TWO CONTEXT REGIMES WITH OPPOSITE SIGNS: MEASURING THE RETRIEVAL TAX ON NATURAL QUESTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieved context can raise average question-answering accuracy while losing answers that a model gives without documents. We measure both directions on 468 Natural Questions examples with a retrieval tax and a rescue rate, each paired against a closed-book cell under the same prompt contract, the system instruction that states how documents may be used. Under exact dense retrieval over 2.68 million passages, five passages lose 14.0% of closed-book-correct answers and repair 49.1% of the incorrect ones. Inside the same kind of genuine rankings, replacing every answer-bearing passage by the next non-answer passage raises the tax from 0.133 to 0.593 for Qwen3-8B in preregistered within-question runs, and by 0.33 to 0.46 on four models and two contracts; blinded semantic labels that adjudicate each question's reference once for all its answers give $+0.313$, with identification bounds above zero. Masking only the gold-answer spans in the unchanged passages produces about half of the containment-scored increase on two models and, under semantic labels, exceeds a matched non-answer mask on complete cases and on questions with an adequate reference; both edits keep their effect at a fourfold reasoning budget scored on the full answer. Uninformative placeholder context costs 43–62% of known answers, but Qwen and Mistral decline by citing the context while Llama-3.1-8B mostly answers "I don't know", so on identical outputs two refusal detectors rank Llama-3.1-8B lowest and highest in known-question over-refusal.

1 INTRODUCTION

Retrieval-augmented generation (RAG) gives a language model access to non-parametric evidence (Lewis et al., 2020; Guu et al., 2020; Karpukhin et al., 2020). Its benefit depends on the retrieved context. Retrieval can hurt, a mean can hide per-question wins and losses, and retrieval-augmented models can over-refuse questions they know (Mallen et al., 2023; Cuconasu et al., 2024; Yang et al., 2026; Zhou et al., 2026); robustness metrics, audits of broken answers and refusal scores each depend on how they are anchored and detected (Park & Lee, 2024; Yoran et al., 2024; Chen et al., 2024). We ask what the anchor, the context construction and the measurement instrument do to the conclusion: under which construction the sign of the effect changes, whether the harm takes the form of abstention or of a corrupted answer, and whether the prompt, the correctness metric or the refusal detector changes the conclusion drawn from the same saved outputs.

We answer it with three measurements on one population of 468 NQ questions, summarized in Table 1, each under stated prompt *contracts*: system instructions that say whether and how the documents may be used. Each comparison is paired within the setting that produced it, because absolute exact match (EM) is sensitive to checkpoint, prompt wrapper, answer extraction and question set.

We define the **retrieval tax** as the fraction of questions answered correctly in the matched closed-book cell that become incorrect with documents, and the **rescue rate** as the fraction of closed-book-incorrect questions that become correct. With closed-book accuracy a, accuracy with documents is $A_t = a + (1 - a)\rho - a\tau$, so the two rates carry different denominators and both are needed to read a change in the mean. We use these conditional rates, as in per-question accounting (Yang et al., 2026; Wang et al., 2023), to read answer-string interventions and a same-output audit. Anchoring

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.

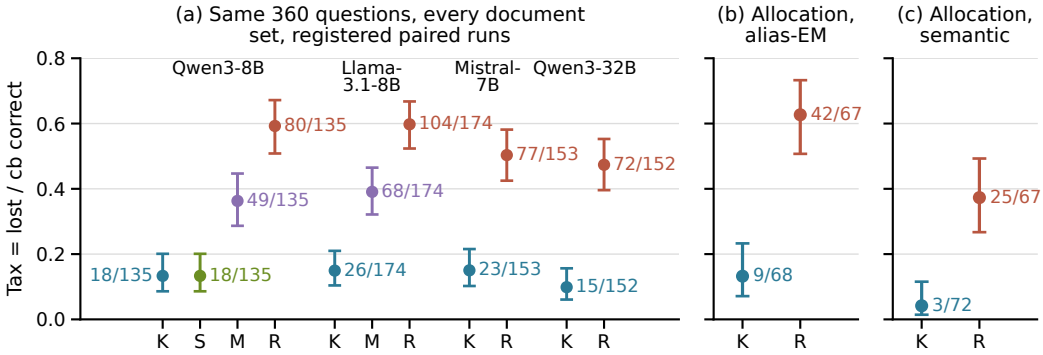

Figure 1: Replacing or masking the answer-bearing passages of genuine retrieved rankings under the neutral contract, 360 eligible questions, Qwen3-8B unless labelled. Points are the tax, lost answers over closed-book-correct questions; whiskers are 95% Wilson intervals. K, S, M, R: keep, matched non-answer mask, answer mask, removal. (a) Every question run with each document set under a registered protocol, for four models; alias-EM, paired differences $+0.459$, $+0.448$, $+0.353$ and $+0.375$. *Mask*: the original passages with every gold-alias span masked, $+0.230$ and $+0.241$; S masks matched non-answer spans instead. (b) Seeded 180/180 allocation, alias-EM. (c) The same allocation on the complete cases of two blinded runs of one LLM annotator, the pairs both runs label alike; of the 75 closed-book answers agreed correct per arm, 3 in keep and 8 in removal lack an agreed with-document label and are excluded. Section 4.4 and Table 17 give details.

