THE DECISION VALUE OF INFERRED STRUCTURE IN BUDGETED EVIDENCE ACQUISITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Model-inferred structures guide which evidence a retrieval-augmented system reads under a limited budget. When these structures omit useful evidence, the effect on answer quality may depend on whether later reading can recover it. We investigate this relationship through controlled comparisons on MuSiQue, starting from identical evidence states and independently varying structure and passage selection. We introduce structural defects by removing unread supporting passages from the structure while keeping them in the candidate pool. On affected states in a confirmation experiment using new source components, the same defect reduces answer F1 by 0.22 with a lexical selector and 0.01 with a strong model selector. This asymmetry persists with a second model endpoint and with 100-passage retrieved pools. On a sealed split of 1,668 questions, opened once for the final evaluation, the same defect reduces F1 by 0.21 with the lexical selector and 0.01 with the model selector, and structure changes the model selector's F1 by -0.01 (within ± 0.03). The model selector more often finds the omitted evidence after returning to the full candidate pool. Structure improves the lexical selector but adds little to the strong model selector ($+0.004$, bounded within ± 0.03), and none of the tested reliability-based controllers reliably outperforms using the strong selector throughout. Across MuSiQue and 2WikiMultihopQA, structure-free policies also remain on the quality–cost frontier at every tested budget. These findings identify fallback recovery as a key factor in the cost of structural defects and show that the benefit of explicit structure depends on the selector used.

1 INTRODUCTION

A retrieval-augmented system that answers a multi-hop question rarely reads everything it retrieved: a retriever returns candidates, a reader reads a few, and a budget (context length, latency or the price of a call) decides how many. A growing family of systems lets the model impose a *structure* on that decision before reading: it decomposes the question into sub-questions, assigns candidates to them, and reads to cover the structure (Trivedi et al., 2023; Zhu et al., 2025). The structure is a guess: it drops supporting passages, invents sub-questions and misassigns candidates. The obvious engineering response is a controller that trusts the structure when it looks reliable and falls back to a structure-free selector otherwise, in the spirit of routing by question complexity (Jeong et al., 2024) or by learned retrieval mode (Guo et al., 2026).

This paper asks what a defect in such a structure costs. A reliability-tracking controller presupposes that the structure's accuracy determines its *decision value*, the change in final answer quality that following it produces from a given evidence state; that is our motivation, not our claim. A passage the structure omits stays in the candidate pool and may be read during fallback (Figure 1), so we evaluate defects by their effect on final answer quality under different continuation policies.

We do so with a protocol that isolates the reading decision from retrieval recall (Section 2): a fixed candidate set, a reveal budget B, and a frozen model that answers from what was read. From identical evidence states (*same-state counterfactuals*) we run alternative continuation policies to the final answer, varying structure (off or on) and selector (a free lexical ranker or the model naming the next title) independently, with a controlled defect that hides the remaining supporting passages from the structure. On the affected states of a confirmation on 200 fresh, unscreened source components, the same defect costs the lexical selector 0.22 F1 and the model selector 0.01 (0.21 against 0.01

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.

on the sealed split); the path records place the difference in the fallback phase, where the model selector re-reads the hidden passage far more often, and the asymmetry replicates with a second model endpoint and on 100-passage retrieved pools (Sections 3.3 and 5). Structure lifts the lexical selector and adds at most a small effect on top of the model selector.

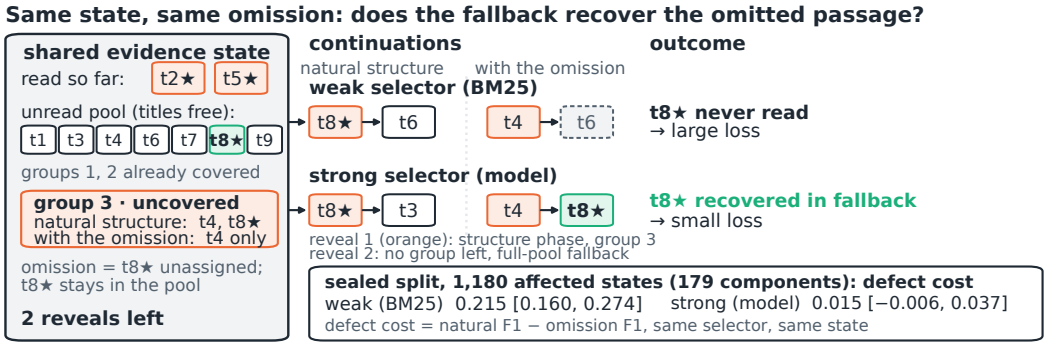


Figure 1: One state, one omission, two selectors (illustrative paths, not a recorded trajectory). The omission removes the unread support from its group only; it stays in the pool, and whether the fallback re-reads it decides the loss. Boxed: sealed-split means on the 1,180 affected states (Section 5). Stars mark gold supports for the reader, never shown to a policy.

Two further findings follow. *Support retention is not decision value*: a gate on the gold label "did the decomposition keep every supporting passage?", the reliability proxy a controller would most want, did not reliably beat a random gate on any of three models, and no gate we ran (gold, visible or bought) reliably beats always letting the model name the next title (Section 3.1). *Per-state routing has a small hindsight gain that visible-feature policies do not recover*: the two continuation policies differ by more than twice the mean repeat noise in about a quarter of decision states, in both directions, yet policies learned on the state's visible features score at or below always-model out of sample (Section 4). Structure also has a price: it is cheaper than the model's continuation only while its decomposition goes uncharged, and structure-free selectors hold the quality–cost frontier at every budget on both tasks and both model endpoints (Sections 3.4 and 5).

Our contribution is empirical and methodological rather than a new architecture: a measurement of what a defect in an inferred structure costs under different selectors, with path records that place the difference in the fallback phase, obtained with an instrument (same-state continuation counterfactuals with a controlled defect, pre-specified rules and independent audits). For system design it says: spend on the selector first, and judge a structure by the outcome of its continuation rather than by its accuracy alone. Section 6 states what this adds to recent work; Section 7 what follows for system design and what the protocol does not license.

2 Setting: Budgeted Reading with an Inferred Structure

Protocol. We adapt MuSiQue (Trivedi et al., 2022) and 2WikiMultihopQA (Ho et al., 2020) as *budgeted reading*. For each question the system sees the question and the titles of the candidate paragraphs supplied with the dataset (twenty for MuSiQue, about ten for 2Wiki, including the gold supports and distractors). Reading a paragraph body costs one unit of the reveal budget $B \in \{1, 2, 4\}$. After the budget is spent a frozen answer model returns a short answer from the paragraphs read, and answer F1 against the gold answer is the primary metric. The candidate set is fixed on purpose: it separates the reading decision from retrieval recall (the stage after retrieval in a reranker-plus-reader pipeline), so a difference between policies is a difference in what they chose to read; corpus retrieval enters only through the retrieved-pool variant of Section 5.

Models and cost. Every policy calls the same frozen model (DeepSeek's `deepseek-flash`, thinking disabled, temperature 0, JSON output), with replications on a larger model of the same family (`deepseek-v4-pro`) and on a second endpoint, a third-party relay that labels its model `kimi-k2.6` (Moonshot); the served model is known only by that returned name (Appendix A, Sec-

tion 5). A policy's *deployment cost* charges every call a stand-alone copy would need, at frozen list prices: the answer call, each guidance call (naming a title, writing a query, rewriting a sub-question) and, for structure policies, the decomposition call. Costs are per question; cache hits between policies and provider prefix caching are not credited (the latter would discount the sequential guidance calls more than the single decomposition call, so the structure-free points are placed conservatively).

Policies. Table 2 (Appendix A) lists the policies with their call counts. *Structure-free* selectors (no explicit decomposition or grouping constraint) pick the next title from the full pool: BM25 over titles with the question and the bodies read so far as the query (Robertson & Zaragoza, 2009), an IRCoT-style policy that asks the model for the next reasoning sentence and retrieves with it (Trivedi et al., 2023), and the strongest, *model-named*, which asks the model to name the next title given the question, what was read and the unread titles. *Structure* policies first call the model once to decompose the question into ordered sub-questions and assign every candidate to one sub-question or to none; *coverage* then reads, for the first sub-question that no revealed paragraph is assigned to (*uncovered*) and that has candidates left, the best BM25 match inside its group, and falls back to the full pool when no group is left. Controls fill the same group sizes at random, rewrite each sub-question with entities from what was read in the manner of ChainRAG (Zhu et al., 2025) with or without restricting the search to the assigned group (*masked*), or alternate one coverage step with one full-pool BM25 step (*alternating*). *Gated* controllers follow coverage while two visible signals (how well the sub-questions anchor to their assigned titles, and whether the last paragraph read matches its sub-question) say the structure is reliable, and switch to model-named selection otherwise; their controls are a random gate matched to the gated controller's development fallback rate, a complexity gate and a diagnostic gate on the gold label. The inferred structure is sparse: it drops a supporting paragraph in 39–40% of MuSiQue questions and, from a decision state, directs about one reading step in five before falling back.

Same-state counterfactuals. A *state* s is a question with its candidate titles, one recorded decomposition, the ordered paragraphs read so far and the remaining budget. From s we execute a continuation policy π (a rule for all remaining reveals) to the final answer and record F1. The *continuation-policy value* of π against a reference policy ρ is

$$\Delta(s; \pi, \rho) = \mathbb{E}\big[\mathrm{F1} \mid s, \pi\big] - \mathbb{E}\big[\mathrm{F1} \mid s, \rho\big],$$

estimated from three independent repetitions of each policy in one session (hosted models are not deterministic at temperature 0; Appendix E measures the drift). Both branches share the state, so decomposition, prefix and budget are fixed by construction and the difference is attributable to the rest of the policy; it answers "which remaining policy is better from here", not the marginal value of the next read alone (a one-step deviation is reported separately). The structure factor is a composite intervention (sub-question queries, groups, visit order, fallback rule) and effects attributed to it belong to that package. Each question contributes two states: the coverage policy's length-one prefix and the model-named policy's length-one prefix, regenerated as its length-two prefix when the two coincide, so prefix lengths one and two are mixed (146 of 300 development states and 1,196 of 3,336 sealed-split states have length two) and the remaining budget is not an exogenous dose. Questions are public MuSiQue train questions, disjoint by source component from every development batch; the development pool keeps the states where the two policies would read different next paragraphs (258 states, 95 components), plus agreeing states as controls; policies that agree now may diverge later: one decision-state population among several. To vary structure and selector independently (Section 3.3) we execute four cells per state: full-pool BM25, coverage with BM25 in the group, full-pool model-named, and coverage with the model naming the title in the group, one targeting rule for both selectors. Two further cells repeat the structure cells with every unread support hidden from the structure, a gold-reading positive control, never a policy (Figure 2). The injected defect is an omission; the commission errors of a natural decomposition, invented sub-questions and misassigned candidates, are present in every natural cell and are not varied.

