contrasts against it nor the epoch contrast exists, and the rows are read under the weaker reading, also fixed before the run, for the case without it. A chain-against-frozen row is read only as whether a Gibbs rebuild differs from the frozen build, never as an effect of the partition, because it carries the construction and the epoch together. Each such row is read beside the chain-against-chain row, the placebo axis, which compares the chains with each other with the construction and the epoch shared: to the extent the chains differ, a chain-against-frozen difference is uninterpretable by as much. The weaker reading's own same-epoch contrasts, the capacity contrast and the difference in differences, need the contiguous twin, and they are missing as well. An interval that covers zero is not distinguished at this n: it is neither an equivalence claim nor evidence that the constructions agree.

The excluded book, and the refusal. In the second chain the tree of one book failed to build. That book is excluded from every contrast on both sides, so every row with an interval is computed over 59 of the 60 books, and the second chain holds 1,495 records against the first chain's 1,526. The rebuild's analysis program ended with one refusal, and its output marks every row as printed for inspection only. The refusal concerns that book: a check requires every band book of a completed set to carry a record of a fresh tree build, and a book whose build failed has none. The generator of Table 37 accepts that refusal and no other, and it stops if any refusal names a book other than those that failed to build and were excluded.

Accounting. The rebuild driver's own per-book ledger and the launcher's pod ledger both bill this rebuild, and they are not the same number: the driver's ledger accrues $105.28 over the sets that ran, and the pod ledger, which bills the pod from boot to stop, records $110.06 for the pod time it logged. Summed over the sets that ran the wall clock is 62.29 hours, and the first pod start and the last pod stop are 65.12 hours apart; that window also holds model load and every stretch in which the pod was up while no set was running. Both ledgers ship with the release.

What this does not license. Nothing here settles a registered comparison or changes a registered verdict, and nothing here identifies a mechanism. Without the same-epoch rebuild no row separates the construction from the epoch; without the contiguous twin no row nets out a change in how many clusters each layer holds; and because the priors, the truncation and the selection rule differ, no row measures the error of the variational approximation. The intervals are unadjusted and descriptive, with no familywise error control.

W READER PROBE: THE FROZEN INDEXES READ BY A LARGER READER

Status. This read-out is **post hoc and exploratory**. It was run after the registered analysis closed, and the driver raised no refusal and no warning over it; it carries no registered status and no error rate, no significance level was spent on it, and its records are never pooled with the confirmatory records. No verdict in this paper rests on it and nothing here reopens a registered comparison.

What was run. The frozen primary-backbone indexes were read again by a same-generation, larger reader (`meta-llama/Meta-Llama-3-70B-Instruct`) over the 1,526 questions and 60 books of the confirmatory band, across 30 conditions, each named in Table 38: the confirmatory arms, the closed-book and oracle references, the registered companion conditions and points from the budget sweep of Appendix G. 28 of them read one of the frozen indexes; the other 2 are the closed-book condition, which reads no document, and the oracle, which reads gold evidence. A condition named without a budget is at the registered reader-token budget, the main-run column of Table 15; the largest sweep point (7,400 tokens) also retrieves top-40 rather than top-20, so those rows are not comparable to the rest on budget alone. No tree was rebuilt and no summary was rewritten; the summaries in every tree remain the frozen `Llama-3-8B-Instruct` summarizer's, so what changes between the two accuracy columns below is which weights answered. Prompt identity is measured rather than assumed: before the paid run, 750 prompts (25 per condition) were looked up in the frozen run's reader cache and each was already a key of it, while deliberately perturbed prompts missed. That lookup fixes the raw string. The wrapper around it is the same too, by a separate check made before this run rather than by this run's own artefacts: the published tokenizer configurations of the frozen checkpoint and of the probe checkpoint carry the same chat template, so the readers differ in their weights and not in what they were handed. The artefact records 0 symmetric book exclusions. The evaluation band is the confirmatory question set itself (the same books and questions, not the registered analysis), and the generator refuses to write this appendix unless the frozen column reproduces this paper's own question-pooled accuracy for each of the 7 conditions for which the body prints that same statistic, so the two columns are read on the same questions, the same books and the same records. Accuracies the body prints for the remaining conditions are book means, a different statistic, so they are not the digits this column reproduces and will not match it. Of the 30 frozen record files behind that column, 15 are pinned to the registered run by content hash; the budget-sweep files are matched by path and record count, not by hash.

The floor first. Closed-book accuracy is 0.478 under the probe reader against 0.389 under the frozen reader, a book-paired difference of +0.078 [+0.036, +0.121]. That difference is the context in which everything else here is read: the larger reader answers more of this benchmark with no document at all. Every contrast below is nevertheless computed on the full paired evaluation set, including the questions the closed-book condition answers correctly. No residual subset is analysed: the questions the readers miss are not the same set, so a residual subset would not be the same population under each of them.

How to read these tables. An interval covering zero is *not distinguished at this n*: it is neither an equivalence claim nor evidence that the indexes agree. The contrasts §7 names are read here with their intervals under both readers. The DP arm over Leiden reads +0.010 [−0.011, +0.031] under the frozen reader and +0.001 [−0.019, +0.020] under the probe; the DP arm over flat dense reads +0.006 [−0.019, +0.031] and −0.009 [−0.031, +0.013]; the DP arm over the closed-book floor reads +0.193 [+0.143, +0.242] and +0.113 [+0.070, +0.157]. All three are rows of Table 39 and are read by the rule just stated; whether the reader-to-reader difference is itself distinguished is a separate question, answered by the gap-shift column below and by no other line here. Of the 13 contrasts, 10 have a probe-reader interval covering zero, and 2 change sign between the readers, each carrying * with both of its intervals printed so that the sign change is read against them rather than on its own. They are DP mixture minus Document-order repack (−0.003 under the probe reader, +0.015

Table 38: Reader gain per condition: the same frozen index read by the probe reader and by the frozen reader. The accuracy columns pool over the band; the last column is the book-paired difference (probe minus frozen) with its descriptive 95% interval. Post hoc and exploratory, no registered status, no error rate, never pooled with the confirmatory records.

Condition	probe reader	frozen reader	book-mean paired Δ [95% CI]
Closed book	0.478	0.389	+0.078 [+0.036, +0.121]
Flat dense	0.587	0.562	+0.014 [−0.018, +0.046]
Flat BM25	0.559	0.548	+0.010 [−0.020, +0.040]
Leiden	0.579	0.562	+0.008 [−0.022, +0.039]
DP mixture	0.575	0.567	−0.001 [−0.034, +0.033]
Leiden, UMAP-matched	0.587	0.556	+0.017 [−0.016, +0.050]
GMM+BIC	0.567	0.553	+0.007 [−0.021, +0.035]
Fixed-block contiguous	0.596	0.573	+0.010 [−0.026, +0.045]
Confidence-reranked	0.564	0.551	+0.006 [−0.022, +0.035]
Oracle evidence	0.775	0.750	+0.012 [−0.007, +0.031]
Document-order repack	0.579	0.560	+0.018 [−0.018, +0.053]
Chapter-boundary	0.599	0.585	+0.010 [−0.023, +0.042]
Leiden, two-stage-UMAP matched	0.564	0.555	−0.002 [−0.034, +0.031]
Contiguous at DP capacity	0.579	0.574	−0.000 [−0.035, +0.035]
GMM+BIC at DP capacity	0.562	0.550	+0.002 [−0.030, +0.034]
Flat dense at 1,000 tokens	0.559	0.553	−0.007 [−0.042, +0.028]
Leiden at 1,000 tokens	0.539	0.546	−0.011 [−0.048, +0.025]
Leiden, UMAP-matched at 1,000 tokens	0.541	0.550	−0.018 [−0.055, +0.019]
DP mixture at 1,000 tokens	0.533	0.550	−0.015 [−0.047, +0.016]
Fixed-block contiguous at 1,000 tokens	0.549	0.560	−0.022 [−0.056, +0.012]
Flat dense at 4,000 tokens	0.587	0.562	+0.014 [−0.018, +0.046]
Leiden at 4,000 tokens	0.588	0.564	+0.015 [−0.017, +0.046]
Leiden, UMAP-matched at 4,000 tokens	0.588	0.557	+0.015 [−0.017, +0.046]
DP mixture at 4,000 tokens	0.598	0.558	+0.026 [−0.009, +0.061]
Fixed-block contiguous at 4,000 tokens	0.607	0.577	+0.018 [−0.011, +0.047]
Flat dense at 7,400 tokens	0.621	0.579	+0.041 [+0.009, +0.073]
Leiden at 7,400 tokens	0.620	0.558	+0.048 [+0.022, +0.073]
Leiden, UMAP-matched at 7,400 tokens	0.617	0.545	+0.057 [+0.033, +0.082]
DP mixture at 7,400 tokens	0.620	0.557	+0.044 [+0.017, +0.070]
Fixed-block contiguous at 7,400 tokens	0.630	0.566	+0.048 [+0.022, +0.074]