Table 1: The three measurements on the same 468 NQ questions; the removal control uses the 360 whose top five hold a replaceable answer-bearing passage. The four configurations are Qwen3-8B, Qwen3-32B, Llama-3.1-8B and Mistral-7B; every tax and rescue is paired against the closed-book cell of the same run and contract. *P3 for Qwen3-8B only. †Answer spans on Qwen3-8B and Llama-3.1-8B; matched non-answer spans on Qwen3-8B.

Setting	Models	Contracts	Context manipulation	Records
Ladder, §4.1–4.3	all four	P2, P3	none, placeholder, recall 30/60%, oracle	$4 \times 5{,}616$
Removal control, §4.4	all four	P2, P3*	genuine top five kept, replaced, masked†	936; 720/arm
Factorial, §4.5–4.6	Qwen3.5-9B	P1–P4	exact dense top $T \in \{0, 1, 5\}$ of 2.68M	5,616

at the matched closed-book cell counts the answers a model supplies without documents, which a retrieval-anchored retention cannot see, as Appendices E and J show. Correctness is scored by strict short-answer EM and by a containment-based alias-EM, the removal control also by blinded semantic annotation, and refusal by two lexical detectors.

Our contributions are:

- **Answer-string interventions inside genuine rankings.** Replacing every answer-bearing top-5 passage by the next non-answer passage of the same ranking, with document count, contract and model fixed, raises the containment-scored tax from 0.133 to 0.593 on the same 135 Qwen3-8B questions in registered within-question runs, a difference of $+0.459\,[0.370, 0.548]$; it rises by $+0.448$, $+0.353$ and $+0.375$ on Llama-3.1-8B, Mistral-7B and Qwen3-32B, and by $+0.331$ under the permissive contract; blinded semantic labels under a registered adjudication of each question's reference give $+0.313\,[0.224, 0.408]$ on Qwen3-8B. Masking only the alias spans produces about half the containment-scored effect on Qwen3-8B and Llama-3.1-8B and about a third on complete cases under adjudicated labels, while an approximately matched non-answer mask that preserves answer support leaves the Qwen3-8B tax at its keep value; the replacement effect persists within strata of length and relevance change, and both edits keep their effect with a fourfold reasoning budget.
- **A same-output audit of three measurement choices.** The contract alone moves the closed-book baseline by 0.38 alias-EM, and a retrieval-anchored retention (Park & Lee,

2024) of 0.88–0.93 answers a different question than the 15–21% of closed-book-correct answers already lost at the retrieval anchor. Strict EM records a significant decline for the document-only contract, 11 gains against 38 losses, that containment scores as retained, because the answers lengthen around the gold string.

- **The form of the loss decides the detector comparison.** Placeholder context costs 43–62% of closed-book-correct answers in every configuration; Qwen and Mistral runs decline by citing the context, while Llama-3.1-8B mostly declines generically, so Llama-3.1-8B has the lowest known-question over-refusal under one detector, 0.020, and the highest under the other, 0.446; blinded semantic labels side with the broader rule.

2 Related Work

Retrieval that hurts. Retrieval-augmented generation supplies non-parametric evidence to a language model (Lewis et al., 2020; Karpukhin et al., 2020), and this evidence can also reduce answer quality. Irrelevant or distracting context degrades reasoning (Shi et al., 2023; Yang et al., 2025); long inputs are used unevenly (Liu et al., 2024), and input length alone costs accuracy even with perfect retrieval (Du et al., 2025). For open-domain QA, retrieval can hurt on questions about popular entities that the model already answers (Mallen et al., 2023; Maekawa et al., 2024), highly ranked non-answer passages can harm while random ones sometimes help (Cuconasu et al., 2024), and a reproduction shows that the latter effect appears or disappears with the prompt and decoding limits (Mazuryk et al., 2026). Amiraz et al. (2025) score the distracting effect of single irrelevant passages and fine-tune on hard distractors; CUB (Hagström et al., 2026) benchmarks context-manipulation methods under relevant, irrelevant and conflicting contexts and finds synthetic contexts easier than natural ones. Wu et al. (2024) elicit closed-book answers and show that irrelevant inputs mislead them or induce uncertainty, mostly with option-based evaluation, and Yeh & Li (2026) find that instruction-tuned models abstain under irrelevant context on questions they otherwise answer. We neither score distractors nor evaluate methods: we replace or mask the answer-bearing passages of genuine rankings with the slot count fixed, pair each question with its own closed-book cell on four models, audit free-form outputs, and show that the *form* of an abstention, context-citing or generic, decides how lexical detectors rank the models. Robustness methods filter context with an NLI model and back off to the no-retrieval answer, or fine-tune on mixed context (Yoran et al., 2024).