Statistics and pre-specification. MuSiQue questions are grouped by source component (questions composed from the same single-hop items); we average seeds within a question, questions within a component, and resample components 1,000 times for every interval. Each run had a design document with decision rules fixed before it started and passed an independent replay audit with zero violations (Reproducibility Statement; interrupted runs under a chain audit, Appendix E);

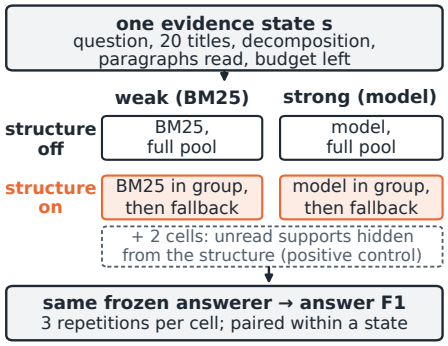


Figure 2: Same-state factorial: four continuations of one state (structure off or on, selector weak or strong) end at the same frozen answerer; two more cells hide every unread support from the structure. Three repetitions per cell, paired within a state.

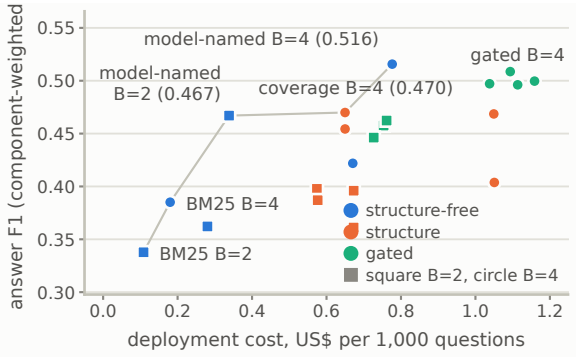


Figure 3: Answer F1 against deployment cost per 1,000 questions on the source-isolated validation set (296 questions, 123 components, `deepseek-flash`), at two (squares) and four (circles) reveals; the gray line joins the Pareto frontier; the random policy (F1 ≤ 0.09) is omitted.

rule outcomes are reported as they fell. Data roles were fixed in advance: 50 first-batch and 403 further questions from the same 35 components for development; 296 questions from 123 untouched components for the gated controller's validation; the public train split for same-state experiments; the sealed split was opened once, after the claims were frozen, for the final evaluation (Section 5).

3 What determines the value of structure

Three results organise Sections 3–5: structure lifts a weak selector and adds at most a small effect on top of a strong one; the same omission costs the two selectors very differently, and the path records place the difference in the fallback phase; no gate tested beats always using the strong selector, and structure-free selectors hold the quality–cost frontier. Table 1 leads with the final evaluation, then follows the order the runs were made.

3.1 Support retention is not decision value

The natural first test is whether structure policies lose where the structure errs. On the development data they do not: in the 403-question extension the coverage-versus-model gap is the same whether the decomposition dropped a support or kept them all (-0.021 against -0.024 at $B = 4$). On the two-hop-dominated validation set the loss *is* concentrated in dropped-support questions (-0.187 against $+0.010$); the relation between structural error and outcome depends on the question population, which already argues against a fixed reliability threshold.

The decisive test is the controller. A gate frozen mechanically from the development signal analysis was run once on the 296 validation questions under three pre-fixed rules. It loses to always-model-named selection (-0.020 and -0.035 on the two models) while costing more, and is indistinguishable from a random gate matched to its development fallback rate on the first model and worse on the second. The diagnostic that matters most is the gold support-retention gate, which switches on the true label "a support was dropped": it is not reliably better than the frozen gate (Table 1), than the random gate or than always using the model (Appendix E). The label is a proxy for one kind of structural error; it says nothing about whether a dropped support is still recoverable from the pool or whether the current decision depends on it, and the model selector is strong either way. The rule class "follow the structure iff the proxy holds" showed no reliable gain even with an oracle proxy; that bounds the rules we ran, not the information in the label.

Table 1: Final evaluation (top) and the comparisons that decided each development step, in the order run (left minus right, answer F1, 95% component-bootstrap intervals); "Outcome" is the pre-specified rule's verdict. Full tables: Appendices C–E.

Step (questions / components)	Comparison (left − right)	B	Δ [95%]	Outcome
Final, sealed split (2,898 states / 386)	interaction, strong − weak	–	-0.057 $[-0.079, -0.038]$	replicates, half-width 0.020
	structure − BM25 (weak selector)	–	$+0.048$ $[+0.030, +0.069]$	helps weak selector
	structure+model − model	–	-0.010 $[-0.018, -0.003]$	within ± 0.03
path-changing (1,180 / 179)	net loss difference, strong − weak	–	$+0.200$ $[+0.145, +0.259]$	recovery supported
Gate batch (50 / 35)	oracle − random	4	$+0.475$ $[+0.340, +0.611]$	signal present
	mask on − mask off	2	$+0.024$ $[-0.017, +0.088]$	harm test did not fire
	coverage − term overlap	2	$+0.071$ $[-0.057, +0.214]$	insufficient evidence
Baselines (50 / 35)	model-named − BM25	2	$+0.114$ $[+0.014, +0.236]$	strongest structure-free
	coverage − model-named	2	-0.037 $[-0.111, +0.014]$	undecided, extend
Extension (403 / 32)	model-named − BM25	2	$+0.131$ $[+0.051, +0.219]$	
	coverage − model-named	4	-0.032 $[-0.081, +0.011]$	undecided (16% cheaper)
	rewrite full − masked	2	-0.027 $[-0.058, -0.001]$	masking helps
Validation, flash (296 / 123)	gated − model-named	4	-0.020 $[-0.049, +0.005]$	empirical line
	gated − random gate	4	-0.003 $[-0.022, +0.013]$	
	gold support-retention gate − gated	4	$+0.020$ $[-0.008, +0.050]$	
Validation, pro (296 / 123)	gated − model-named	4	-0.035 $[-0.067, -0.009]$	empirical line
	gated − random gate	4	-0.026 $[-0.055, -0.001]$	
Validation, relay kimi (296 / 123)	gated − model-named	4	-0.002 $[-0.041, +0.038]$	empirical line
	gated − random gate	4	$+0.008$ $[-0.031, +0.046]$	
	gold support-retention gate − gated	4	-0.016 $[-0.058, +0.020]$	
Same-state (258 states / 95)	structure − model, remaining reveals	–	-0.071 $[-0.117, -0.032]$	action space exists
	hindsight per-state switch gain	–	$+0.028$ $[+0.012, +0.044]$	in-sample reference
	two-repetition switch rule, held out	–	$+0.017$ $[+0.001, +0.033]$	below the 0.02 threshold
	visible-feature policies, held out	–	-0.007 to -0.003	no positive gain
	bought judgement gate − model	–	-0.040 $[-0.079, -0.008]$	signal harmful
Reference matrix (258 / 95)	structure − free BM25	–	$+0.008$ $[-0.017, +0.032]$	not supported
	model pick then structure − model	–	$+0.005$ $[-0.021, +0.031]$	non-inferior, $1.27\times$ cost
Factorial (174 states / 73)	interaction, strong − weak	–	-0.047 $[-0.094, -0.004]$	negative, magnitude open
path-changing (46 / 24)	defect cost, weak selector	–	-0.273 $[-0.428, -0.132]$	control detected
	net loss difference, strong − weak	–	$+0.235$ $[+0.111, +0.367]$	recovery supported
selector-neutral (61 / 31)	net loss difference, strong − weak	–	$+0.144$ $[+0.055, +0.258]$	post hoc
Confirmation v5.3c (662 states / 200)	interaction, strong − weak	–	-0.049 $[-0.080, -0.019]$	H3 met
	structure − BM25 (weak selector)	–	$+0.053$ $[+0.024, +0.083]$	helps weak selector
	structure+model − model	–	$+0.004$ $[-0.005, +0.014]$	within ± 0.03
path-changing (139 / 84)	net loss difference, strong − weak	–	$+0.211$ $[+0.113, +0.305]$	recovery supported
Retrieved pools (305 states / 101)	structure+model − model	–	$+0.000$ $[-0.014, +0.016]$	within ± 0.03
path-changing (61 / 39)	net loss difference, strong − weak	–	$+0.232$ $[+0.110, +0.356]$	recovery supported
2Wiki (150 / 147)	coverage − BM25	2	$+0.007$ $[-0.046, +0.060]$	simple methods suffice
	rewrite full − masked	2	$+0.034$ $[+0.005, +0.071]$	masking hurts here

3.2 Same-state counterfactuals: an action space exists but is hard to see

Figure 5 (Appendix E) gives the continuation values from the 258 decision states. Following the structure for all remaining reveals costs 0.071 F1 against following the model, one structure step then the model 0.012, and stopping 0.134: the remaining budget has real value. 23% of states (59 of 258) differ from the reference by more than twice the mean over states of the paired repeat standard deviation (0.042; 17 favouring structure, 42 against), the pre-declared heterogeneity rule: a common-threshold heuristic, not a per-state test (Appendix E), that says an action space may exist. The positive control (structure with the remaining supports hidden from it) is detected, -0.094; its paired difference against the natural structure, -0.022 $[-0.044, -0.002]$, is small because hiding a support changes the path in only 33 of 258 states, and large there (-0.227).

3.3 Varying structure and selector independently

Structure wins where the model selector scores low (0.26 against 0.75 where structure loses), partly mechanically: with the structure branch fixed, its advantage is ordered by the selectors' scores. We then ran the two-by-two design of Section 2 from the same states, with pre-declared predictions.