under the frozen one) and DP mixture minus Flat dense (−0.009 and +0.006), and none of their intervals excludes zero, so at this n neither flip is distinguished from noise. Both were computed after these tables were first read, and the reason for each is on the record. For DP mixture minus Document-order repack: the reader-side line this appendix is pointed at from Appendix S reports a comparison of that shape (Laitenberger et al., 2025), the registered contrast set carried no row for it, and the pooled accuracies already printed beside it had the readers ordering the pair oppositely. The direction was therefore in view when the interval was commissioned. For DP mixture minus Flat dense: the registered floor criterion of §4 compares exactly that pair under the frozen reader, and this table reported probe re-reads of the contrasts around it without reporting the probe re-read of that registered comparison. Withholding either interval would have been the selective omission this appendix's rule forbids, so both are computed and flagged as outcome-informed rather than presented as planned rows. The gap shift is descriptive on the accuracy scale: it says how a paired contrast differs between these readers, it needs no additivity assumption to compute, and it does not isolate an effect of model size. Carrying it to a third reader would assume the two effects add, which this design does not test. The gap-shift column holds 1 interval excluding zero, DP mixture minus Closed book at −0.079 [−0.130, −0.028]: the DP arm's margin over the closed-book floor is narrower under the probe reader. That is the floor paragraph above restated as an interval, not a statement about index structure, and the exposure caveat there is why it is read as a floor. In the column that reports the reader gain per condition (Table 38, drawn as Figure 8), 6 of the 30 intervals exclude zero: the closed-book row and all 5 conditions at 7,400 tokens. Two rows of that table coincide by construction: the flat arm's records at the sweep point just above the registered budget repeat the main run's, because top-20 retrieval over short chunks already exhausts its candidate list below the registered budget, so a larger cap leaves its context unchanged, and Appendix G reports all 1,526 predictions identical. 69 unadjusted descriptive 95% intervals are printed here with no

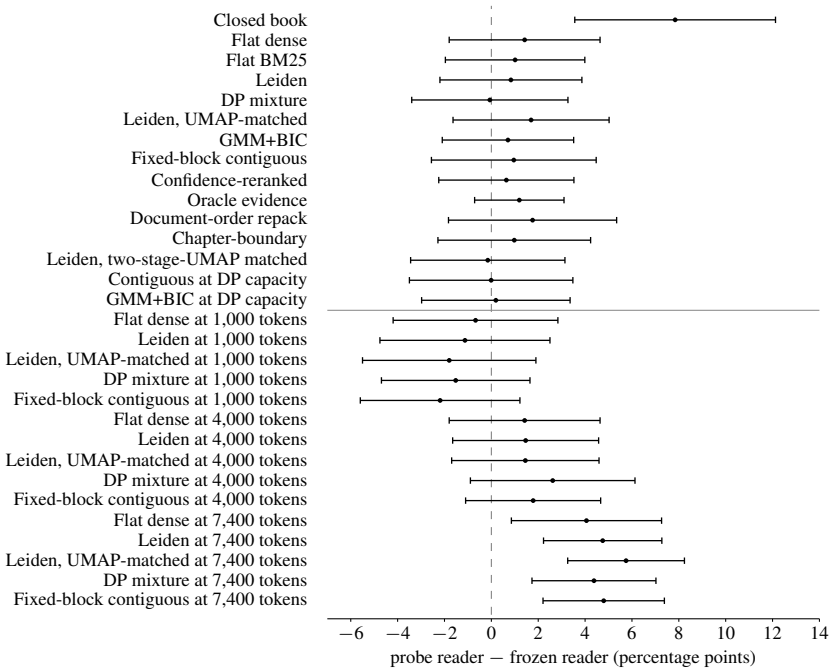

Figure 8: The last column of Table 38, drawn: the book-paired accuracy difference between these readers for each of the 30 conditions, with its descriptive 95% interval. Conditions appear in the order the run reports them; the dashed line is no difference and the rule separates the conditions the tables report from the budget sweep beneath them. 6 of the 30 intervals exclude zero; which conditions those are is read below the tables. Post hoc and exploratory, no registered status, no error rate, never pooled with the confirmatory records.

familywise error control and no confirmatory claim, so the tables are read whole and no single interval in them is a finding. That total includes the 2 late rows: adding them raised the count from 63 to 69.

Answer-format drift. The probe reader leaves far more completions unparseable than the frozen one. Under the frozen registered parse rule, the unparsed count runs from 14 to 227 per condition (largest at DP mixture at 1,000 tokens), that is from 0.9% of a condition's completions (Oracle evidence) to 14.9% (DP mixture at 1,000 tokens), against at most 6 under the frozen reader, and it is higher in 30 of the 30 conditions. An unparseable completion is scored wrong in every condition alike (Appendix Z), so this depresses the probe reader's accuracies, and it does so unevenly across conditions. The probe specification's reading rule, fixed before any paid call and not registered, treats a raised unparsed share as the reader's answer format drifting from the frozen prompt's expectation rather than as reader capability; either way it is a property of this probe and not of the indexes, and it is one more reason to read the probe column against the frozen column beside it rather than against the tables of the body. Appendix Y asks what these completions are, and finds abstentions rather than lettered answers the rule misses.

What this does not license. Nothing here settles a registered comparison: the registered verdicts stand as §4 states them, and this probe was never designed or powered to move one. A single larger reader is not a claim about stronger readers as a class. The readers differ in their weights and in whatever their post-training implies, and this design holds none of that fixed, so no difference between the columns is attributable to model size. The probe also inherits every exposure caveat of §7: a reader that answers more with no document at all may be recalling the benchmark rather than reading better, which is why the closed-book row is printed first and read as a floor rather than as a baseline.

Accounting. Two ledgers bill this probe and they are not the same number: the driver's own per-condition ledger accrues $162.21 over this probe's conditions, and the launcher's pod ledger, which bills the pod from boot to stop, records $172.31 over this probe's pod sessions; both released ledgers also hold later rows, which these sums leave out. Summed over completed conditions the wall clock

Table 39: Registered-shape contrasts under each reader in turn, and the gap shift between them. A row label reads *condition − reference* and may wrap onto a second line; the sign follows that order. Each cell is a book-paired difference with its descriptive 95% interval. The Oracle evidence row is computed on all 1,526 questions, including those without gold evidence spans, which the oracle gap of Table 7 drops. The gap shift is the probe reader's contrast minus the frozen reader's, * marks a sign change between the readers, and † marks the 2 contrasts computed after these tables were first read; the text of Appendix W names them and gives the reason for each. Post hoc and exploratory, no registered status, no error rate, never pooled with the confirmatory records.