Per-question measurement. Per-question accounting has conditioned on retrieval-correct answers (RAD; Park & Lee, 2024), on win/lose transitions and closed-book knowledge states (SURE, SKR; Yang et al., 2026; Wang et al., 2023; Zhou et al., 2026), on context sufficiency (Joren et al., 2025), or on a manual audit of retrieval-induced errors (Yoran et al., 2024). Unlike these, (i) the answer-bearing passages are replaced, or their alias spans masked, *inside* genuine rankings, by a seeded allocation and within the same questions on four models and two contracts, with document count fixed as in Section 4.4, rather than stratified post hoc or varied across document categories, and read with lexical and blinded semantic labels. (ii) On identical outputs we show how the closed-book anchor, strict versus containment scoring and the refusal detector change a conclusion.

Measuring correctness and refusal. Lexical matching misjudges many LLM answers (Kamalloo et al., 2023), and prompt formatting alone moves accuracy (Sclar et al., 2024). RGB scores rejection by a fixed phrase and by an LLM judge, and its model ranking already changes between them (Chen et al., 2024); we show that on closed-book-correct questions the refusal ranking of four configurations reverses with the detector and locate the source of the reversal in the output forms.

Parametric knowledge and retrieval policies. Conflict and counterfactual studies separate contextual from parametric knowledge (Neeman et al., 2023; Xu et al., 2024; Tao et al., 2024; Ren et al., 2025); retriever gains need not become answer gains (Mathur et al., 2026). Adaptive methods decide when and how to retrieve (Asai et al., 2024; Jeong et al., 2024; Moskvoretskii et al., 2025); the tax and rescue are what such a policy, or the back-off of Yoran et al. (2024), must estimate.

3 RECALL-CONTROLLED DIAGNOSTIC

3.1 POPULATION AND RECALL LADDER

We use the 468 questions among the first 500 of the NQ-open validation split (Kwiatkowski et al., 2019) for which the first 500,000 passages of the BEIR/NQ corpus (Thakur et al., 2021) contain a passage with a normalized gold-answer string. This 500K pool determines eligibility and supplies the answer-bearing passages of the oracle tiers. Non-oracle documents at every ladder tier come from a separate 10,000-passage placeholder index whose entries all follow one synthetic template naming a topic and asserting relevance to a subject, described in Appendix I; they contain no encyclopedic text, although bare-number answers collide with template slots on 71 of 468 questions.

The six tiers are **closed book** with no document, **placeholder** with five placeholder passages, **recall30** and **recall60**, where 28.6% and 60.7% of questions receive an answer-bearing passage, **oracle1** with one answer-bearing passage and **oracle5** with that passage plus four placeholder passages. The same 468 questions occur in every cell. Answer-bearing allocation at the intermediate tiers is a deterministic hash: question q is answer-bearing at tier t with target rate r exactly when the first eight hexadecimal digits of $\mathrm{SHA1}(t:q)$, read as an integer modulo 100, fall below $100r$, giving the 134 and 284 questions behind the two rates. Oracle retrieval is answer-aware and serves only as a diagnostic boundary. Because placeholder passages carry no encyclopedic content, the ladder measures the cost of *uninformative* context, not what a weak genuine retriever would do; since the template calls itself synthetic and asserts relevance, the placeholder tier is a specific context treatment.

3.2 MODELS AND PROMPT CONTRACTS

The ladder is reported for four configurations. Qwen3-8B uses its native thinking path; Qwen3-32B, Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3 use a generic chain-of-thought path that appends a brief think instruction to the system prompt and elicits the answer with a fixed suffix after the reasoning, as given in Appendix G. Cross-configuration comparisons therefore vary checkpoint and scaffolding together, and we speak of run configurations rather than model families or scale. Each run contains 468×6 tiers $\times 2$ contracts $= 5{,}616$ records. Decoding is greedy with seed 42, a 256-token reasoning budget, at most 128 answer tokens, the model's chat template and a 4,096-token input limit. Placeholder retrieval combines BM25 and BGE-large-en-v1.5 dense scores (Xiao et al., 2024; Robertson & Zaragoza, 2009), each normalized by its top score and weighted equally.

The four system-prompt contracts are: P1 requires document-only answers and a statement when the documents lack the answer; P2 asks for a short direct answer; P3 additionally licenses the model's own knowledge; P4 instructs the model to ignore the documents. P2–P4 all request a short direct answer and P1 does not, so the contracts differ in answer format as well as evidence policy, and we treat each as a complete prompt treatment; Appendix G gives the exact strings. The ladder uses P2 and P3; all four are compared in the factorial of Section 4.5, a separate setting never pooled with the ladder. Every tax and rescue rate is computed against the same contract's own closed-book cell.