Figure 4 (hollow markers) shows the result on the 174 decision states that still had a support to find, beside the confirmation of Section 5 (filled). The selector effect is large under both structure settings ($+0.18$ with structure off, $+0.14$ with it on); the structure effect is $+0.021$ $[-0.019, +0.066]$ under the weak selector and -0.026 $[-0.064, +0.005]$ under the strong one, and their difference, the interaction, is -0.047 $[-0.094, -0.004]$: the structure effect is smaller under the strong selector. A negative interaction says only that; neither simple effect is established here. The effect concentrates where the structure carries evidence. On the 46 states (24 components) where hiding the remaining

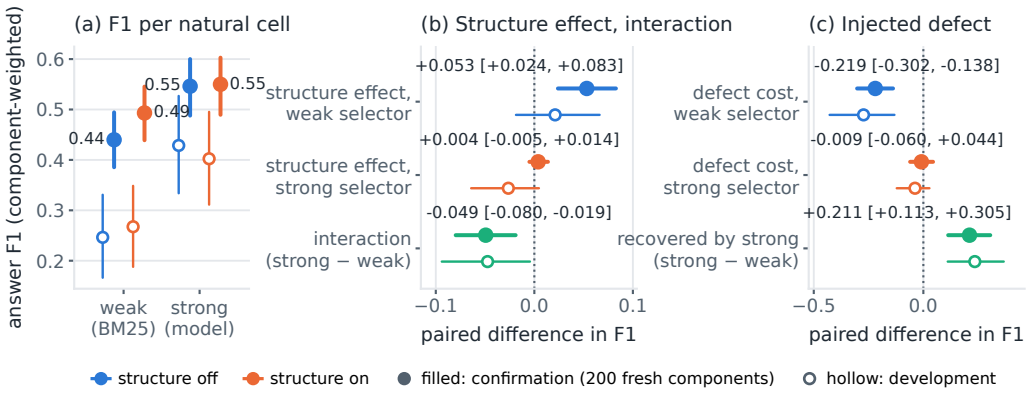


Figure 4: Structure and selector varied independently from the same states; 95% component intervals. Filled: confirmation on 200 unscreened components (662 states; (c) 139 affected). Hollow: development (174 screened states, 73 components; (c) 46). (a) Answer F1 per natural cell. (b) Structure effect per selector and the difference. (c) Defect cost per selector and net loss difference.

supports changes the weak structure policy's path, the natural structure lifts the weak selector by $+0.24$ with no detected effect on the strong one (-0.01 $[-0.07, +0.04]$), and the injected defect costs the weak selector 0.273 F1 $[0.132, 0.428]$ and the strong one 0.038 $[-0.028, 0.122]$; the difference of the two net losses is $+0.235$ $[+0.111, +0.367]$ (strong minus weak; positive when the strong selector loses less). That subset is defined by the weak policy; under selector-neutral definitions, on the 61 states (31 components) where the defect changes either policy's path the difference is $+0.144$ $[+0.055, +0.258]$; on the 49 states where it changes the strong policy's path, $+0.043$ $[-0.040, +0.139]$. The natural baselines are close (0.48 and 0.50), and the path records support the recovery reading: under the defect the strong selector re-reads the hidden supports in the full-pool phase in 71% of cases against 40% for the weak one on the selector-neutral subset (Appendix E gives the weak-defined one). Three recorded features of the paths co-vary with the recovery difference (observed strata, not manipulated conditions): whether the fallback phase is reached (60 of the 61 affected states reach it; the one that does not shows no difference), how much budget the structured phase leaves ($+0.18$ when the natural structure directs fewer than two reads, $+0.02$ when two or more; with two reveals left it never reaches a second group and every difference is zero), and whether the selector finds the passage in the pool, where the two selectors differ. A defect removes a passage from the structured phase, not from the pool; the path records attribute most of its cost to what the rest of the policy fails to recover (Figure 1), without excluding other channels (reading order, distractors read, answerer sensitivity).

3.4 The quality–cost frontier

Figure 3 plots answer F1 against deployment cost on the validation set. The frontier is BM25 at one, two and four reveals and model-named selection at two and four; coverage at four reveals is the only structure point on it, and only just: 0.003 F1 above model-named selection at two reveals at $1.9\times$ its cost, 0.03–0.05 below model-named selection at four reveals at 16% lower cost, and on the frontier in 54% of component draws. Every gated variant, and the IRCoT-style policy at both budgets (0.36 and 0.42 F1), is dominated by a cheaper structure-free point. Under component resampling the BM25 and model-named points at two and four reveals stay on the frontier in at least 75% of draws on every MuSiQue run; coverage membership runs from 93% on the 3–4-hop development extension to 7% on the larger model, and no gated point exceeds 44% on any run (Appendix E): the cheaper structure trade-off belongs to the development questions, not the protocol.

4 What per-state routing could gain: three estimands

By the heuristic of Section 3.2 a controller has a candidate decision in about a quarter of the states. Three estimands with different information sets say what a switch between the two continuation policies could gain; none bounds every controller.

Hindsight, a noisy rule, and visible-feature policies. The hindsight estimand $\mathbb{E}_s[\max\{\Delta(s), 0\}]$, choosing per state from the observed three-repetition means, is $+0.028$ $[+0.012, +0.044]$ over always-model; the positive part of noisy means is optimistic in-sample, so this is a reference for switching by state identity, not a population bound. A rule that chooses on two repetitions and is scored on the third gains $+0.017$ $[+0.001, +0.033]$, below the 0.02 threshold fixed before the run; this is one noisy rule, not the best policy. Policies over the state's 24 visible features (title anchoring, group sizes, last-paragraph match), learned on all but one component and scored on it, gain -0.007 $[-0.014, 0.000]$ for a single-feature threshold, -0.005 for a two-feature conjunction and -0.003 $[-0.020, +0.016]$ for a ridge linear score (best in-sample rule $+0.007$): none of the hindsight gain; richer learners or features might.

What a bought signal buys. Any signal beyond the visible features costs a call. One was tested at the same states and budget: the frozen model sees the observation and its own decomposition and says whether to follow structure or the model selector; its choice selects the recorded continuation of the same repetition. It chose structure in 84% of states, scored -0.040 F1 $[-0.079, -0.008]$ against always-model at $1.37\times$ the cost, and picked the better policy in 45% of the 62 states where they differ; it beats a random gate matched to its own 84% structure rate ($+0.020$ $[+0.001, +0.043]$), so it carries information, but less than its structure bias costs. No signal tested, gold label, visible features or bought judgement, reliably beats always-model (gold gate against it: $+0.001$ $[-0.017, +0.017]$ on the first model, negative on the others; Appendix E); richer signals are untested.

Two cost questions. Given an existing structure, following it costs 29% of the model's continuation (free BM25 steps), 103% once the recorded decomposition is charged (106% component-weighted; two or three remaining reveals), and the one hybrid matching the model's quality, one model pick then structure ($+0.005$ $[-0.021, +0.031]$), costs $1.27\times$ as much. Neither includes the shared prefix; whether to build a structure at all is answered only by the whole-trajectory frontier (Section 3.4); longer trajectories are untested.

5 Robustness across tasks, models and sessions

Second task. On 2WikiMultihopQA, with about ten candidates per question, free BM25 already reaches 0.62 F1 at two reveals and every non-random policy is within 0.70–0.71 at four; structure adds nothing over BM25 ($+0.007$), semantic grouping matters ($+0.50$ over random groups) and the masking effect changes sign: restricting rewritten queries to the assigned group helps on MuSiQue ($+0.027$ $[+0.001, +0.058]$ at $B = 2$ on the development extension; $+0.065$ $[+0.016, +0.117]$ at $B = 4$ on the validation set) and hurts on 2Wiki (-0.034 $[-0.071, -0.005]$ at $B = 2$); at $B = 2$ both intervals exclude zero, the MuSiQue one only just, from 1,000 draws. Task, pool size and hop count are confounded across datasets, so the reversal is an observation, not a mechanism; the same-state defect experiments are MuSiQue-only.

Second model endpoint. The second endpoint is a third-party relay whose served model is known only by the returned name `kimi-k2.6` (Appendix A); "second family" is the relay's label, not verified provenance. A 50-question calibration reproduced every direction of the first endpoint (Appendix E). The one-shot transfer of the gated controller to the validation set, rule unchanged, reproduces the verdict (Table 1; the gate costs $1.47\times$ model-named selection); coverage's F1 matches `deepseek-flash` (0.474 against 0.470) while the model selector gains $+0.065$ $[+0.011, +0.120]$, so coverage trails it by more (-0.107 $[-0.173, -0.044]$) and masking helps again ($+0.076$ $[+0.029, +0.123]$). The factorial with `kimi-k2.6` as selector and answerer on the same 300 states reproduces the defect asymmetry, not the population interaction: the defect costs the weak selector 0.334 $[0.164, 0.510]$ and the strong one 0.117 $[0.018, 0.237]$ (difference $+0.217$ $[+0.041, +0.386]$; $+0.152$ $[+0.019, +0.287]$ selector-neutral; re-read 71% against 34%); no structure effect on the strong selector is detected (-0.009 $[-0.036, +0.015]$); the interaction is -0.014 $[-0.069, +0.039]$ on the primary subset and -0.237 $[-0.406, -0.069]$ on path-changing states. Unlike with `deepseek-flash`, the strong selector here pays a detectable price; re-answering both endpoints' paths with both answerers does not isolate a path or answerer contribution, while the net loss difference is positive in all four cells (Appendix E).

Confirmation on new sources without the decision-state screen. The factorial was repeated on 200 fresh source components (452 questions, 904 states, none screened; pre-specified primary subset: 662 states with a support still unread). The interaction replicates and tightens: -0.049 $[-0.080, -0.019]$. Structure lifts the weak selector by $+0.053$ $[+0.024, +0.083]$ and adds at most a small effect on the strong one, $+0.004$ $[-0.005, +0.014]$, bounded within ± 0.03. The defect costs the weak selector 0.219 $[0.138, 0.302]$ and the strong one 0.009 $[-0.044, 0.060]$ on the 139 path-changing states (84 components), a difference of $+0.211$ $[+0.113, +0.305]$ ($+0.152$ $[+0.065, +0.242]$ under the selector-neutral definition); the strong selector re-reads the hidden supports in the full pool in 90% of cases against 56% (87% against 51% selector-neutral). Two things did not replicate: on the *diverging* states alone no interaction is detected (-0.002 $[-0.037, +0.030]$); it is carried by states where structure and model agree on the next read (-0.068 $[-0.104, -0.036]$), where the structure hands the weak selector the strong one's choice. The recovery difference depends less on the budget the structure used ($+0.16$ against $+0.09$ selector-neutral): of the three recorded features, selector ability and reaching the fallback phase replicate, budget is suggestive.

Retrieved candidate pools. The factorial was also run on 150 further components whose pool is the top 100 paragraphs BM25 retrieves from an 84,559-paragraph train-split corpus (gold recall 0.66). On the 305 states with an in-pool support left to find, structure adds at most a small effect on top of the model selector ($+0.000$ $[-0.014, +0.016]$), lifts BM25 by $+0.043$ $[+0.017, +0.072]$, and the interaction is -0.042 $[-0.079, -0.008]$; on the 61 path-changing states the defect costs the weak selector 0.243 and the strong one 0.011 (difference $+0.232$ $[+0.110, +0.356]$; re-read 82% against 41%). Pool recall stratifies the effects (an observed stratum, not a manipulated one): where the pool holds every support they are present ($+0.062$ $[+0.010, +0.114]$ on 204 states); where it misses one (101 states) the three natural-structure effects lie within ± 0.03 while the two defect effects are imprecise (Appendix E). With 100 candidates the decomposition dominates cost: the structure continuation costs $1.8\times$ the model's with it charged, $0.14\times$ without.