Contrast	probe reader	frozen reader	gap shift
DP mixture − Leiden	+0.001 [−0.019, +0.020]	+0.010 [−0.011, +0.031]	−0.009 [−0.035, +0.017]
DP mixture − GMM+BIC	+0.009 [−0.012, +0.029]	+0.016 [−0.005, +0.038]	−0.008 [−0.037, +0.022]
DP mixture − Fixed-block contiguous	−0.011 [−0.036, +0.014]	−0.000 [−0.026, +0.025]	−0.010 [−0.045, +0.024]
Confidence-reranked − DP mixture	−0.011 [−0.025, +0.003]	−0.018 [−0.034, −0.002]	+0.007 [−0.012, +0.026]
Leiden − Flat BM25	+0.002 [−0.029, +0.033]	+0.004 [−0.025, +0.033]	−0.002 [−0.032, +0.029]
Oracle evidence − DP mixture	+0.192 [+0.160, +0.223]	+0.179 [+0.145, +0.213]	+0.013 [−0.024, +0.049]
DP mixture − Closed book	+0.113 [+0.070, +0.157]	+0.193 [+0.143, +0.242]	−0.079 [−0.130, −0.028]
DP mixture − Contiguous at DP capacity	−0.002 [−0.023, +0.018]	−0.002 [−0.025, +0.022]	−0.001 [−0.029, +0.028]
DP mixture − GMM+BIC at DP capacity	+0.014 [−0.010, +0.037]	+0.016 [−0.007, +0.039]	−0.003 [−0.033, +0.028]
Document-order repack − Flat dense	−0.005 [−0.024, +0.013]	−0.009 [−0.029, +0.011]	+0.003 [−0.025, +0.032]
Leiden, two-stage-UMAP matched − Leiden, UMAP-matched	−0.023 [−0.043, −0.002]	−0.004 [−0.023, +0.014]	−0.018 [−0.045, +0.008]
DP mixture − Document-order repack*†	−0.003 [−0.024, +0.017]	+0.015 [−0.012, +0.041]	−0.018 [−0.046, +0.010]
DP mixture − Flat dense*†	−0.009 [−0.031, +0.013]	+0.006 [−0.019, +0.031]	−0.015 [−0.042, +0.013]

is 17.00 hours, and the first pod start and the last pod stop are 18.84 hours apart; that window also holds model load, preflight, the gaps between conditions and the interruptions the run was resumed from, and it is charged only while a pod is up. Resumption cannot silently shorten a condition: every record file on both sides of these tables holds exactly 1,526 records with no missing, duplicate, out-of-band or error rows, and the generator that writes this appendix refuses to write it otherwise.

The second registration's reader run. The second registration budgets a reader run on both backbones as part of an item of its execution plan that also covers its stronger flat controls (Appendix L.1); that run has no hypothesis or analysis rule of its own. Under that item the frozen primary-backbone indexes were read again by `Qwen2.5-7B-Instruct` and by `Llama-3.3-70B-Instruct`, in 8 conditions over the same 1,526 questions and 60 books, with the driver, parse rule and reading rule of this appendix; each reader's summary labels its run post hoc and exploratory and carries 29 unadjusted intervals. Of these readers, `Qwen2.5-7B-Instruct` is the second registration's robustness-check backbone, while `Llama-3.3-70B-Instruct` is not among its backbones and the registration does not name it. Table 40 prints every contrast either run computed; each summary's other intervals are its per-condition reader gains, which the table does not print and the released summaries carry. Against the frozen reader, the sign of the difference changes in 4 of the 7 contrasts under `Qwen2.5-7B-Instruct` and in 3 under `Llama-3.3-70B-Instruct`, and every arm-against-arm interval covers zero under both: at this n those differences are not distinguished from zero. The gap shift excludes zero in 1 contrast under the first (confidence reranking against the DP index, +0.025 $[+0.005, +0.044]$) and in 3 under the second: the same contrast (+0.023), the DP

arm against GMM+BIC (−0.032 [−0.058, −0.006]), and the DP arm's margin over the closed-book floor (−0.101 [−0.154, −0.049]), under a reader whose question-pooled closed-book accuracy is 0.521 against the frozen reader's 0.389. Under the second reader 0.46–4.59% of the completions per condition were unparseable and scored wrong; under the first, none were. As with the probe above, nothing here settles a registered comparison and no difference between the columns is attributable to model size.

Table 40: The second registration's reader run: the first 7 contrasts of Table 39 for each of its readers, with the frozen reader's column repeated. Each cell is a book-paired difference with its descriptive 95% interval; * marks a sign change between the readers. Post hoc and exploratory, no error rate, never pooled with the confirmatory records.

Contrast	probe reader	frozen reader	gap shift
Qwen2.5 7B Instruct			
DP mixture − Leiden*	−0.002 [−0.017, +0.014]	+0.010 [−0.011, +0.031]	−0.011 [−0.037, +0.014]
DP mixture − GMM+BIC*	−0.011 [−0.029, +0.007]	+0.016 [−0.005, +0.038]	−0.027 [−0.055, +0.001]
DP mixture − Fixed-block contiguous	−0.003 [−0.021, +0.015]	−0.000 [−0.026, +0.025]	−0.003 [−0.032, +0.026]
Confidence-reranked − DP mixture*	+0.007 [−0.010, +0.023]	−0.018 [−0.034, −0.002]	+0.025 [+0.005, +0.044]
Leiden − Flat BM25*	−0.012 [−0.042, +0.017]	+0.004 [−0.025, +0.033]	−0.016 [−0.048, +0.016]
Oracle evidence − DP mixture	+0.155 [+0.126, +0.184]	+0.179 [+0.145, +0.213]	−0.024 [−0.057, +0.009]
DP mixture − Closed book	+0.165 [+0.128, +0.203]	+0.193 [+0.143, +0.242]	−0.027 [−0.070, +0.016]
Llama 3.3 70B Instruct			
DP mixture − Leiden	+0.009 [−0.012, +0.029]	+0.010 [−0.011, +0.031]	−0.001 [−0.026, +0.024]
DP mixture − GMM+BIC*	−0.016 [−0.034, +0.003]	+0.016 [−0.005, +0.038]	−0.032 [−0.058, −0.006]
DP mixture − Fixed-block contiguous	−0.010 [−0.032, +0.012]	−0.000 [−0.026, +0.025]	−0.009 [−0.038, +0.020]
Confidence-reranked − DP mixture*	+0.005 [−0.007, +0.017]	−0.018 [−0.034, −0.002]	+0.023 [+0.005, +0.042]
Leiden − Flat BM25*	−0.006 [−0.035, +0.023]	+0.004 [−0.025, +0.033]	−0.010 [−0.042, +0.022]
Oracle evidence − DP mixture	+0.207 [+0.178, +0.237]	+0.179 [+0.145, +0.213]	+0.028 [−0.006, +0.062]
DP mixture − Closed book	+0.091 [+0.051, +0.131]	+0.193 [+0.143, +0.242]	−0.101 [−0.154, −0.049]

X EXECUTED FAILURE AUDIT: THE THREE LABELS, POST HOC

Appendix P defines three labels (*reader residual*, *displacement*, *evidence absence*) that apply one per incorrectly answered question, disjoint by design, so that they partition the error event. This section executes that assignment. It was computed after the confirmatory verdict, it spends no error budget, it gates nothing, and no registered claim in this paper rests on it. We report it because a frame that is never executed is untested prose, and because the one class it was built to single out, displacement, is the mechanism this paper's central comparison is about.