3.3 METRICS AND VALIDATION

Strict EM requires the normalized generation to equal a gold alias; **alias-EM** requires a normalized gold alias to appear as a contiguous token span in the generation. For refusal we use two detectors on the same outputs. The **context-phrase detector** (CCR, for context-citing refusal) fires on 22 negated patterns stating that the provided context lacks the answer or that the question cannot be answered. A further 23rd pattern, "based on the provided documents", carries no negation and also matches affirmative answers; we omit it, as Appendix H details. The **clause detector** counts an output as a refusal when no clause still asserts an answer after hedges are removed, so it also counts generic declines such as "I don't know"; it is still lexical and misses bare declines such as "Unknown". We also split declines into context-attributed, generic and answer-following ones.

Alias-EM agrees with blinded model judgments on stratified samples of 258 Qwen3-8B ladder outputs and 230 factorial outputs under P2 at $T \in \{0, 5\}$, at a reweighted Cohen's $\kappa = 0.700$ and 0.620 respectively, as Appendix B reports; CCR agrees at $\kappa = 0.788$ under P2, the contract on which we use it quantitatively. Unless labelled semantic, every tax and rescue is a rate of the stated lexical

label; blinded semantic labels cover the removal control, the P2 factorial and the P2 closed-book and placeholder ladder cells of Sections 4.3–4.5.

4 RESULTS

4.1 RECALL CHANGES BOTH CORRECTNESS AND REFUSAL

In the Qwen3-8B P2 ladder of Table 12 in Appendix C, alias-EM falls from 0.344 closed-book to 0.226 under placeholder context and rises to 0.620 with one oracle passage, while CCR moves from 0 to 0.534 and back to 0.109. Oracle5 keeps the answer-bearing passage but adds four placeholder passages: alias-EM stays high while CCR rises to 0.201, so answer presence is not the only variable.

4.2 THE NET CHANGE FLIPS SIGN BETWEEN RECALL TIERS

Table 11 gives the paired tax and the net change in alias-EM at every tier for all four configurations. The pattern is the same in each. The tax falls monotonically from the placeholder tier through recall30 and recall60 to oracle1, $0.447 \rightarrow 0.385 \rightarrow 0.273 \rightarrow 0.186$ for Qwen3-8B and $0.623 \rightarrow 0.500 \rightarrow 0.314 \rightarrow 0.206$ for Llama-3.1-8B, while rescue rises from 0.045–0.069 to 0.379–0.518. The net change is significantly negative at the placeholder tier and significantly positive at oracle1 in all four configurations, with exact McNemar tests of gains against losses giving $p < 10^{-8}$ and $p < 2 \times 10^{-6}$. Between them, it is significantly negative at recall30 for Mistral and Llama, $p = 0.031$ and 2×10^{-7}, and significantly positive at recall60 for Qwen3-8B, Mistral and Qwen3-32B, $p \leq 0.017$; Figure 2 shows question-bootstrap intervals. Even with one answer-bearing passage, 15–21% of closed-book-correct answers are lost, so a positive mean never implies that the tax is zero.

Where the sign changes also depends on the contract. Under the permissive P3, each configuration's placeholder tax, measured against its own P3 closed-book cell, is lower than under P2: 0.171, 0.330, 0.338 and 0.188 for Qwen3-8B, Mistral, Llama and Qwen3-32B. The closed-book accuracies behind the two denominators are close, P2 against P3 being $0.344/0.363$, $0.380/0.395$, $0.436/0.481$ and $0.391/0.397$, and for both Qwen runs the net change is already significantly positive at recall30, $+0.071$ with $p = 6 \times 10^{-4}$ and $+0.062$ with $p = 0.002$.

4.3 UNINFORMATIVE CONTEXT: ONE LOSS, DIFFERENT FORMS

Placeholder context costs a large share of closed-book-correct answers in every configuration: the alias-EM tax is $72/161 = 0.447$ for Qwen3-8B, $76/178 = 0.427$ for Mistral-7B, $127/204 = 0.623$ for Llama-3.1-8B and $82/183 = 0.448$ for Qwen3-32B; strict-EM taxes in Appendix C are 0.600–0.773, and the blinded semantic labels of Table 10 give 0.38–0.60.

The form of the loss differs, as Table 2 shows. For Qwen3-8B, Qwen3-32B and Mistral, most lost answers are declines that name the context, 52/72, 62/82 and 47/76, and almost none are generic. For Llama-3.1-8B, 88 of the 127 lost answers are generic declines, led by 42 verbatim "I don't know", and only 2 cite the context; Table 16 gives examples. The two detectors respond to these forms differently. On each configuration's own lost subset, the context-phrase detector ranges from 0.750 for Qwen3-8B to 0.024 for Llama, a ratio of $31.8\times$; the clause detector ranges from 0.756 for Qwen3-32B to 0.645 for Mistral, a ratio of $1.17\times$, and gives Qwen3-8B and Llama 0.722 and 0.709. The rules agree closely on Qwen3-8B, $\kappa_{\text{all}} = 0.851$, and barely above chance on Llama, 0.083.