Final evaluation on the sealed split. With every claim frozen, the factorial ran once on the sealed MuSiQue split: 1,668 questions from 386 source components, sharing no single-hop item, question or support passage with the development data, all used, none screened (3,336 states, 60,048 branches, `deepseek-flash`). Five confirmatory checks and one descriptive check were pre-specified (Appendix E); all five held. On the 2,898 states with a support still unread the interaction is -0.057 $[-0.079, -0.038]$, the first run within the pre-specified half-width of 0.03 (0.020); structure lifts the weak selector by $+0.048$ $[+0.030, +0.069]$ and changes the strong one by -0.010 $[-0.018, -0.003]$, within ± 0.03. On the 1,180 path-changing states (179 components) the defect costs the weak selector 0.215 $[0.160, 0.274]$ and the strong one 0.015 $[-0.006, 0.037]$, a difference of $+0.200$ $[+0.145, +0.259]$; the strong selector re-reads the hidden support in the full pool in 81% of cases against 44%. The interaction also appears on the diverging states alone (-0.036 $[-0.059, -0.010]$); on the 47 states (12 components) that never reached the full pool no reliable difference is detected (-0.01 $[-0.08, +0.05]$, the descriptive check; the interval does not bound it).

6 Related work

Inferred structure and its value. Iterative retrieval interleaves reasoning and retrieval (Trivedi et al., 2023); ChainRAG repairs entities lost in decomposition by progressive rewriting (Zhu et al., 2025); both appear here as restricted adaptations with the same information access. RouteRAG learns when to retrieve from text or graph and when to stop, and its ablation shows a stronger dense retriever narrowing the gain of graph retrieval ($+6.4$ to $+3.0$ F1, 3B model) (Guo et al., 2026); a stage-aware study finds decomposition helps reranking and hurts first-stage retrieval (Yin et al., 2026); per-chunk verification breaks multi-hop chains (Kumar, 2026), and candidate masking shows the opposite sign on MuSiQue and the same on 2Wiki. Each finds the value of structure conditional on retriever or stage; none reports the mechanism at a single evidence state (a controlled defect, its effect under two selectors from the same prefix, the recovery path and its conditions), which this paper adds. Rational metareasoning (Hay et al., 2012; Callaway et al., 2018) asks the value-of-computation question; our synthetic study (Appendix B) motivated the coverage policies.

Routing and the information it uses. Adaptive-RAG routes by question complexity before retrieval (Jeong et al., 2024); Self-RAG (Asai et al., 2024) and FLARE (Jiang et al., 2023) let the genera-

tor's own reflection tokens or confidence decide when to retrieve, a question of whether to read at all, where our protocol fixes the number of reads and asks what to read, so a confidence-triggered stopping rule is a complement we do not test. Routing by learned preference (Ong et al., 2025) and cost-sensitive deferral (Mao et al., 2024) supply the standard learners and the estimand distinction we use (post-hoc selection gain $\mathbb{E}[\max(\Delta, 0)]$ against the gain from visible information); a multimodal study finds retrieval signals do not improve query-only routing (Li et al., 2026); a per-query router over commit, resample, retrieve and defer draws its strongest features from executing the interventions themselves (eight calls per query, a cost reported beside the utility it optimises), with the smallest gain on its most capable model (Yin & Zhang, 2026); routing on executed interventions is not new; the bought signal of Section 4 prices one such call inside the deployment cost at fixed states rather than proposing a router.

7 Implications and limitations

The result in one sentence. In this protocol the selector moderates both the value of structure and the cost of its defects: structure lifts the weak selector and adds at most a small effect to the strong one, the same omission costs the strong selector far less, and the path records place that difference in what the fallback recovers; neither follows from the structure's accuracy alone.

For system design. (i) In this protocol, at the budgets tested, spend on the selector first: the model naming the next title dominates structure policies of equal or higher cost, and adding structure to it neither helps nor pays. (ii) A weak, cheap selector gains from structure where it carries evidence the selector would not have found, and the fixed cost must amortise over more than two or three reads. (iii) A support-retention gate did not pay: even an oracle proxy gave no reliable gain over a random gate or the model selector; reachability and recoverability were not tested as gates. (iv) Report structure-guided policies against a model-selector frontier under full call accounting; against BM25 alone with the decomposition uncharged they mislead. (v) The cost measure charges calls, not latency (sub-question retrievals can run in parallel), and does not price hosted-model drift (the decomposition is reproduced in 16–32% of questions across sessions, Appendix E).

What the protocol does and does not license. The candidate set is fixed and small: the conclusions concern the reading decision given retrieved candidates, not retrieval over a corpus; 100 retrieved candidates did not change the result (Section 5); denser or larger pools remain untested. The same-state defect experiments are MuSiQue-only; 2Wiki supplies baselines and the masking observation, not a test of the mechanism; the injected defect is an omission, and commission errors were not injected, only stratified after the fact (Appendix E). MuSiQue is public and may be in the training data of every model used; the recorded branches argue against answer recall (with no gold support read the answerer scores 0.10 F1 and abstains in 70% of branches, against 0.68 with every support read), every mechanism estimate is a same-state contrast within one model, and a selector recognising a title from world knowledge exercises the ability we measure. The controllers run are threshold, conjunction and ridge policies over visible features, a gold-label gate and a bought judgement; trained routers were not run, and any per-state switch between the two continuations is bounded in-sample by the hindsight estimand of Section 4. The development pool is screened to diverging states; the unscreened confirmation reproduces the interaction at the population level but not on its diverging states alone (the sealed split does, at -0.036), so the mechanism is stated as what the structure hands the weak selector rather than as divergence. The affected-state subsets are defined by the policies' own recorded paths (the recovery difference is smaller under a selector-neutral definition). The second endpoint reproduces the gate verdict and the defect asymmetry but not the population-level interaction on the screened states, confirmed only on unscreened states of the first endpoint; its model identity rests on the relay's returned name. An exploratory medium-strength (local 8B) selector failed its pre-specified admission test and is reported in Appendix E only as a bound on the mechanism. Development intervals are wide (95 to 123 components); the sealed split, opened once, puts the interaction below -0.03 and the strong-selector effect inside ± 0.03; hosted-model drift is bounded but not removed.

Reproducibility Statement

Every run in this paper is defined by a frozen configuration and a design document written before the run; the runner records SHA-256 hashes of its sources and inputs, every request with its token usage and returned model identity, and a reservation journal that caps calls, tokens and cost (in the answerer-swap run concurrent clients overwrote the shared journal, and caps and cost were verified from the request logs instead; Appendix E). An independent auditor replays each run from the recorded requests, checks that prompts contain only observations, recomputes every score, interval and rule outcome, and verifies isolation rules. Runners and auditors execute under a fixed Python hash seed so that ties are broken identically. The supplementary material contains the code (data preparation, reveal environments and policies, budgeted model clients, runners and analysis scripts) and the frozen configuration of every run; its README gives the commands that recreate the data and re-run each experiment. Result directories are named in Appendix E; the request logs, audit reports and design documents behind them are not part of the submission package.

AI Usage Statement

An AI assistant (Claude, Anthropic) was used under the authors' direction for research execution (experimental runners, auditors and analysis scripts), for retrieval and discovery (literature search and bibliographic checks), and to draft and revise sections of this manuscript; AI-assisted internal reviews of the literature difference table and of the manuscript are on record. The authors set the research questions, decision rules, data roles and every frozen configuration and cost, and are responsible for all statements.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Frederick Callaway, Sayan Gul, Paul M. Krueger, Thomas L. Griffiths, and Falk Lieder. Learning to select computations. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 776–785. AUAI Press, 2018. URL https://auai.org/uai2018/proceedings/papers/269.pdf.

Yucan Guo, Miao Su, Saiping Guan, Zihao Sun, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. RouteRAG: Efficient retrieval-augmented generation from text and graph via reinforcement learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30042–30059, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1502. URL https://aclanthology.org/2026.findings-acl.1502/.

Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2012. URL https://people.eecs.berkeley.edu/~russell/papers/uai12-meta.pdf.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023. Association for Computational Linguistics.

Randhir Kumar. Verification without sufficiency: Per-chunk filtering fails on multi-hop RAG, and decomposition repairs it. arXiv preprint arXiv:2608.00585, 2026. URL https://arxiv.org/abs/2608.00585.

Qiaomu Li, Qiuyuan Zhang, and Nong Ming. Beyond the query: Do retrieval signals improve adaptive multimodal RAG routing? arXiv preprint arXiv:2609.12437, 2026. URL https://arxiv.org/abs/2609.12437. Preprint dated 2026-09-11; author list taken from the arXiv record.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/mao24d.html.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

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.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Bochao Yin, Xuan Lu, Zhengyu Qi, and Xiaoyu Shen. When should queries be decomposed? a stage-aware study of query decomposition for multi-condition retrieval. arXiv preprint arXiv:2606.08577, 2026. URL https://arxiv.org/abs/2606.08577. Preprint dated 2026-06-07; author list taken from the arXiv record.

Shiyuan Yin and Ruizhi Zhang. Failure-mode-aware uncertainty intervention routing for large language models. *Knowledge-Based Systems*, 351:116685, 2026. doi: 10.1016/j.knosys.2026.116685. URL https://www.sciencedirect.com/science/article/pii/S0950705126014115.

Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, and Wei Hu. Mitigating lost-in-retrieval problems in retrieval augmented multi-hop question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22362–22375, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1089. URL https://aclanthology.org/2025.acl-long.1089/.

A PROTOCOL DETAILS

Table 2: Policies compared. All share the same frozen answer model, candidate set and budget; "calls" counts model calls per question at $B = 4$ besides the answer call (the decomposition counts as one).

Family	Policy	Selection rule	Calls
structure-free	BM25	best BM25 match over titles, query = question + bodies read	0
	IRCoT-style	model writes the next reasoning sentence, BM25 retrieves with it	4
	model-named	model names the next unread title	4
structure	coverage	decomposition; BM25 inside the group of the first uncovered sub-question (no revealed paragraph assigned to it)	1
	random groups	same group sizes, random membership (control)	1
	rewrite (full / masked)	sub-question rewritten with entities read (from the second reveal on); full pool / assigned group	4
	alternating	coverage rule on the first and third reveal, full-pool BM25 on the second and fourth	1
gated	gated	coverage while visible signals pass, else model-named	1–5
	random gate	same fallback rate, random trigger (control)	1–5
	gold retention gate	trigger = "a support was dropped" (gold; diagnostic)	1–5
oracle	oracle	gold supports first (diagnostic ceiling)	0

Reveal episode. The observation available to a policy at any step is the question, the list of candidate IDs with titles, and the bodies of the paragraphs already revealed. Prompts to the model contain only this observation (plus, for structure policies, the model's own decomposition); the audit rejects any recorded prompt that differs from the observation-only construction. Answer scoring uses token-level F1 after lower-casing, punctuation removal and article stripping; support recall is the share of gold supports among revealed paragraphs.

Decomposition. One call asks the model to split the question into ordered sub-questions and to assign every candidate ID to one sub-question or to `null`; an invalid, incomplete or duplicated assignment is a format failure and the question is treated as undecomposed. Structure policies charge this call.