What can be audited, and what cannot. Assigning a label requires knowing which leaves carry a question's gold evidence. We locate a gold snippet by exact substring search over the book's own leaf text under a normalization looser than the recall metric's (letters and digits only, casefolded, versus casefold with whitespace collapsed), so localized and recall-matchable snippet counts are not on the same footing; localization succeeds for every snippet of 748 of the 1519 evidence-bearing questions (49.2%). Of the 771 that fail, 686 have a snippet that does not occur verbatim in the harness's extracted text (the same difference that sets the recoverability reference of Appendix O), while 84 have one below the length floor and 1 has one that occurs more than once in the book. Localisation reads only the book text and the leaf chunking, both shared by every arm, so the audited set is *identical across arms*; differences between rows of Table 41 are not confounded by which questions

could be audited. The minimum snippet length is 40 characters. Re-running the whole audit at 5 settings of that threshold, spanning 24 to 80 characters, moves the audited count over 545–791 and leaves every reading below intact: at each of those settings, and under both coverage definitions, displacement remains the smallest of the three labels in every arm and evidence absence the largest.

The audited half is not the full band. It differs from the unlocalisable remainder exactly as the mechanism predicts: a question with n snippets is audited only if all n are found, so multi-snippet questions drop out faster. Audited questions carry 1.278 gold snippets on average against 2.097 for the remainder; multi-hop items are 0.3061 of the audited set against 0.4643 of the remainder, and detail items 0.2968 against 0.1855. Accuracy is correspondingly higher on the audited half (DP mixture 0.5922 against 0.5409; flat dense 0.5735 against 0.5499). The absolute shares in Table 41 are therefore *not* estimates for the full band and we do not extrapolate them. Comparing rows does not equalize the denominators either: each row of Table 41 is conditioned on that arm's own wrong answers, so its denominator is the n that row prints. The same audit read unconditionally (Table 42) does put every arm on one denominator, the audited questions themselves, and that makes it the cleaner descriptive comparison; it removes the differing denominators but not the selection, and neither table generalizes to the full band.

Table 41: Executed three-label decomposition of wrong answers: post hoc, exploratory, no registered status. Rates are book means with each contributing book weighted equally, and the contributing set is not the whole corpus: the three shares are conditional on an incorrect answer and sum to one, so a book on which the arm makes no audited error cannot enter the mean even as a zero. That set holds 54–56 of the 60 books, arm by arm. The audited question set is the same in every arm; the book set entering each book mean is not. n is how many of the 748 audited questions the arm answers incorrectly. Two coverage definitions are reported and neither is cited alone: (A) counts only leaves retrieved directly, (B) counts the leaf closure of every retrieved node. Each triple sums to one by construction; the printed shares are rounded independently, so a printed triple can differ from one in the last digit. Absolute levels do not transfer to the full band (see text); the comparison down a column is over a fixed question set. Shares are conditional on an incorrect answer, so the denominators differ by arm and the shares are not risks. The flat dense row retrieves no summaries, so its displacement is zero by construction and the two definitions coincide. †Under (B) this column is not reader failure: leaf-closure coverage does not establish that the evidence survived summarization, so the (B) column also holds summaries that never carried the fact. The label name is kept for parity with (A); the event it names is wider.

		(A) retrieved leaves			(B) leaf closure		
Condition	n	Residual	Displ.	Absence	Residual†	Displ.	Absence
Flat dense	319	0.2968	0.0000	0.7032	0.2968	0.0000	0.7032
GMM+BIC	315	0.2583	0.0066	0.7351	0.3055	0.0056	0.6889
GMM+BIC at DP capacity	318	0.2429	0.0364	0.7207	0.3290	0.0288	0.6422
Leiden	313	0.2726	0.0116	0.7159	0.3336	0.0032	0.6632
Leiden, UMAP-matched	318	0.2510	0.0188	0.7302	0.3178	0.0174	0.6648
Leiden, two-stage-UMAP matched	313	0.2382	0.0085	0.7533	0.2736	0.0044	0.7220
DP mixture	305	0.2056	0.0329	0.7615	0.3001	0.0223	0.6777
Fixed-block contiguous	301	0.2641	0.0134	0.7225	0.3997	0.0066	0.5937
Contiguous at DP capacity	299	0.2390	0.0385	0.7225	0.3458	0.0147	0.6395
Chapter-boundary	300	0.2592	0.0138	0.7271	0.4298	0.0068	0.5634

What the labels say. Under either definition, displacement (the class the hierarchy is supposed to trade against, where summary nodes crowded out leaves that would have covered the evidence) is the smallest of the three in every arm: at most about 0.0385 of an arm's errors under (A) and about 0.0288 under (B), bounds taken over the whole threshold sweep, whose maxima are attained at the headline setting. The sweep widens the support for the reading without changing its value. The DP mixture carries more of it than Leiden under both definitions (0.0329 against 0.0116 under (A), 0.0223 against 0.0032 under (B)), the direction its larger summary share predicts, but the magnitude is small beside the class that dominates every row: evidence absence, 0.7032–0.7615 under (A) and 0.5634–0.7220 under (B). The flat dense floor retrieves no summaries and so has displacement zero by construction (and, for the same reason, the two definitions coincide on it), yet it still fails

Table 42: Unconditional audit risks (post hoc, exploratory; not preregistered, and spending no alpha). Each cell is the book-mean of that book's errors of one label over that book's audited questions, so the three cells of a row sum to that arm's audited error risk rather than to one. Denominator: the same 748 questions in every arm, over the 57 books that carry one; localisation is arm-invariant, so all arms are compared on the same audited subset, but the audited subsample is measurably biased and these absolute risks do not extrapolate to the full band. Both coverage definitions are reported, and neither is cited alone. Contrast Table 41, whose shares are conditional on an incorrect answer. †Under (B) this column is not reader failure but the wider event defined in Appendix P: covered-by-ancestry and still wrong, which includes summaries that dropped the fact.

	(A) direct-leaf			(B) leaf-closure		
Arm	Reader	Displ.	Absence	Reader†	Displ.	Absence
Flat dense	0.1351	0.0000	0.2971	0.1351	0.0000	0.2971
GMM+BIC	0.1108	0.0034	0.3034	0.1338	0.0028	0.2811
GMM (matched)	0.1109	0.0128	0.2897	0.1427	0.0103	0.2603
Leiden	0.1230	0.0064	0.2859	0.1498	0.0013	0.2641
UMAP-Leiden	0.1172	0.0087	0.2914	0.1477	0.0081	0.2616
UMAP-Leiden (2-stage)	0.1004	0.0052	0.2968	0.1191	0.0026	0.2806
DP mixture	0.0862	0.0162	0.2979	0.1205	0.0107	0.2691
Contiguous	0.1242	0.0062	0.2822	0.1806	0.0026	0.2294
Contiguous (cap)	0.1095	0.0138	0.2877	0.1560	0.0059	0.2492
Chapter	0.1164	0.0071	0.2845	0.1827	0.0021	0.2231

for evidence absence on 0.7032 of its errors. These shares are conditional on an incorrect answer, so they are not risks and their denominators differ by arm; the same audit read on the common denominator of all 748 audited questions (Table 42) puts displacement at about 0.0162 or below under (A) and 0.0107 or below under (B), inside a total audited error risk of 0.4003–0.4322 under (A). On that reading the class that exists only because there is a tree accounts for a small part of what the tree arms lose. The split itself has precedent: a diagnostic benchmark separates evidence-construction failure from reader synthesis in a different setting (Wu et al., 2026); these rows apply that split under a fixed budget with the construction operator varied.