The divergence has two sources. Generic declines, which the phrase rule cannot match, drive it for Llama-3.1-8B, where 3 of its 88 are phrase hits. Incomplete phrase coverage of context-citing declines drives it for Qwen3-32B and Mistral: the phrase rule matches 51 of 52 such declines for Qwen3-8B but 26 of 47 for Mistral and 21 of 62 for Qwen3-32B. The cross-configuration comparison of "refusal" is therefore decided by the detection rule, on identical outputs, and the same holds for known-question over-refusal over all closed-book-correct questions: Llama scores $4/204 = 0.020$ under the phrase rule, the lowest of the four, and $91/204 = 0.446$ under the clause rule, the highest, as Appendix E reports.

Which rule the semantic labels support. We had all P2 closed-book and placeholder outputs of the four configurations labelled blind to configuration by the same two-run annotation as the removal

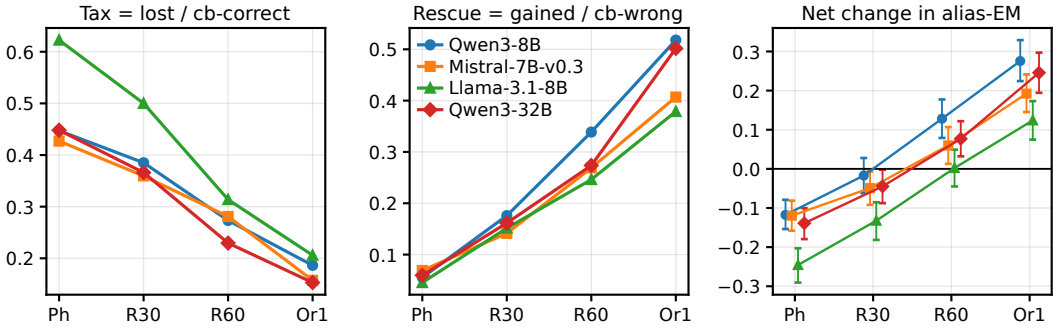

Figure 2: Paired tax, rescue and net change in alias-EM across ladder tiers, P2 contract, four run configurations on the same 468 questions. Net-change whiskers are 95% question-bootstrap intervals over 10,000 resamples. The net change is significantly below zero at the placeholder tier and above zero at oracle1 in every configuration, and its observed sign changes between recall30 and recall60; the tax stays above 0.15 even with an oracle passage.

control, described in Appendix B. The labels agree with the clause rule at $\kappa = 0.89$–0.97 per configuration and with the phrase rule only for Qwen3-8B, at 0.87, against 0.50 for Mistral, 0.30 for Qwen3-32B and 0.06 for Llama. They mark 71–79% of the lost answers as refusals in every configuration, shown in column Sem., and 89 of Llama's 97 lost-answer refusals as generic. Over questions the labels count as known, over-refusal at the placeholder tier is $0.450\,[0.390, 0.510]$ for Llama-3.1-8B against $0.292\,[0.230, 0.354]$ for Mistral in Table 10, the reverse of the phrase rule's ordering of the same two runs, 0.020 against 0.185. Both runs, the agreed labels and the reason for each missing one are released for all 3,744 outputs with a script that recomputes the Sem. column, Table 10 and these agreements, as listed in Appendix A.

Table 2: The same lost answers at the P2 placeholder tier under both detectors. *Regex* and *clause* are refusal rates on the lost subset; *ctx* and *gen* count lost answers in the lexical categories "declines citing the context" and "declines generically", without an answer cue; κ_{all} is detector agreement over all 468 placeholder-tier outputs. *Sem.*: fraction of lost answers that both blinded annotation runs label a refusal.

Configuration	Lost	Regex	Clause	Sem.	Ctx	Gen	κ_{all}
Qwen3-8B	72	0.750	0.722	0.722	52	0	0.851
Mistral-7B-v0.3	76	0.368	0.645	0.711	47	2	0.573
Llama-3.1-8B	127	0.024	0.709	0.764	2	88	0.083
Qwen3-32B	82	0.293	0.756	0.793	62	0	0.306

4.4 Replacing Answer Passages Inside Genuine Rankings

In the ladder the alternative to an answer-bearing passage is a placeholder, as Appendix D shows. This control uses genuine retrieved passages throughout, with Qwen3-8B under P2; by the bounds of Appendix G, no input nears the token limit. Of the 377 questions whose top-5 exact dense retrieval over the 500K pool contains an answer-bearing passage, 17 are ineligible because ranks 6–20 hold too few non-answer passages to replace every answer-bearing position, a property of the ranking alone. The remaining 360 were split 180/180 before generation by shuffling the sorted question identifiers with seed 42 and assigning the first half to removal. In the removal arm every answer-bearing top-5 position is replaced, in rank order, by the highest-ranked unused non-answer passage from ranks 6–20; document count, contract, model, decoder and scoring rule are fixed. The contrast is keeping versus replacing the original answer-bearing passages, not answer presence alone.