Coverage query. A sub-question counts as covered as soon as any revealed paragraph is assigned to it, whether or not that paragraph answers it; a sub-question whose group has no unrevealed candidate is skipped. Inside the group of the first uncovered sub-question, coverage scores the group's titles by BM25 with that sub-question's text plus the bodies of the paragraphs already read as the query; when no title in the group shares a term with this query, the question plus the bodies read is used instead and the step is recorded as a query fallback. Once no group is left, the full pool is scored with the question plus the bodies read, the structure-free BM25 query. Document frequencies and the average title length come from the question's full candidate set, so scores do not drift as the pool shrinks. The alternating control applies the coverage rule on the first and third of four reveals and scores the full pool on the second and fourth.

Guidance calls. Model-named selection asks for exactly one unread candidate ID; the IRCoT-style policy asks for the next reasoning sentence and uses it as a BM25 query; rewriting asks for a self-contained query for one sub-question given the paragraphs read. A guidance response cut off at the token limit is paid, recorded empty and handled by the policy's parse-failure fallback (BM25 over the question and bodies read).

Gate signals. The frozen gate uses two visible features computed from the decomposition and the observation: the mean lexical anchoring of sub-questions to their assigned titles (post-decomposition gate) and the lexical match between the last revealed paragraph and its sub-question (step gate). Thresholds were selected mechanically on the 403-question development extension by leave-one-component-out switching value and frozen before the validation run; the random gate uses the realised fallback rates (0.62 post, 0.75 step).

Second endpoint: adapter and identity. The second endpoint is a third-party relay that exposes only the Anthropic Messages protocol. The served model is identified solely by the name the relay returns, `kimi-k2.6`, with a null system fingerprint on every response; no credential independent of the relay establishes the upstream model, its version or its update date, so the paper reports a replication on a second endpoint and treats the family label as the relay's. The relay does not honour the request to disable thinking; an assistant prefill of the expected JSON opening suppresses thinking entirely (no thinking block in 6,082 calibration responses) and is part of the recorded model identity. Costs reported for this endpoint (the validation transfer and its frontier) are at Moonshot's list price for `kimi-k2.6`, US$0.95 per million input and US$4.00 per million output tokens, every prompt token, including relay cache reads, at the cache-miss rate; the factorial's manifest carries an earlier accounting assumption (US$0.6 / 2.5 per million; US$1.66 there, US$2.63 at the list price), and no cost from that run is reported. The relay refused every request for three of the 296 validation questions as "high risk" (98 answer requests and 83 guidance requests in total, two of the questions sharing a paragraph about a head of government); they are recorded as empty answers, disclosed with the K1 results, and excluding those questions changes no reported comparison beyond the third decimal.

B THE SYNTHETIC PRECURSOR STUDY

Before the text experiments we studied the internal-computation analogue on synthetic tasks with a known finite answer model, where a planner with at most B questions and a computation budget evaluates two-step question plans and a dependency graph, corrupted on purpose, may guide which plans it evaluates first. Four rounds of controlled experiments gave the following, each against an exhaustive two-step reference: a wrong graph hurts when used to prune plans and still hurts when used only to order evaluations; a learned predictor of which unevaluated plan is worth evaluating did not beat a fixed evaluation order in two attempts; dynamically concentrating computation on the currently best first question failed because it ignored first questions that are worthless alone but valuable as the start of a sequence; reserving a small evaluation batch for every first question before concentrating repaired that failure, lowering realised error from 20.3% to 14.5% on 320 fresh task groups (-5.8 pp $[-9.1, -2.7]$), with the gain confined to task families where the useful first question has no value on its own; replacing the coverage rule's root selection by plain UCB1 gave higher model risk ($+0.044$ $[+0.036, +0.052]$) and an inconclusive realised-error difference ($+2.3$ pp $[-0.5, +5.3]$); and under answer-model misspecification the plan with lower predicted risk could produce more actual errors. The coverage-before-concentration finding motivated the coverage policies of the text protocol; the finding that a wrong structure hurts even when used only to order motivated the reliability gate whose failure Section 3.1 reports.

Tasks, answer model and graph. A task has $k = 4$ independent latent bits (16 equiprobable worlds) and $m = 6$ three-valued queries (answers 0, 1 or unknown), each reading one family-specific binary signal of the world under a random query permutation and random answer flips; the target is a two-valued label of the world. The four families fix the signal set and the target: *direct* (signals $a, b, c, d, b{\oplus}c, c{\oplus}d$; target a), *redundant* (a, a, a, b, c, d; target a), *xor* (a, b, c, d, c, d; target $a{\oplus}b$) and *adaptive* ($a, b, c, d, d, 1{-}d$; target b if $a = 0$ else c), so that in the xor and adaptive families the useful first query has no value on its own. The public answer model assumes answers conditionally independent given the world with a 5% wrong and a 5% unknown rate; the realised channel is noiseless by default (a mild misspecification) and, on the misspecification axis, has its error and unknown rates raised in four levels. The clean priority graph scores each query pair by the normalised expected reduction in Bayes risk from asking the pair under the public model, using no realised answers; the *noisy* corruption sets each pair's score to zero with probability 0.25 (a missing edge), to one with probability 0.20 (a spurious edge) and otherwise jitters it by up to ± 0.12, and the *adversarial* corruption inverts every score. The planner evaluates two-step plans (a first query and a second query conditional on its answer) under a work budget P counted in deterministic operation units, plans from the posterior after the answers actually obtained, and is scored by the realised error of its terminal prediction beside the model risk it planned for; the full two-step reference is the same planner without the work budget. The close-out rounds used 320 fresh task groups at $P = 6{,}081$, a scale sweep over (k, m) cells with the same construction, and a second generator with six latent bits, eight queries, non-uniform product priors and query-dependent noise for the misspecification

axis (64 groups, noisy graph, $P \in \{12{,}000, 24{,}000, 48{,}000\}$). Family labels are never available to a policy.

C External gate: first measurement batch

Protocol `musique-gate-v2`; run directory `results/musique_gate_v2_paid_batch1/` (2,350 episodes, 50 decompositions, 2,400 logical requests; the replay audit passed with zero violations).

Table 3: First MuSiQue batch (50 questions, 35 source components): answer F1 and support recall with 95% component-bootstrap intervals. The batch ran in stages: B1 (oracle, random and lexical: the signal check), B2 (decomposition) and B3 (the structure policies with their own lexical and random controls, executed again in that stage). Every pre-specified pair was computed within one stage (coverage − lexical within B3); rows from different stages are not to be subtracted. All policies use the same frozen answer model; the oracle reads gold and is a diagnostic bound. `mask on/off` share scorer, tie keys and decomposition and differ only in whether unassigned candidates may be revealed.

Stage	Policy	B	Answer F1	Support recall
B1	oracle	2	0.389 [0.245, 0.535]	0.838 [0.771, 0.910]
B1	random	2	0.061 [0.025, 0.105]	0.106 [0.076, 0.137]
B1	lexical (term overlap)	2	0.174 [0.074, 0.288]	0.443 [0.362, 0.529]
B3	lexical (term overlap)	2	0.181 [0.079, 0.310]	0.417 [0.326, 0.507]
B3	random	2	0.058 [0.025, 0.097]	0.106 [0.070, 0.150]
B3	mask on	2	0.190 [0.076, 0.310]	0.525 [0.433, 0.621]
B3	mask off	2	0.167 [0.062, 0.274]	0.436 [0.340, 0.529]
B3	hard prune	2	0.240 [0.117, 0.381]	0.571 [0.477, 0.672]
B3	soft order	2	0.252 [0.126, 0.400]	0.588 [0.492, 0.690]
B3	coverage	2	0.252 [0.126, 0.400]	0.588 [0.492, 0.690]
B1	oracle	4	0.580 [0.439, 0.710]	1.000
B1	random	4	0.104 [0.055, 0.164]	0.188 [0.146, 0.233]
B1	lexical (term overlap)	4	0.260 [0.136, 0.400]	0.615 [0.522, 0.712]
B3	lexical (term overlap)	4	0.231 [0.114, 0.352]	0.612 [0.514, 0.710]
B3	random	4	0.097 [0.043, 0.161]	0.196 [0.152, 0.245]
B3	mask on	4	0.313 [0.186, 0.456]	0.676 [0.579, 0.772]
B3	mask off	4	0.243 [0.124, 0.379]	0.594 [0.502, 0.689]
B3	hard prune	4	0.329 [0.200, 0.471]	0.698 [0.608, 0.788]
B3	soft order	4	0.317 [0.183, 0.464]	0.725 [0.628, 0.815]
B3	coverage	4	0.368 [0.235, 0.515]	0.757 [0.665, 0.844]

Pre-specified rules and outcomes: oracle minus random F1 at $B = 4$ at least 0.10 with interval above zero—passed ($+0.475$ $[+0.340, +0.611]$); decomposition format valid in at least 90%—passed (100%); at least 10% of questions with a gold support assigned to no sub-question—met (38.6%, component weighted); mask on/off trajectories differ in at least 10% of questions at $B = 2$—met (50%); mask on minus mask off F1 interval entirely below zero (harm)—not met ($+0.024$ $[-0.017, +0.088]$); coverage minus lexical F1 interval lower bound above -0.03—not met ($+0.071$ $[-0.057, +0.214]$, paired within stage B3). Verdict: insufficient evidence to expand. Failure to meet a rule is not evidence of equivalence; the intervals are exploratory and uncorrected. The dominant error at every budget is a missing support paragraph followed by abstention; with all supports revealed the model still abstains on 20% of questions, and 79% of questions contain a support whose title shares no token with the question or any gold sub-question.

D External gate: graph-free baselines and mechanism controls

Protocol `musique-baselines-v3`; run directory `results/musique_baselines_v3_paid_batch1b/` (2,250 episodes, 50 decompositions, 3,400 logical requests; the replay audit passed with zero violations). The design, the nine pre-declared comparisons and the decision rule

were written before the run; a first attempt was rerun from cache after a documented amendment. Same 50 questions and model as Appendix C.

Table 4: Second MuSiQue batch: answer F1 with 95% component-bootstrap intervals and per-question deployment cost at frozen peak list prices (US$). `ircot_lite` and `rewrite` are restricted adaptations of the published mechanisms to this title-only protocol, not reproductions.