This is a decomposition of where the errors sit, not a causal attribution: the labels are assigned under the disjoint definitions of Appendix P, and the identification of what moves them remains the contrast map of Bookkeeping identity 1. Figure 9 draws the same cells as bars, which is how the row total becomes visible: Table 42 prints the three parts and leaves the sum to the reader.

Controls. The audit's first control checks its localization against a registered instrument whose matching rule shares no code with that localization; the two read the gold snippets through one loader, and the control's population is selected under the matcher's normalization. On 2010 records spread over all 10 audited arms (restricted to questions whose every gold snippet sits inside one leaf), definition-(A) coverage and the registered gold-evidence substring matcher, run on the reassembled reader context, returned the same verdict 2010 times, with no disagreement. A fault in localization, in the leaf-boundary convention, in the budget prefix rule or in the record inputs breaks that agreement in one direction or the other. Span localization is checked for minimality (every one of 956 snippets is refound from its own recorded span, and all 378 single-leaf deletions from the 182 spans that cross more than one leaf are rejected) and for adjacency dependence: after shuffling leaf order within a book, 1 of 255 adjacency-dependent spans still localized. The record-shape filter the audit depends on was exercised against 10 known-unparseable records drawn from a condition family the audit never touches, and caught 10. Every rate above is emitted only if all of these pass, together with a closure check that each triple sums to one and a cross-check that the error counts implied by Table 41 agree with the accuracies quoted above. Those arithmetic checks and the three readings are applied to the headline audit and to each threshold setting alike, not to the headline alone.

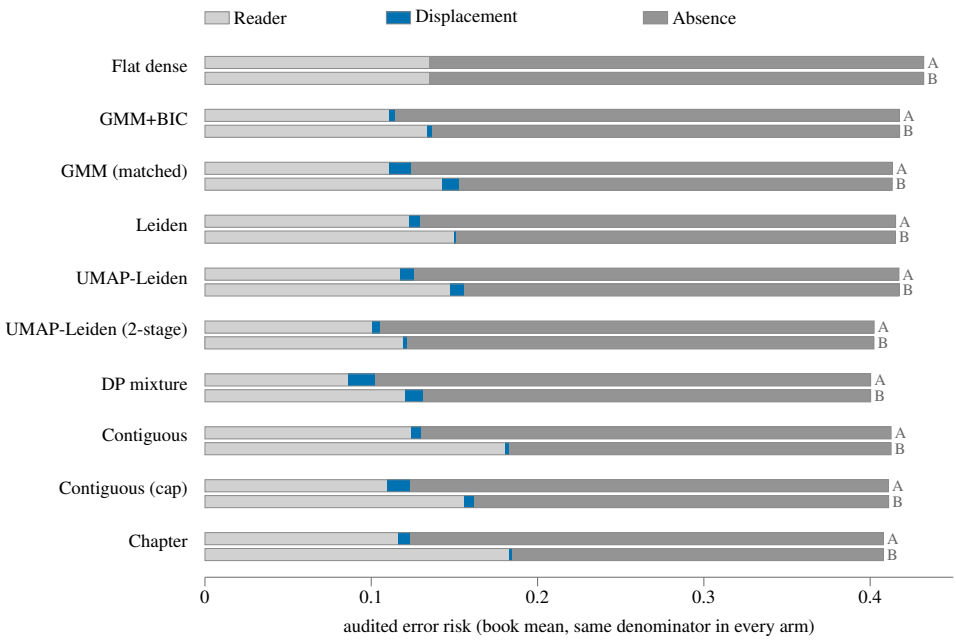

Figure 9: The cells of Table 42 drawn as stacked bars, upper bar (A) direct-leaf and lower bar (B) leaf-closure for each arm. *What the table does not give.* The table prints the three parts and not their sum, so the arm's audited error risk – the length of the whole bar – is the one quantity a reader has to compute before comparing arms; drawn, the two coverage definitions give the same length within each arm (the same underlying value; the printed parts are rounded separately, so their sums can differ in the last place) and similar magnitudes across arms, which places the comparison in the composition of the bar rather than in its length. *Reading rules.* Segments are stacked in the printed column order and the bar is the row total, so nothing here is a proportion of anything but the audited denominator. The displacement segment is drawn in the accent colour because it is the class that exists only because there is a tree; it is the smallest of the three in every bar and in the flat arm it is zero by construction, so that bar has two segments. Every segment length comes from a cell of Table 42; the figure adds only their stacked sum, and the axis carries no interval because these risks are point estimates on an audited subsample that is measurably biased.

Y What the probe reader's unparsed completions are

Status. **Post hoc and exploratory**, like the probe it reads (Appendix W): no registered status, no error rate, never pooled with the confirmatory records. It changes no registered number. The frozen column below is the probe's own, and the generator that writes this appendix refuses to write it unless that column equals the driver's summary.

Question. Appendix W reports that a same-generation, larger reader leaves 14 to 227 of each condition's completions unparseable under the frozen parse rule, and that each such completion is scored wrong; over the 30 conditions this probe reads, 4,203 of 45,780 completions (9.2%) are unparsed. The probe specification's reading rule calls this answer-format drift. Two things could sit behind an unparsed completion: an answer written in a shape the frozen rule declines to read (a lowercase letter, an option quoted rather than lettered), or no answer at all. They have different consequences for the probe column, so we asked which it is.

Method. A lenient re-parse with 4 rules, fixed before the first pass over the probe's completions and not revised after it; the rule file the release ships hashes to the value the shipped aggregate records, and the generator checks that. The rules: a lowercase letter after the final "answer:" marker, closed by a bracket, a period or a colon, or bare only when generation finished under the cap; a bracketed lowercase letter opening the completion or a line, or a completion that is nothing but a lowercase letter and its closer; the letter after "option" or "choice" when it closes the clause; and a completion that is the beginning of a single option text, or begins with a single one, both sides at least a fixed minimum length after a standard (SQuAD-style) answer normalisation. The rules see the completion and the option texts, never the answer key; a letter they recover is scored afterwards against the

same key the registered run used. Two candidates from one rule count as ambiguous, not as a pick. Contrasts are book-paired over the same 60 books as Appendix W, and the generator checks that every contrast pairs all of them. A record no rule resolves enters two columns: the frozen score, which is the registered rule and scores such a record wrong, and an all-unparsed-correct scenario in which every unresolved record is scored right. For a condition's accuracy on its own the two bracket it and that scenario is a ceiling, a bound no parser could exceed; for a contrast neither column bounds the difference, because the scenario moves both sides at once. It is printed so that the shift it produces can be read against the frozen column. The counts, the shape flags and the point deltas below are recomputed from a released per-record file that carries, for each record, the frozen parse outcome, the token count and a few boolean shape predicates, and no completion text; the intervals are the driver library's own, copied from the shipped aggregate, and the distinct-string census below, and the no-marker and indefinite-article counts of the "answer … letter" shape, are read from that aggregate too, because the strings they count do not ship. The shape predicates, a post-hoc abstention phrase list among them, were written after the zero-recovery run, to say what the completions are; they score nothing and move no bound.