Seeded allocation. Closed-book alias-EM is 0.378 in the keep arm and 0.372 in the removal arm. The tax rises from $9/68 = 0.132\,[0.071, 0.233]$ to $42/67 = 0.627\,[0.507, 0.733]$, a difference of $+0.495$ with 95% question-bootstrap interval $[0.351, 0.631]$, shown in Table 3 and Figure 1b.

Registered within-question replication. We then ran every one of the 360 questions with *both* document sets, under a protocol, endpoint and decision rule registered, with the design frozen, before generation, as Appendix A details; this paired containment-tax difference is the primary estimand, the analyses of Table 17 are confirmatory and all others exploratory. The closed-book outputs of the two runs are identical for all 360 questions, so both arms share one set of 135 closed-book-correct questions. The tax is $18/135 = 0.133\,[0.086, 0.201]$ with the original top-5 and $80/135 = 0.593\,[0.508, 0.672]$ with the answer-bearing passages replaced, a paired difference of $+0.459$ with 95% question-bootstrap interval $[0.370, 0.548]$; 64 questions stay correct only with the original documents and 2 only with the replacements, exact McNemar $p = 6 \times 10^{-17}$, which meets the registered criterion; rescue falls from 0.662 to 0.053. Strict EM and a second 180/180 allocation with the registered seed agree, as reported in Appendix A.

Three further models. Under the same registration, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3 and Qwen3-32B were run with the generic chain-of-thought path of their ladder runs rather than native thinking; each gives identical closed-book answers in its two runs. For Llama-3.1-8B the tax rises from $26/174 = 0.149$ to $104/174 = 0.598$, a paired difference of $+0.448\,[0.368, 0.529]$ with 83 against 5 discordant questions; for Mistral-7B from $23/153 = 0.150$ to $77/153 = 0.503$, $+0.353\,[0.275, 0.431]$ with 56 against 2; for Qwen3-32B from $15/152 = 0.099$ to $72/152 = 0.474$, $+0.375\,[0.289, 0.461]$ with 59 against 2. All three meet the registered criterion, and the strict-EM differences of all four removal runs are positive with intervals excluding 0. The effects persist without the questions where the scoring normalization and the raw-substring presence rule disagree, and in every stratum of replacement count, length change and retrieval-score change, including a length-matched subset with $+0.357$ to $+0.481$; Appendix A gives these analyses.

A permissive contract. A fourth registered run repeats the Qwen3-8B paired design under P3. Its closed-book outputs are identical in both runs, and on the 139 closed-book-correct questions the tax rises from $16/139 = 0.115$ to $62/139 = 0.446$, $+0.331\,[0.245, 0.417]$ with 50 against 4 discordant and $p = 4 \times 10^{-11}$, which meets the registered criterion. Removing the answer strings costs known answers even when the prompt invites the model to use its own knowledge.

Alias-span masking. The replacement changes answer status, passage content, relevance and length together. A third registered run, frozen before generation, keeps the five original passages in their order and replaces every occurrence of a gold alias, widened to the enclosing word, by three underscores, as specified in Appendix A. Its closed-book outputs are again identical for 360/360 questions. On the same 135 questions the Qwen3-8B containment tax rises from 0.133 to $49/135 = 0.363\,[0.287, 0.447]$, a paired difference of $+0.230\,[0.156, 0.311]$ with 33 against 2 discordant and $p = 4 \times 10^{-8}$, which meets the registered criterion. The mask effect is $0.50\,[0.35, 0.67]$ of the removal effect under containment and, in an exploratory analysis, $0.37\,[0.18, 0.57]$ under the adjudicated labels below, a ratio of two interventions rather than a mediation share. On Llama-3.1-8B the same frozen mask, a registered secondary analysis, gives $0.149 \to 68/174 = 0.391$, $+0.241\,[0.161, 0.322]$ with 51 against 9 and $p = 3 \times 10^{-8}$, $0.54\,[0.38, 0.71]$ of its removal effect, and under strict EM $+0.275\,[0.183, 0.367]$. A fourth registered run instead masks non-answer word runs approximately matched in number, length and position: the Qwen3-8B tax stays at $18/135$, with five questions changing in each direction, and the answer mask exceeds it by $+0.230\,[0.156, 0.311]$ with 34 against 3 and $p = 1 \times 10^{-7}$, meeting the registered criterion. Registered blinded labels of all 1,080 document sets show that, against keep, the matched mask leaves answer support unchanged at $-0.006\,[-0.015, 0.000]$ while the answer mask removes it in 178 of 329 sets. Both edits remove text, so the contrast compares two edits of the same passages that differ in answer support.