Policy	$B = 2$ F1	$B = 4$ F1	cost B=2 / B=4
oracle (diagnostic)	0.389 [0.245, 0.535]	0.542 [0.402, 0.676]	0.00012 / 0.00020
random	0.034 [0.012, 0.062]	0.163 [0.088, 0.240]	0.00011 / 0.00018
lexical (term overlap)	0.152 [0.057, 0.267]	0.207 [0.095, 0.336]	0.00012 / 0.00018
BM25 over titles	0.157 [0.059, 0.271]	0.226 [0.112, 0.355]	0.00012 / 0.00019
IRCoT-style query, BM25	0.193 [0.076, 0.321]	0.263 [0.134, 0.400]	0.00031 / 0.00073
model names next title	0.271 [0.143, 0.407]	0.417 [0.280, 0.564]	0.00035 / 0.00081
coverage of sub-questions, BM25	0.235 [0.115, 0.377]	0.346 [0.221, 0.493]	0.00061 / 0.00068
same group sizes, random members	0.034 [0.008, 0.066]	0.155 [0.069, 0.255]	0.00060 / 0.00068
alternating structure / full pool	0.149 [0.056, 0.262]	0.303 [0.187, 0.449]	0.00060 / 0.00068
ChainRAG-style rewrite, full pool	0.179 [0.076, 0.298]	0.285 [0.156, 0.420]	0.00071 / 0.00112
rewrite, assigned group only	0.206 [0.094, 0.341]	0.346 [0.214, 0.484]	0.00071 / 0.00112

Pre-declared comparisons at $B = 2$ (left minus right, F1): BM25 − lexical $+0.005$ $[0.000, +0.014]$; model-named title − BM25 $+0.114$ $[+0.014, +0.236]$; IRCoT-style − BM25 $+0.036$ $[0.000, +0.114]$; coverage − BM25 $+0.077$ $[-0.037, +0.199]$ (support recall $+0.123$ $[+0.046, +0.198]$); coverage − random groups $+0.200$ $[+0.079, +0.339]$; rewrite − BM25 $+0.021$ $[-0.036, +0.100]$; rewrite full pool − rewrite masked -0.027 $[-0.104, +0.038]$; alternating − coverage -0.086 $[-0.195, +0.002]$; coverage − model-named title -0.037 $[-0.111, +0.014]$. Decision rule — rule 1, *simple methods suffice*: a structure-free policy is non-inferior to every structure policy; rule 2, *structure earns its cost*: coverage beats model-named selection at equal or lower cost — outcome undecided (no structure policy has an interval entirely below $+0.03$ against the best structure-free policy at both budgets, and none beats it at $B = 2$); the 403-question development extension runs next. Same-question noise calibration: re-running the oracle with different random filler paragraphs changed 7 of 50 predictions and moved its F1 from 0.585 to 0.544, so differences within about 0.05 F1 on 50 questions are at the noise level. (A later six-wave measurement, below, attributes part of such cross-batch differences to session drift of the answerer itself.)

E External evidence after the first batches

All runs below use the frozen answer model, the title-visible reveal protocol, task-then-component averaging and 1,000 component bootstraps; every run passed its independent replay audit with zero violations, runners and auditors executing under a fixed Python hash seed. Design documents were written and frozen before each paid run; decision rules are reported as they fell. The code and the frozen configuration of every run below are in the supplementary material.

Development extension (403 questions, 32 components). Answer F1 at B=2 / B=4: model-named title 0.285 / 0.390; coverage 0.206 / 0.357; masked rewrite 0.167 / 0.354; BM25 0.154 / 0.230; random 0.046 / 0.091; oracle 0.358 / 0.516. Pre-declared pairs at B=2 (left minus right): model-named − BM25 $+0.131$ $[+0.051, +0.219]$; coverage − BM25 $+0.051$ $[-0.013, +0.115]$ (support recall $+0.133$ $[+0.066, +0.209]$); coverage − random groups $+0.152$ $[+0.073, +0.235]$; rewrite full − rewrite masked -0.027 $[-0.058, -0.001]$; alternating − coverage -0.039 $[-0.078, -0.005]$; coverage − model-named -0.079 $[-0.148, -0.015]$ (at B=4: -0.032 $[-0.081, +0.011]$, cost -16%). Decomposition drops a gold support in 39.9% of questions. Decision rule: undecided (rule 1 missed by 0.004 on one policy). Protocol `musique-baselines-v3`.

Gated controller, source-isolated validation (296 questions, 123 components; two models). Gate frozen mechanically from the extension: post-gate `mean_anchor` < 1.583, step-gate `last_sub_match` < 0.523, fallback to model-named selection; random-gate control at the realized fallback rates (0.62 post, 0.75 step).

Table 5: Gated controller at $B{=}4$, answer F1 with 95% component-bootstrap intervals and deployment cost per question (US$, frozen peak list prices). `gate_oracle` reads gold and is a diagnostic; `gate_complexity` degenerated to always-fallback under the frozen threshold.

Policy	dev 403 (flash)	validation 296 (flash)	validation 296 (pro)	cost flash / pro
model-named title	0.390 [0.279, 0.489]	0.516 [0.442, 0.588]	0.558 [0.483, 0.632]	0.00078 / 0.00332
coverage	0.357 [0.247, 0.460]	0.470 [0.389, 0.550]	0.441 [0.364, 0.522]	0.00065 / 0.00254
gated	0.396 [0.282, 0.497]	0.496 [0.423, 0.569]	0.523 [0.450, 0.593]	0.00111 / 0.00455
random gate	0.398 [0.288, 0.500]	0.500 [0.427, 0.572]	0.548 [0.476, 0.621]	0.00116 / 0.00468
gold support-retention gate	0.369 [0.258, 0.469]	0.516 [0.443, 0.589]	0.507 [0.434, 0.583]	0.00090 / 0.00362
BM25	0.230 [0.138, 0.328]	0.385 [0.312, 0.468]	0.396 [0.319, 0.474]	0.00018 / 0.00078

Pre-declared rules on the validation set (both models): rule 1, gated minus coverage lower bound > 0 at $B{=}4$ — flash $+0.026$ $[-0.024, +0.073]$ fails, pro $+0.082$ $[+0.022, +0.134]$ passes; rule 2, gated minus model-named lower bound > -0.03 and not costlier, or lower bound > 0 — fails on both (-0.020 $[-0.049, +0.005]$; -0.035 $[-0.067, -0.009]$); rule 3, gated minus random gate lower bound > 0 — fails on both (-0.003 $[-0.022, +0.013]$; -0.026 $[-0.055, -0.001]$). Gold support-retention gate minus gated: $+0.020$ $[-0.008, +0.050]$ and -0.015 $[-0.056, +0.029]$; gold support-retention gate minus random gate at $B{=}4$ (offline paired component bootstrap, same interval procedure): $+0.017$ $[-0.002, +0.037]$ flash, -0.041 $[-0.072, -0.011]$ pro, -0.008 $[-0.041, +0.022]$ relay endpoint; minus model-named: $+0.001$ $[-0.017, +0.017]$, -0.051 $[-0.079, -0.024]$, -0.018 $[-0.049, +0.009]$. Outcome: empirical line. On this 90%-two-hop reserved set the coverage loss is concentrated in questions whose decomposition dropped a support (-0.187 against $+0.010$ when all supports were kept, flash), unlike the 3–4-hop extension; the gate removes most of that loss but so does the random gate. Protocol `musique-controller-v4`.

Common-prefix branching (public train split, 100 components, 258 diverging states). Protocol `musique-prefix-branching-v5.1`. From identical states (same question, titles, recorded decomposition, revealed prefix and remaining budget), three independent repetitions of each continuation. States: per question the coverage policy's length-one prefix and the model-named policy's length-one prefix; when both coincide the second is regenerated as the model-named policy's length-two prefix (300 kept states: 124 coverage and 30 model-named at length one, 146 model-named at length two; the same rule gives 452 / 78 / 374 in the confirmation, 333 / 74 / 259 in the retrieved-pool run and 1,668 / 472 / 1,196 on the sealed split), so prefix length and remaining budget are tied to the two policies' agreement on the first read. Relative to "model-named title for every remaining reveal": following the structure for every remaining reveal -0.071 $[-0.117, -0.032]$ at 29% of the continuation cost (about 103% state-weighted, 106% component-weighted, once the recorded decomposition is charged; neither figure includes generating the shared prefix, and the cost ratios are state means while F1 is component-weighted); following it with all remaining supports hidden from the structure (gold positive control) -0.094 $[-0.145, -0.048]$; one structure step then model -0.012 $[-0.035, +0.007]$; stopping -0.134 $[-0.192, -0.081]$. The mean over states of the paired repeat SD is 0.042; 59 states (23%) differ from the reference by more than twice this common threshold, 17 in favour of structure and 42 against (pre-declared rule: at least 20% of states and five in each direction). The threshold is common to all states by design and the rule is a heterogeneity heuristic, not a per-state test: with three repetitions the paired difference is identical across repetitions in 227 of the 258 states, so per-state noise is not estimable, and the verified complementarity is the held-out switch gain below. Naive post-hoc per-state switching would gain $+0.027$ $[+0.012, +0.044]$; selecting on other repetitions and scoring on the held-out one, $+0.017$ $[+0.001, +0.033]$, below the pre-declared 0.02. No visible feature separates the 17 favourable states (AUC $\leq$ 0.62). The positive control's paired difference against the natural structure is -0.022 $[-0.044, -0.002]$: hiding the supports changed the trajectory in only 33 of the 258 states (there it cost -0.227 $[-0.392, -0.061]$); elsewhere the hidden support was already revealed, already unassigned, or not reached within the remaining budget, so most of the control's -0.094 is the natural structure's own deficit.

Selector-strength and first-pick matrix (same 300 states, six continuations, three repetitions). Protocol `musique-prefix-branching-v5.2`; all branches re-executed in one session (a first run was discarded after its audit failed on three tied-BM25 branches; the numbers are from the complete re-run under a fixed hash seed). Answer F1 on the 258 diverging states: model-named every

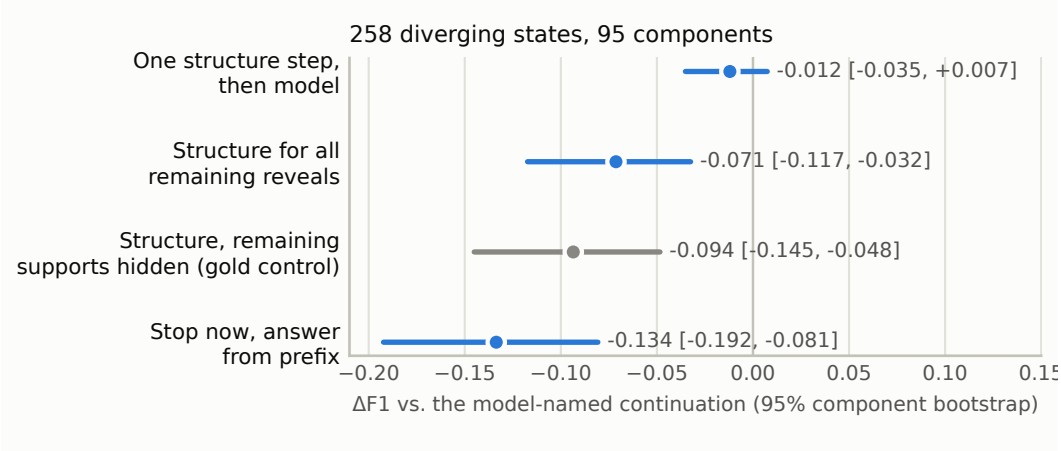


Figure 5: Same-state continuations on the 258 decision states (95 components). Each row is the component-weighted difference in final answer F1 between a continuation and "model-named title for every remaining reveal", with a 95% component bootstrap interval (three independent repetitions per branch). Gray: the positive control, structure followed after every unread support was hidden from it.

step 0.509; one model pick then structure 0.514; one structure step then model 0.500; structure every step 0.434; free BM25 every step 0.425; uniform random 0.412. Pre-declared pairs: structure − BM25 $+0.008$ $[-0.017, +0.032]$ (rule: lower bound > 0, fails); structure − model-named -0.075 $[-0.125, -0.029]$ (rule: upper bound < 0, holds; replicates -0.071 above); one model pick then structure − model-named $+0.005$ $[-0.021, +0.031]$ at $1.27\times$ the deployment cost, because at two or three remaining reveals the decomposition fee exceeds the saved selection calls (rule: cost $\leq 50\%$, fails); one model pick then structure − structure only $+0.081$ $[+0.045, +0.123]$. Neither pre-declared hypothesis (structure value rises as the reference selector weakens; model value concentrates in the first pick at lower cost) is supported. The monotone ordering of the three structure-minus-reference differences holds but is algebraically implied by the references' own scores. The structure continuation was in its structure phase for only 18.9% of steps: inferred assignments are sparse and exhausted after one or two reveals, which bounds any structure effect measurable from these states.