Result. The rules recover 0 of the 4,203 unparsed completions; 0 are ambiguous and 4,203 unresolved, so the low bound is the frozen column itself. What most of the unparsed completions contain instead is an abstention: 3,839 of them (91.3%) match a phrase on the post-hoc list. Its families overlap, so 3,839 is their union rather than their sum: 3,772 match a "none of the above" phrase, 1,205 a "not mentioned in the text" phrase and 11 a "cannot" phrase, against 47 of the 41,577 parsed completions (0.1%). They are templated: the 4,203 strings comprise 1,196 distinct completions, and the most frequent one alone accounts for 805 (19.2%). 3,226 (76.8%) ran to the registered 16-token cap, against 162 of the parsed ones; 616 contain the word "answer" at all and 21 an "answer … letter" shape. That shape is not a lettered answer the rules missed: the shape predicate accepts "answer is" and a bare "answer" and ignores letter case, whereas the first rule needs the "answer:" marker, and 21 of the 21 carry no such marker; 21 of them match on the letter *a* followed by a word, that is, on the indefinite article, and 20 ran to the cap.

Reading. In 91.3% of the probe reader's unparsed completions the completion is an abstention phrase rather than a letter; the remainder match neither a letter shape the rules read nor a phrase on the list. The registered rule scores every unparsed completion as wrong, in every condition alike. For the part the census covers, the asymmetry across conditions in Appendix W is therefore an asymmetry in how often the larger reader abstains over each index, a property of that reader against these prompts rather than an artefact these 4 rules could repair. Table 44 gives the per-condition counts. Figure 10 sets that table's unparsed share against the reader gain Table 38 reports for the same condition, a pairing neither table can show because the two columns sit in different appendices. The points fall from upper left to lower right, and the reader budgets form a ladder along that fall, the 7,400-token conditions at the upper left and the smallest-budget ones at the lower right. Both quantities therefore move with the budget, and the figure describes that joint movement rather than identifying either quantity as a cause of the other. The all-unparsed-correct column of Table 43 shows what counting every unresolved completion as right would do. Of the 13 contrasts, 10 cover zero under the frozen rule and under that scenario alike and are not distinguished at this n under either; 2 (Oracle evidence − DP mixture; DP mixture − Closed book) exclude zero under both; and the interval of 1 (Leiden, two-stage-UMAP matched − Leiden, UMAP-matched) covers zero under the all-unparsed-correct scenario only. The largest shift of a point estimate is 0.100 (Oracle evidence − DP mixture), a contrast that involves Oracle evidence, the condition that abstains least. Nothing here bears on the registered verdicts.

Table 43: Book-paired contrasts of the reader probe under the frozen parse rule and under the all-unparsed-correct scenario that scores every unparsed completion as right. That scenario moves both sides of a contrast, so its column is a scenario for the difference and not a bound on it. A row label reads *condition* − *reference*; unparsed counts are out of 1,526 records per condition; each cell is a book-paired difference with its descriptive 95% interval, and the last column says whether that interval covers zero. Post hoc and exploratory; no registered status, no error rate, never pooled with the confirmatory records. The frozen column is Table 39's probe column.

Contrast	unparsed	frozen Δ	all-unparsed-correct Δ	covers 0
DP mixture − Leiden	162 / 141	+0.001 [−0.019, +0.020]	+0.015 [−0.006, +0.036]	yes / yes
DP mixture − GMM+BIC	162 / 178	+0.009 [−0.012, +0.029]	+0.005 [−0.018, +0.029]	yes / yes
DP mixture − Fixed-block contiguous	162 / 128	−0.011 [−0.036, +0.014]	+0.008 [−0.015, +0.030]	yes / yes
Confidence-reranked − DP mixture	169 / 162	−0.011 [−0.025, +0.003]	−0.006 [−0.020, +0.009]	yes / yes
Leiden − Flat BM25	141 / 176	+0.002 [−0.029, +0.033]	−0.017 [−0.040, +0.005]	yes / yes
Oracle evidence − DP mixture	14 / 162	+0.192 [+0.160, +0.223]	+0.092 [+0.064, +0.120]	no / no
DP mixture − Closed book	162 / 116	+0.113 [+0.070, +0.157]	+0.137 [+0.092, +0.183]	no / no
DP mixture − Contiguous at DP capacity	162 / 159	−0.002 [−0.023, +0.018]	+0.002 [−0.019, +0.023]	yes / yes
DP mixture − GMM+BIC at DP capacity	162 / 175	+0.014 [−0.010, +0.037]	+0.010 [−0.013, +0.032]	yes / yes
Document-order repack − Flat dense	139 / 138	−0.005 [−0.024, +0.013]	−0.006 [−0.025, +0.013]	yes / yes
Leiden, two-stage-UMAP matched − Leiden, UMAP-matched	153 / 137	−0.023 [−0.043, −0.002]	−0.021 [−0.043, +0.001]	no / yes
DP mixture − Document-order repack	162 / 139	−0.003 [−0.024, +0.017]	+0.013 [−0.011, +0.037]	yes / yes
DP mixture − Flat dense	162 / 138	−0.009 [−0.031, +0.013]	+0.007 [−0.015, +0.029]	yes / yes

Table 44: Per condition: completions the frozen rule left unparsed, how many of them the lenient rules recovered or found ambiguous, how many match an abstention phrase, how many ran to the 16-token cap, and the question-pooled accuracy under the frozen rule and under the ceiling. The unparsed count is out of 1,526 records per condition; the recovered, ambiguous, abstention and cap counts are out of that row's unparsed count. Post hoc and exploratory; no registered status, no error rate, never pooled with the confirmatory records.

Condition	unparsed	recov.	ambig.	abstain	at cap	frozen	ceiling
Chapter-boundary	146	0	0	136	117	0.599	0.695
Closed book	116	0	0	55	92	0.478	0.554
Confidence-reranked	169	0	0	154	125	0.564	0.674
Contiguous at DP capacity	159	0	0	150	121	0.579	0.683
DP mixture	162	0	0	149	119	0.575	0.681
DP mixture at 1,000 tokens	227	0	0	222	151	0.533	0.682
DP mixture at 4,000 tokens	138	0	0	123	87	0.598	0.689
DP mixture at 7,400 tokens	111	0	0	94	90	0.620	0.693
Document-order repack	139	0	0	125	124	0.579	0.670
Fixed-block contiguous	128	0	0	119	109	0.596	0.680
Fixed-block contiguous at 1,000 tokens	184	0	0	174	136	0.549	0.670
Fixed-block contiguous at 4,000 tokens	115	0	0	102	86	0.607	0.682
Fixed-block contiguous at 7,400 tokens	86	0	0	77	65	0.630	0.687
Flat BM25	176	0	0	165	153	0.559	0.674
Flat dense	138	0	0	128	119	0.587	0.677
Flat dense at 1,000 tokens	155	0	0	145	124	0.559	0.661
Flat dense at 4,000 tokens	138	0	0	128	119	0.587	0.677
Flat dense at 7,400 tokens	101	0	0	91	68	0.621	0.687
GMM+BIC	178	0	0	169	122	0.567	0.683
GMM+BIC at DP capacity	175	0	0	162	129	0.562	0.677
Leiden	141	0	0	130	114	0.579	0.671
Leiden at 1,000 tokens	184	0	0	175	147	0.539	0.659
Leiden at 4,000 tokens	129	0	0	118	104	0.588	0.672
Leiden at 7,400 tokens	102	0	0	92	68	0.620	0.687
Leiden, UMAP-matched	137	0	0	126	108	0.587	0.677
Leiden, UMAP-matched at 1,000 tokens	176	0	0	167	129	0.541	0.657
Leiden, UMAP-matched at 4,000 tokens	130	0	0	119	98	0.588	0.674
Leiden, UMAP-matched at 7,400 tokens	96	0	0	89	72	0.617	0.680
Leiden, two-stage-UMAP matched	153	0	0	143	119	0.564	0.664
Oracle evidence	14	0	0	12	11	0.775	0.784