Semantic labels. Containment scoring counts reformatted correct answers as losses, as when "December 1972" becomes "December 14, 1972". We therefore had all 936 outputs of the allocation run, both cells of all 468 questions, and all 2,880 outputs of the Qwen3-8B and Llama-3.1-8B paired runs labelled blind to arm, cell and model by two independent runs of one LLM annotator, following rubrics registered before labelling and described in Appendix B; a label is used when both runs agree on correctness and answer status. On these complete cases the paired tax is $6/133 = 0.045$ against $50/133 = 0.376$ for Qwen3-8B, $+0.331\,[0.248, 0.414]$ with 44 against 0 discordant, and $15/187 = 0.080$ against $79/187 = 0.422$ for Llama-3.1-8B, $+0.342\,[0.273, 0.417]$ with 67 against 3, as in Tables 6 and 17. Two open-weight judges of other families give $+0.107\,[0.040, 0.173]$ and $+0.187\,[0.107, 0.280]$ on Qwen3-8B under the two registered merge rules, meeting the registered rule. A further registered run adjudicated each question's reference once under a written policy and applied it to all its answers, blind to model and arm. On the 147 complete pairs

of the 157 Qwen3-8B questions adjudicated correct closed-book, the tax rises from $10/147$ to $56/147$, $+0.313\,[0.224, 0.408]$ with 53 against 7 and $p = 8 \times 10^{-10}$, with identification bounds $[0.074, 0.463]$ over unresolved labels, which meets the registered rule; Llama-3.1-8B gives $+0.341$. A second, different adjudicator on the same blind items gives $+0.320\,[0.235, 0.405]$ with bounds $[0.036, 0.464]$, also meeting its registered rule. The answer mask exceeds the matched mask by $+0.116\,[0.052, 0.181]$ on the complete cases, and by $+0.105\,[0.044, 0.175]$ on the 114 questions whose reference both runs mark adequate, where its identification bounds $[0.083, 0.150]$ also lie above zero; we therefore state the semantic mask contrast for these two target sets, and Table 17 gives the bounds of every target. Table 8 lists every paired-run semantic estimate with its target set.

A larger reasoning budget. A registered rerun of Qwen3-8B on 180 paired questions with a 1,024-token budget, which cuts the with-document budget hits from 90–144 to 22–62 of 180, gives $+0.493\,[0.343, 0.627]$ for removal and $+0.194\,[0.090, 0.299]$ for the mask, both meeting the registered criterion, against $+0.448$ and $+0.224$ at 256 tokens; scoring the full answer rather than its first line gives $+0.485$ and $+0.191$.

Table 3: The answer-removal control: per-arm transitions against the matched closed-book cell. *Paired*: the registered runs giving all 360 questions both document sets; *allocation*: the seeded 180/180 split of Qwen3-8B; *mask*: the keep passages with the gold-alias spans masked; *sham*: matched non-answer spans masked instead. Tax is lost over closed-book-correct questions, cb, with 95% Wilson interval; rescue is gained over cb-wrong questions; McNemar compares gains with losses within an arm. Answer presence in each document set is given in Appendix A.

Setting	Arm	With docs	Tax, lost/cb	Wilson 95%	Rescue	McNemar p
Qwen3-8B, paired	Keep	266/360	$18/135 = 0.133$	$[0.086, 0.201]$	$149/225$	7×10^{-27}
	Removal	67/360	$80/135 = 0.593$	$[0.508, 0.672]$	$12/225$	2×10^{-13}
	Mask	112/360	$49/135 = 0.363$	$[0.287, 0.447]$	$26/225$	0.011
	Sham	267/360	$18/135 = 0.133$	$[0.086, 0.201]$	$150/225$	4×10^{-27}
Llama-3.1-8B, paired	Keep	240/360	$26/174 = 0.149$	$[0.104, 0.210]$	$92/186$	8×10^{-10}
	Removal	83/360	$104/174 = 0.598$	$[0.523, 0.668]$	$13/186$	8×10^{-19}
	Mask	133/360	$68/174 = 0.391$	$[0.321, 0.465]$	$27/186$	3×10^{-5}
Mistral-7B, paired	Keep	236/360	$23/153 = 0.150$	$[0.102, 0.215]$	$106/207$	6×10^{-14}
	Removal	87/360	$77/153 = 0.503$	$[0.425, 0.581]$	$11/207$	2×10^{-13}
Qwen3-32B, paired	Keep	275/360	$15/152 = 0.099$	$[0.061, 0.156]$	$138/208$	4×10^{-26}
	Removal	96/360	$72/152 = 0.474$	$[0.396, 0.553]$	$16/208$	1×10^{-9}
Allocation, Qwen3-8B	Keep	132/180	$9/68 = 0.132$	$[0.071, 0.233]$	$73/112$	1.4×10^{-13}
	Removal	32/180	$42/67 = 0.627$	$[0.507, 0.733]$	$7/113$	3.6×10^{-7}

4.5 A Single-Setting Four-Contract Factorial

The factorial crosses all four contracts with genuine retrieval in one evaluation setting. The same 468 questions pass through every cell with one model, Qwen3.5-9B in text-only mode, greedy decoding at seed 42, a 256-token reasoning budget and exact dense retrieval over the full 2.68M-passage BEIR/NQ corpus, at $T \in \{0, 1, 5, 20\}$ documents. A retrieved passage contains a gold answer string for 81.0% of questions at top-5 and for 77.4% under a stricter token-span match. The analysis uses $T \in \{0, 1, 5\}$; at $T = 20$ the input limit can truncate the question, which follows the context.