Structure $\times$ selector factorial (same 300 states, six cells, three repetitions). Protocol `musique-prefix-branching-v5.3`, written after the second review of the literature table and frozen before the run. Two factors from identical states: structure off (full-pool selection) or on (the first uncovered sub-question's group, same targeting rule for both selectors, full pool when no group is left); selector weak (BM25, free) or strong (the model names the next title; one call per non-forced reveal). Two further cells repeat the structure-on cells with every unrevealed support hidden from the structure (reads gold). Pre-declared primary subset: diverging states with at least one support still unrevealed (174 states, 132 questions, 73 components). Cell F1: BM25 0.246, structure+BM25 0.267, model 0.429, structure+model 0.402. Structure effect under the weak selector $+0.021$ $[-0.019, +0.066]$, under the strong selector -0.026 $[-0.064, +0.005]$; interaction -0.047 $[-0.094, -0.004]$ (H3, upper bound < 0: passes; precision, half-width ≤ 0.03: not met at 0.045). Interpretation follows the sign of the interval; "no interaction" would require the whole interval inside ± 0.03, which it is not. On the 46 states where hiding the supports changes the deterministic structure path (pre-declared control subset), the defect costs the weak selector -0.273 $[-0.428, -0.132]$ (H4a, at least 0.10 with upper bound < 0) and the strong selector -0.038 $[-0.122, +0.028]$; the difference of net losses, $+0.235$ $[+0.111, +0.367]$ (H4b, lower bound > 0), is compatible with recovery and the path records support that reading: under the defect the strong selector re-reads the hidden supports in the full-pool phase in 65% of cases against 34% for the weak one ($+0.310$ $[+0.145, +0.473]$), while the two natural baselines are close (0.481 and 0.495). On those states the natural structure lifts the weak selector by $+0.238$ $[+0.104, +0.384]$ and leaves the strong one unchanged (-0.013 $[-0.069, +0.044]$). The structure was in its structure phase for 20%

of all steps; on states with two reveals left every difference is zero. Cost: structure+model 0.00108 per branch against model 0.00068. Development data; the follow-up rule asks for a confirmation on new source components without the divergence screen.

Answerer swap (protocol K3, same 46 path-changing states, 24 components). The recorded paths of the `deepseek-flash` and relay `kimi-k2.6` factorials were re-answered by both answerers in one session. Defect cost under the strong selector by (paths, answerer): `flash` paths with the `flash` answerer -0.052 $[-0.129, 0.000]$, `flash` paths with `kimi` -0.035 $[-0.134, +0.061]$, `kimi` paths with `flash` -0.020 $[-0.145, +0.083]$, `kimi` paths with `kimi` -0.107 $[-0.229, -0.006]$; the net loss difference is positive with its lower bound above zero in all four cells ($+0.207$, $+0.323$, $+0.239$, $+0.251$), so the asymmetry survives swapping either the paths' endpoint or the answerer. The pre-declared decomposition into a path effect and an answerer effect has intervals containing zero in both orders, so no single attribution is made. Re-answering the `flash` paths with `flash` reproduced the recorded F1 within 0.000 $[-0.005, +0.006]$ on these states.

Second-endpoint calibration (50 questions, relay `kimi-k2.6`). Before the one-shot transfer, a 50-question calibration reproduced every direction of the first endpoint: a stronger selector (model-named $+0.151$ $[+0.040, +0.263]$ at $B = 2$), an equal answerer, coverage trailing by more (-0.12 against -0.07), masking not hurting, and the frozen gate below a random gate (-0.076 at $B = 4$).

Second task (2WikiMultihopQA-20K dev, 150 questions, 147 components). Protocol `twowiki-baselines-v3`; about ten candidates per question. F1 at B=2: BM25 0.615, coverage 0.622, model-named 0.658, IRCoT-style 0.663, rewrite full 0.656, rewrite masked 0.622; at B=4 every non-random policy is within 0.698–0.713 (oracle 0.719). Coverage − BM25 $+0.007$ $[-0.046, +0.060]$; rewrite full − rewrite masked $+0.034$ $[+0.005, +0.071]$ (masking hurts here); coverage − random groups $+0.497$ $[+0.416, +0.574]$. Decision rule: simple methods suffice.

Answer and decomposition drift across sessions. Protocol `musique-answer-drift-v1`. The 50 decomposition prompts and 179 answer prompts of the second batch were re-issued verbatim in six waves 90 minutes apart under the same alias, fingerprint and temperature 0. Per wave, 10–13% of answers differed from the original recording and 7–8% from the first wave; component-weighted F1 moved by at most 0.022. Of 179 prompts, 148 never changed, 9 changed in all six waves and 22 flickered; 18 flipped between an answer and abstention. Decompositions reproduced the original assignment in only 16–32% of questions per wave, while the gold-side flag "a support was dropped" flipped in 2–6%. All paired comparisons in this paper are therefore within-session; cross-batch differences of a few points are not interpreted.

Quality–cost frontier. Five runs are covered (`quality_cost_frontiers_2026-09-17.json`, `frontier_membership_bootstrap_2026-09-17.json` and `frontier_membership_bootstrap_kimi_k1_2026-09-17.json`): the MuSiQue development extension (`deepseek-flash`), the validation set with `deepseek-flash`, `deepseek-v4-pro` and `kimi-k2.6`, and 2Wiki (`deepseek-flash`). On every run the point-estimate frontier of answer F1 against deployment cost per question contains BM25 at one, two and four reveals and model-named selection at two and four (on 2Wiki at four only; the cheaper lexical and random selectors join it at some budgets). Structure points on the point frontier: coverage at four reveals on the development extension and on the flash validation set (there 16% cheaper than model-named at four, 0.03–0.05 F1 lower), and the alternating structure/full-pool hybrid at four reveals on the development extension and on 2Wiki; no structure point is on it for `deepseek-v4-pro` or `kimi-k2.6`, and no gated point is on it for any run (Figure 3 shows the flash validation set). Resampling source components 1,000 times and recomputing the frontier: at two and four reveals the BM25 and model-named points stay on it in at least 75% of draws on every MuSiQue run (mostly above 95%; the model-named point at one reveal in at most 8%); coverage at four reveals stays on it in 93% of draws on the 3–4-hop development extension, 54% on the flash validation set, 7% on `deepseek-v4-pro`, 9% on `kimi-k2.6` and 36% on 2Wiki; the hybrid at four in 62% on the development extension, 38% on 2Wiki and at most 18% elsewhere; the most frequent point of the gated family reaches 26% on the flash validation set (repair gate at four), 7% on `deepseek-v4-pro` (random gate) and 44% on `kimi-k2.6` (frozen gate at four).

These are resampling frequencies on the present samples, not probabilities for a population frontier; the cheaper structure points were frequent only on the development extension and are sensitive to sample composition elsewhere.

Retrieved pools (protocol `musique-prefix-branching-v5.4-open`). 150 train-pool components disjoint from every earlier run; 333 questions after dropping 12 whose pool held no support; 666 states, none screened. Pool: the top 100 paragraphs by BM25 (title and body, question as query) from the 84,559 unique paragraphs of the train pool, shuffled into fresh IDs; gold supports mapped by text; recall 0.66 (first-hop supports in the pool 95%, second-hop 39%). Two declared protocol changes: decomposition max tokens 1,200 to 3,000 and a parser that leaves unlisted candidates unassigned (never triggered: all 333 decompositions listed all 100 candidates, none failed; the decomposition assigned 6.2 candidates per question on average and dropped an in-pool support in 12% of questions). Primary subset improvable (305 states / 101 components): structure − BM25 $+0.043$ $[+0.017, +0.072]$, structure+model − model $+0.000$ $[-0.014, +0.016]$, interaction -0.042 $[-0.079, -0.008]$ (half-width 0.035); weak-policy path-changing states (61 / 39): defect cost -0.243 $[-0.354, -0.137]$ weak, -0.011 $[-0.051, +0.021]$ strong, difference $+0.232$ $[+0.110, +0.356]$; selector-neutral union (81 / 48) $+0.135$ $[+0.011, +0.257]$; strong-policy path-changing (52 / 35) -0.010 $[-0.133, +0.119]$. Recall strata: full (204 / 69) interaction -0.063 $[-0.114, -0.015]$, difference $+0.062$ $[+0.010, +0.114]$; partial (101 / 40): structure − BM25 $+0.006$ $[-0.009, +0.024]$, structure+model − model $+0.009$ $[-0.004, +0.026]$, interaction $+0.003$ $[-0.010, +0.017]$ and weak defect cost -0.009 $[-0.023, 0.000]$, each within ± 0.03; strong defect cost -0.029 $[-0.100, 0.000]$ and net loss difference -0.021 $[-0.083, +0.015]$, imprecise. Recall is an observed stratum of the retrieved pool, not a manipulated factor. Continuation cost per branch excluding the shared prefix: BM25 0.00018, coverage 0.00220 (decomposition charged), model-named 0.00125, coverage+model 0.00306 USD. The run was resumed once from its cache after an unreadable response; the chain audit passed.