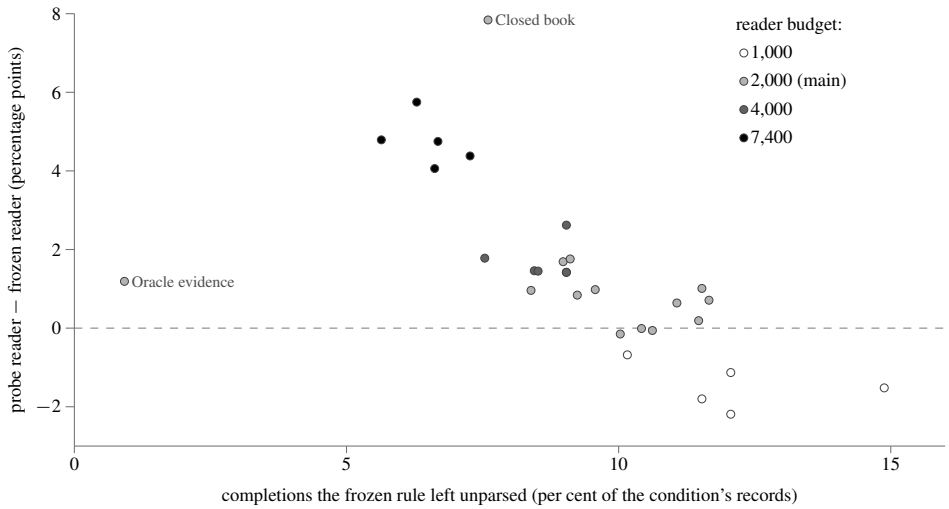

Figure 10: Per condition: the share of the probe reader's completions the frozen rule left unparsed, out of 1,526 records per condition (horizontal; Table 44 gives the counts), against the book-paired reader gain of Table 38 and Figure 8 (vertical), with shade giving the reader budget. Both quantities are already printed, each in its own appendix; what is drawn only here is the pair, which is the relation the paragraph above states in words. The fall from upper left to lower right runs with the budget ladder, so the picture reports a joint movement and identifies neither axis as a cause of the other. The dashed line is no difference. The closed-book and oracle conditions are named. The 30 conditions are drawn on one fewer mark, because the flat arm's 2,000- and 4,000-token points are one output – the first disclosure of Appendix G – and so share every field this figure reads. Post hoc and exploratory; no registered status, no error rate, never pooled with the confirmatory records.

Z REGISTERED PARSER-SHAPE MONITOR: THE FLAGGED ROWS

The registration monitors completion shape under its closed-book memorization gate, and we apply the monitor to every condition counted below: its count (a) is the number of completions in which an option letter appears two or more times (we count two or more option-letter occurrences in total, whether or not the same letter; expected zero), and it commits us to disclosing every such completion verbatim if the count is nonzero. Over the first registration's NovelQA conditions on the primary backbone, its option-order probe aside, the seven confirmatory conditions (10,682 records) plus the non-confirmatory registered ones (35,098 records), the confirmatory count is zero (with 3 unparseable completions, from which no option letter was recovered, out of 10,682); the non-confirmatory count is 122, all of them in the 7,400-token budget-curve conditions, and Table 46 lists each one by its shape. No completion is printed: beside the released per-row correctness, a printed completion would carry information about the benchmark's request-gated answer key, and every row keeps its shape columns (Appendix J records the departure from the registered duty). The release ships this monitor's aggregate with each row's condition, question identifier, parse stage, letter counts and first-versus-last flag, and with the completion text withheld, as for every parsed completion. The paper and the release therefore withhold the same text; the count (a) and its rows are read from the shipped aggregate rather than recounted over completions that do not ship, the first-versus-last disagreement count below follows from the flag, and every column the table prints re-derives from that aggregate.

Parse stages and per-arm unparseable rates. The registered monitor also fixes the parser's *stage* for every completion (`marker`, `last_letter`, `option_text`, or `none`) and the registration asks for the histogram and for the unparseable rate per arm, not only in total. Table 45 prints both for the seven confirmatory conditions: every parsed completion resolved at the `last_letter` stage (10679 of 10682 records; `marker` and `option_text` fired 0 and 0 times), and the 3 unparseable completions all belong to a single condition (Closed book). Over the 23 non-confirmatory conditions (35098 records) there are 10 unparseable completions, in 4 budget-curve conditions (arm0@7400 1; arm2@7400 2; arm3@4000 1; arm3@7400 6). An unparseable completion is scored wrong in every arm alike, by the frozen rule.

Table 45: Parse-stage histogram and unparseable rate per confirmatory condition, from the registered FC0 parse monitor (FC0 is the registered closed-book memorization gate; primary backbone only). The per-condition counts are cross-checked against the analysis code's single unparseable predicate before this table is written.

Condition	n	marker	last letter	option text	none	unparseable
Closed book	1,526	0	1,523	0	3	0.0020
Flat dense	1,526	0	1,526	0	0	0.0000
GMM+BIC tree	1,526	0	1,526	0	0	0.0000
Leiden tree	1,526	0	1,526	0	0	0.0000
Leiden tree, UMAP-matched	1,526	0	1,526	0	0	0.0000
DP-mixture tree	1,526	0	1,526	0	0	0.0000
+ confidence reranking	1,526	0	1,526	0	0	0.0000

Parsing accuracy. A repeated letter does not by itself mean the parser misread the completion: in 114 of the 122 rows the letter occurs exactly twice and in most rows every occurrence is the same letter, so the registered last-letter rule and a first-letter rule land on the same letter. Re-reading each completion under the audit regex, the first and last letter disagree in 15 rows (marked $\times$ in the table). These are the rows where the two readings conflict; in some of them the completion itself names different letters in different places, so no mechanical rule picks one of them over the others. They arise from two mechanisms, both identified by reading the completions, which the table does not print: an option whose text begins with the article "A" (the parser takes the trailing article rather than the chosen letter), and truncation residue at the 16-token completion cap, where the cut-off prose ends in a stray capital. The release withholds those completions too, so neither mechanism can be checked against it. Both mechanisms are specific to the long-budget regime, and no confirmatory completion triggers the monitor. That is the monitor reporting nothing on the confirmatory set, not a validation of those completions: this monitor watches parsed shape, so a confirmatory answer mis-parsed in a way the monitor does not watch would not appear here.

Sensitivity to the reading rule. The registered rule reads the last letter of the completion. Reading the first instead changes the scored letter on the 15 rows marked × in Table 46 and on no other monitored row. That reach is a set of records, not a re-scored accuracy, and no re-scored accuracy is reported here: on the rows whose completion names different letters in different places, no gold-blind rule decides which letter was intended, and this monitor adjudicates none of them. Three things bound the sensitivity. Every one of those rows sits in the long-budget conditions and none in the confirmatory ones, as the count above records; the largest count in any single arm is 8 of that arm's 1,526 questions. Agreement between the two readings is not validation of either: 16 monitored completions name a second option letter while the readings conflict on only 15, so in the remaining completion the second letter sits between a first and a last that agree. And the monitored population is defined by two or more option-letter occurrences, so a completion with at most one never enters it: this bounds the rows the monitor sees, not every record.

The table below lists every monitored row by its shape and prints no completion, which leaves the registered duty unmet for every row (Appendix J). The release adds a machine-readable copy of the same rows (`evidence/fc0_disclosure_index.jsonl`), carrying each row's condition, query identifier, occurrence and distinct-letter counts, its first-versus-last disagreement flag, and whether this table prints its completion, which it does for no row. No row leaves this appendix; the rows gain a second form that joins to the released per-query records on the query identifier.

Table 46: All 122 monitored rows (count (a): an option letter appearing ≥ 2 times), each listed by its shape with its completion withheld from print: beside the released per-row correctness a completion would carry information about the benchmark's request-gated answer key (a departure from the registered duty; Appendix J). Every row parsed at the last-letter stage; × marks the 15 rows whose first and last letter disagree (the two readings conflict).

Condition	Query	Occ.	Distinct	First ≠ last	Completion
arm0@7400	Q1569	2	1		*withheld*
arm2@7400	Q0201	2	1		*withheld*
arm2@7400	Q0212	2	1		*withheld*
arm2@7400	Q0318	2	1		*withheld*
arm2@7400	Q0652	2	2	×	*withheld*
arm2@7400	Q0658	2	1		*withheld*
arm2@7400	Q0705	2	1		*withheld*
arm2@7400	Q0706	3	1		*withheld*
arm2@7400	Q0841	2	1		*withheld*
arm2@7400	Q1221	2	1		*withheld*
arm2@7400	Q1249	2	1		*withheld*
arm2@7400	Q1356	2	1		*withheld*
arm2@7400	Q1489	2	1		*withheld*
arm2@7400	Q1569	2	1		*withheld*
arm2@7400	Q1570	2	1		*withheld*
arm2@7400	Q1596	2	1		*withheld*
arm2@7400	Q1625	2	1		*withheld*
arm2@7400	Q1630	2	1		*withheld*
arm2@7400	Q2269	2	1		*withheld*
arm2p@7400	Q0194	2	1		*withheld*
arm2p@7400	Q0200	2	1		*withheld*
arm2p@7400	Q0201	2	1		*withheld*
arm2p@7400	Q0372	2	1		*withheld*
arm2p@7400	Q0652	3	1		*withheld*
arm2p@7400	Q0665	2	1		*withheld*
arm2p@7400	Q0689	2	1		*withheld*
arm2p@7400	Q0700	2	2	×	*withheld*
arm2p@7400	Q0706	3	1		*withheld*
arm2p@7400	Q0841	2	1		*withheld*
arm2p@7400	Q1221	2	1		*withheld*

Table 46 continued

Condition	Query	Occ.	Distinct	First ≠ last	Completion
arm2p@7400	Q1359	2	1		*withheld*
arm2p@7400	Q1367	2	1		*withheld*
arm2p@7400	Q1378	2	1		*withheld*
arm2p@7400	Q1408	2	1		*withheld*
arm2p@7400	Q1472	2	1		*withheld*
arm2p@7400	Q1568	2	1		*withheld*
arm2p@7400	Q1569	2	1		*withheld*
arm2p@7400	Q1574	2	1		*withheld*
arm2p@7400	Q1584	2	1		*withheld*
arm2p@7400	Q1591	2	1		*withheld*
arm2p@7400	Q1596	2	1		*withheld*
arm2p@7400	Q1901	2	1		*withheld*
arm2p@7400	Q1968	2	1		*withheld*
arm2p@7400	Q2219	3	2		*withheld*
arm3@7400	Q0126	2	1		*withheld*
arm3@7400	Q0316	2	2	×	*withheld*
arm3@7400	Q0532	2	1		*withheld*
arm3@7400	Q0637	2	1		*withheld*
arm3@7400	Q0659	2	2	×	*withheld*
arm3@7400	Q0660	2	1		*withheld*
arm3@7400	Q0706	3	1		*withheld*
arm3@7400	Q0709	2	1		*withheld*
arm3@7400	Q0714	2	2	×	*withheld*
arm3@7400	Q0758	2	1		*withheld*
arm3@7400	Q0770	2	2	×	*withheld*
arm3@7400	Q0780	2	1		*withheld*
arm3@7400	Q0842	2	1		*withheld*
arm3@7400	Q0872	2	1		*withheld*
arm3@7400	Q0879	2	1		*withheld*
arm3@7400	Q0919	2	1		*withheld*
arm3@7400	Q0920	2	2	×	*withheld*
arm3@7400	Q1005	2	1		*withheld*
arm3@7400	Q1139	2	1		*withheld*
arm3@7400	Q1179	2	2	×	*withheld*
arm3@7400	Q1181	2	1		*withheld*
arm3@7400	Q1187	2	1		*withheld*
arm3@7400	Q1201	3	1		*withheld*
arm3@7400	Q1245	2	1		*withheld*
arm3@7400	Q1249	2	1		*withheld*
arm3@7400	Q1320	2	1		*withheld*
arm3@7400	Q1406	2	1		*withheld*
arm3@7400	Q1439	2	1		*withheld*
arm3@7400	Q1588	2	1		*withheld*
arm3@7400	Q1596	2	1		*withheld*
arm3@7400	Q1601	2	2	×	*withheld*
arm3@7400	Q1603	2	1		*withheld*
arm3@7400	Q1646	2	1		*withheld*
arm3@7400	Q1958	2	1		*withheld*
arm3@7400	Q1967	2	1		*withheld*
arm3@7400	Q1977	2	2	×	*withheld*
arm3@7400	Q2040	2	1		*withheld*
arm3@7400	Q2193	2	1		*withheld*
arm3@7400	Q2202	2	1		*withheld*
arm3@7400	Q2209	2	1		*withheld*
arm3@7400	Q2214	2	1		*withheld*

Table 46 continued

Condition	Query	Occ.	Distinct	First ≠ last	Completion
arm3@7400	Q2219	2	1		*withheld*
arm3@7400	Q2238	2	1		*withheld*

Table 46 continued

Condition	Query	Occ.	Distinct	First ≠ last	Completion
arm5@7400	Q0318	2	1		*withheld*
arm5@7400	Q0365	2	1		*withheld*
arm5@7400	Q0367	2	2	×	*withheld*
arm5@7400	Q0375	2	2	×	*withheld*
arm5@7400	Q0637	2	1		*withheld*
arm5@7400	Q0652	2	2	×	*withheld*
arm5@7400	Q0690	2	1		*withheld*
arm5@7400	Q0706	3	1		*withheld*
arm5@7400	Q0709	2	1		*withheld*
arm5@7400	Q0758	2	1		*withheld*
arm5@7400	Q0780	2	1		*withheld*
arm5@7400	Q0835	2	1		*withheld*
arm5@7400	Q0841	2	1		*withheld*
arm5@7400	Q0879	2	1		*withheld*
arm5@7400	Q0907	2	1		*withheld*
arm5@7400	Q1011	2	1		*withheld*
arm5@7400	Q1136	2	1		*withheld*
arm5@7400	Q1142	2	1		*withheld*
arm5@7400	Q1176	2	1		*withheld*
arm5@7400	Q1203	3	1		*withheld*
arm5@7400	Q1221	2	1		*withheld*
arm5@7400	Q1249	2	1		*withheld*
arm5@7400	Q1367	2	1		*withheld*
arm5@7400	Q1435	2	1		*withheld*
arm5@7400	Q1439	2	1		*withheld*
arm5@7400	Q1571	2	1		*withheld*
arm5@7400	Q1584	2	1		*withheld*
arm5@7400	Q1603	2	1		*withheld*
arm5@7400	Q1628	2	1		*withheld*
arm5@7400	Q1898	2	2	×	*withheld*
arm5@7400	Q1901	2	1		*withheld*
arm5@7400	Q1977	2	2	×	*withheld*
arm5@7400	Q2198	2	1		*withheld*
arm5@7400	Q2228	2	1		*withheld*
arm5@7400	Q2249	2	1		*withheld*