Commission errors of the natural structures (observational, post hoc). The injected defect is an omission; the natural decompositions also commit errors, and every natural cell already contains them. From the recorded decompositions (evaluator-side, gold-derived, never shown to a policy) we read three: a non-support candidate assigned to a sub-question (*misassigned*), more sub-questions than gold hops (*extra*), and two hops' supports in one slot (*merged*, among decompositions that keep every support). Prevalence per decomposition, development / confirmation / retrieved pools: misassigned ≥ 1 in 52% / 50% / 69% (on average 2.0 of 3.8, 1.5 of 3.3 and 5.5 of 6.8 assigned candidates); extra in 7% / 15% / 19%; merged in 27% / 21% (no per-hop record in the retrieved-pool run); a support dropped in 40% / 35% / 12%. In the weak structure cell the structure phase read at least one non-support in 35% / 24% / 40% of improvable states (0.37, 0.26 and 0.43 of its 0.68, 0.74 and 0.74 reveals per branch). Stratifying the improvable subset by these recorded properties (observed strata, not manipulated factors; the usual aggregation, with one joint component bootstrap for each between-stratum difference): with against without a misassigned candidate the structure effect under the weak selector differs by $+0.050$ $[-0.024, +0.125]$, -0.028 $[-0.082, +0.024]$ and $+0.047$ $[+0.015, +0.083]$, no consistent direction, and under the strong selector by -0.028 $[-0.090, +0.020]$, $+0.010$ $[-0.007, +0.028]$ and -0.001 $[-0.019, +0.016]$. Merged hops remove the weak selector's lift in the confirmation run (-0.034 against $+0.105$; difference -0.138 $[-0.208, -0.082]$; development -0.105 $[-0.263, +0.054]$), mechanically: one read covers the merged group and the structure phase ends; the strong selector's effect differs by -0.006 $[-0.030, +0.017]$ in the confirmation run and -0.104 $[-0.223, -0.021]$ in development. Extra sub-questions lower the strong selector's structure effect in the confirmation run (-0.038 against $+0.011$; -0.049 $[-0.103, -0.013]$) but not in the retrieved-pool run (-0.004 $[-0.029, +0.022]$; development has 20 such states). Where the structure phase read a non-support (a path stratum), the weak selector's structure effect is lower by -0.042 $[-0.118, +0.044]$, -0.060 $[-0.107, -0.012]$ and -0.032 $[-0.074, +0.008]$, the strong selector's by -0.068 $[-0.168, +0.004]$, -0.002 $[-0.030, +0.028]$ and -0.029 $[-0.074, +0.010]$. On the path-changing states the defect asymmetry is present with and without misassigned candidates (confirmation net loss difference $+0.27$ $[+0.10, +0.44]$ and $+0.19$ $[+0.07, +0.31]$; development $+0.22$ and $+0.26$); these strata are small and their differences imprecise. None of this is a causal cost of commission: decomposition quality co-varies with question difficulty. Design `docs/commission-observational-analysis-design.md`; out-

puts `commission_observational_v53_dev.json`, `_v53c_confirm.json`, `_v54_open.json` under `research/results/`.

Medium-strength selector (protocol `musique-prefix-branching-v5.3-medium`, exploratory). Same 300 states and recorded decompositions as the factorial; three selector levels in one session: free BM25 (weak), `llama3.1:8b-instruct-q8_0` served locally by Ollama 0.34.2 (greedy decoding, JSON-constrained output, 8,192-token context window; medium) and `deepseek-flash` (strong); the answerer is `deepseek-flash` throughout. Nine cells $\times$ 3 repetitions (a local arm runs the six factorial cells with the 8B model naming titles, a flash arm the three model-selector cells), 8,100 branches, one reservation journal per client, audit passed with zero violations. The local model's outputs were identical across repetitions for 98.8% of prompts (near-deterministic, not deterministic) and named an unlisted or unparsable ID in 0.85% of calls (protocol fallback). Pre-specified admission: the strong-versus-weak reading had to replicate H3, H4a and H4b in-session; H4a (-0.281 $[-0.441, -0.136]$) and H4b ($+0.252$ $[+0.121, +0.389]$) replicated, H3 did not (-0.028 $[-0.073, +0.013]$; development -0.047), so the remaining readings are descriptive. On the 46 path-changing states (24 components): net defect cost -0.281 (BM25), -0.144 $[-0.242, -0.052]$ (8B), -0.029 $[-0.091, +0.024]$ (flash); hidden supports re-read in the full pool under the defect 0.342, 0.595, 0.653; adjacent differences of net loss $+0.137$ $[-0.007, +0.288]$ (8B minus BM25) and $+0.115$ $[+0.028, +0.222]$ (flash minus 8B), of re-read share $+0.253$ $[+0.095, +0.418]$ and $+0.058$ $[-0.021, +0.141]$. On the 174 primary states (73 components) the structure effect is $+0.017$ $[-0.026, +0.063]$, $+0.033$ $[+0.013, +0.053]$ and -0.011 $[-0.036, +0.011]$ for the three levels, the structure-free continuations score 0.248, 0.302 and 0.414 F1, and the 8B and flash selectors agree on their first pick in 39% of states. Local inference is priced nominally at a hosted list price (US$0.18 per million tokens) only to keep costs comparable. Development data; an exploratory boundary on the mechanism, not a confirmed dose–response.

Final evaluation on the sealed split. Protocol `musique-prefix-branching-v5.3-final`. The sealed split is the authors' held-out part of the annotated MuSiQue development set, fixed with the data roles on 2026-09-16 and hashed in the processed-data manifest: 1,668 questions from 386 source components (879 two-hop, 510 three-hop, 279 four-hop), each with the 20-candidate pool of the development protocol. A disclosure report generated before the run and hashed by it (`sealed_split_disclosure_2026-09-18.json`) records its isolation from every development question: no shared single-hop source item, no identical normalised question, no shared support-passage text; 2,713 of its 15,468 candidate paragraphs, almost all distractors, also occur in the train pool. The runner refuses a configuration without the key-holder's authorisation token and the loader re-verifies the file hashes before reading. The protocol is that of the confirmation run: one new paid decomposition per question, two candidate states per question (all kept: 1,668 coverage prefixes of length one, 472 model-named prefixes of length one and 1,196 of length two; the divergence screen recorded only as a stratum label), six cells, three repetitions in one session, `deepseek-flash` throughout, 1,000 component-bootstrap draws. Five confirmatory claim checks (S1–S5) and one descriptive check (S6) were written down before the run with their subsets and rules; all five confirmatory checks held. (S1) On the 1,180 path-changing states (179 components) the defect costs the weak selector -0.215 $[-0.274, -0.160]$ (rule: at most -0.10 with the upper bound below zero) and the net loss difference is $+0.200$ $[+0.145, +0.259]$ (lower bound above zero). (S2) Under the defect the strong selector re-reads the hidden support in the full pool in 0.811 of cases against 0.439, a difference of $+0.372$ $[+0.309, +0.433]$. (S3) On the 2,898 states with a support still unread, structure on top of the model selector is -0.010 $[-0.018, -0.003]$, inside ± 0.03. (S4) Structure lifts the weak selector by $+0.048$ $[+0.030, +0.069]$. (S5) The interaction is -0.057 $[-0.079, -0.038]$ with half-width 0.020, within the pre-specified 0.03 for the first time (development 0.045, confirmation 0.030). (S6, descriptive) On the 47 path-changing states (12 components) that never reached the full pool the net loss difference is -0.01 $[-0.08, +0.05]$, against $+0.200$ on the 1,133 that did. Strata: the interaction is -0.036 $[-0.059, -0.010]$ on the 1,599 diverging states with a support unread, -0.074 $[-0.100, -0.047]$ on the 1,384 control states and -0.041 $[-0.056, -0.027]$ on all 3,336; the net loss difference is $+0.153$ $[+0.107, +0.198]$ under the selector-neutral definition (1,413 states), $+0.107$ with two reveals left against $+0.205$ with three, and $+0.200$ against $+0.142$ where the natural structure used fewer than two against two or more reveals. Cell F1 on the 2,898 states: BM25 0.432, coverage 0.479, model-named 0.531, structure+model 0.522, corrupted coverage 0.422, corrupted structure+model 0.517; every natural cell is 0.01–0.03 below the confirmation run (the sealed split has more three- and four-hop

questions) while every paired difference is unchanged to within its interval. The run cost US$12.30 at frozen list prices for 65,083 paid calls across eight segments (below). The sealed split has been read for no other purpose.

Result directories. Under `research/results/`: development extension `musique_baselines_v3_dev_extension_b`; gated controller `musique_controller_v4_dev403_merged`, `musique_controller_v4_validation_merged`, `musique_controller_v4_pro_merged`; same-state branching `musique_prefix_v51_dev`, `musique_prefix_v52_dev_b`, `musique_prefix_v53_dev`, confirmation `musique_prefix_v53c_confirm`, retrieved pools `musique_prefix_v54_open`; final evaluation on the sealed split `musique_prefix_v53f_final` with its segments `_aborted_a` to `_aborted_g` and `sealed_split_disclosure_2026-09-18.json`; medium-strength selector `musique_prefix_v53m_medium`; second endpoint (relay `kimi-k2.6`) `musique_controller_v4_kimi_merged`, `musique_prefix_v53_kimi_b` and the answerer swap `musique_answer_swap_k3_merged`; bought-signal gate `paid_signal_gate_g1`; second task `twowiki_baselines_v3_batch1b`; drift `musique_answer_drift_v1`; frontier `quality_cost_frontiers_2026-09-17.json` and `frontier_membership_bootstrap_2026-09-17.json`. Each directory holds the frozen configuration, request log, run manifest with source hashes, summary and audit.

Interrupted and resumed runs. Four runs were interrupted and resumed from their caches under a frozen resume configuration (same journal and cache root, new output directory): the second-endpoint factorial (`musique_prefix_v53_kimi`, aborted; `musique_prefix_v53_kimi_b`, final), shards 1–3 of the answerer swap (`musique_answer_swap_k3_shard1` to `_shard3`, aborted when the relay balance ran out; the `_c` directories, final), the retrieved-pool run (`musique_prefix_v54_open_aborted_b`; `musique_prefix_v54_open`, final) and the sealed-split final evaluation (`musique_prefix_v53f_final_aborted_a` to `_aborted_g`, aborted on transport errors and on two locks of the reservation journal; `musique_prefix_v53f_final`, final). For the factorial, the retrieved-pool run and the final evaluation the applicable audit is `audit_chain.json` in the final directory: across the aborted and resumed logs it checks that every (phase, repetition, request) key of the resumed run was paid at least once (the logs show exactly once: 4,766, 11,554 and 65,083 keys), that every replayed request matches the recorded response and token usage, that the aborted rows are a prefix of the resumed rows and that the configurations differ only in the resume fields; the single-log `audit.json` of these three directories reports the repetition-independence check as violated because the paid records sit in the earlier log, and the chain audit supersedes it. For the answerer swap the same cross-log checks are part of the merged run's `audit.json` (passed: every (answerer, repetition, request) key paid exactly once across the aborted and resumed shard logs). Two further first attempts were discarded and restarted from scratch. In the answerer swap the runner built one client per (answerer, repetition), and the concurrent clients overwrote the shared reservation journal, so its in-run cap was only partly effective; caps and cost were verified from the request logs (every chain and answerer below its shard cap: at most US$0.33 for `deepseek-flash` and US$0.99 for `kimi-k2.6` per shard against caps of US$3 and US$7).