Contracts do not share a closed-book baseline. Under P1 the $T = 0$ cell scores alias-EM 0.000 with a CCR hit rate of 0.964, so the tax denominator is zero. P2, P3 and P4 score 0.382, 0.387 and 0.387. The instruction alone thus moves the no-document score by 0.38 alias-EM, so each contract needs its own baseline. CCR hits at $T = 5$ fall as the contract is relaxed, $0.156 \to 0.077 \to 0.026 \to 0.000$; outside P2 these are hit-rate differences rather than validated abstention rates.

An ignore-context instruction does not make accuracy independent of the documents. P4 tells the model to ignore the documents, yet its alias-EM rises from 0.387 to 0.513 and 0.603 as documents are added, and it loses 9.4% of its closed-book-correct answers at $T = 5$. The instruction suppresses context-phrase refusals entirely, so an ignore arm is a contract of its own, not an ablation of context.

Table 4: Four prompt contracts on the same 468 NQ questions, with Qwen3.5-9B and exact dense retrieval over 2.68M passages. Tax and rescue are paired against each contract's own $T=0$ cell. The 95% Wilson intervals of the tax are $[0.096, 0.198]$ for P2 and $[0.059, 0.145]$ for P3 and P4; those of the rescue are $[0.532, 0.621]$, $[0.434, 0.549]$, $[0.434, 0.549]$ and $[0.356, 0.469]$ for P1–P4. †P1 forbids answering without documents, so its $T=0$ cell scores zero, its tax is undefined and its rescue equals its $T=5$ accuracy, not a parametric rescue.

	Alias-EM			At $T=5$		CCR	
	0	1	5	Tax	Resc.	0	5
P1	0.000	0.434	0.577	—	0.577†	0.964	0.156
P2	0.382	0.491	0.632	0.140	0.491	0.000	0.077
P3	0.387	0.549	0.652	0.094	0.491	0.002	0.026
P4	0.387	0.513	0.603	0.094	0.411	0.002	0.000

At high recall the tax is small and outweighed by rescue. At $T = 5$ the tax is $25/179 = 0.140$ under P2 and 0.094 under P3 and P4, against rescue of $142/289 = 0.491$, 0.491 and 0.411; under P2 the net gain is $117/468 = 0.250$. Blinded semantic labels of the P2 outputs agree, with tax $25/199 = 0.126$ and rescue $102/146 = 0.699$.

4.6 The Metric Decides Which Contract Improves

Under alias-EM every contract improves from $T = 1$ to $T = 5$, each with exact McNemar $p < 2 \times 10^{-6}$. Under strict EM the contracts separate. P2 improves from 0.310 to 0.372, 59 gains against 30 losses, $p = 0.003$. P1 declines from 0.118 to 0.060, 11 gains against 38 losses, $p = 1.4 \times 10^{-4}$, and token F1 falls from 0.211 to 0.191. The metrics disagree because the answers change form: 36 of the 38 strict losses still contain a gold alias at $T=5$, and on those 38 questions the mean answer length grows from 2.2 to 16.4 words; neither metric alone says if a longer answer still asserts it.

5 Discussion

The protocol. (i) Run a closed-book cell under the same prompt contract as every retrieval cell. (ii) Report the paired tax and rescue with their denominators, not only the mean. (iii) Score correctness with a strict and a containment metric; a disagreement in direction is a finding about the metric. (iv) State the detector's construct and split context-attributed from generic declines before comparing configurations. (v) Intervene on the answer strings inside genuine rankings with document count fixed.

Scope. The results cover 468 filtered single-hop English NQ questions, single greedy runs with intervals over questions, four 7–32B run configurations, one model for the factorial and the matched non-answer mask and two for the answer mask; the four checkpoints share the 360 questions and edits of the removal control, so the replication is across models on a shared question set. The tax is relative to one greedy closed-book output; the removal effect persists on questions closed-book-correct under both contracts or for at least three of the four models, as Appendix A shows. Semantic labels are model annotations. Effect sizes refer to a 256-token reasoning budget, which binds for most Qwen outputs with documents, and first-line extraction; the fourfold budget of Section 4.4, and Llama-3.1-8B and Mistral-7B, where the budget rarely binds, show the same direction. The recall30/recall60 sign change is specific to these two tiers.

6 Conclusion

The retrieval tax depends on the context regime: rescue dominates under high-recall retrieval, replacing the answer-bearing passages raises the tax by 0.33 to 0.46 and masking their alias spans about half as much, also at a fourfold reasoning budget, and uninformative context costs most, in forms that decide which model a refusal detector ranks as the most cautious.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to formulate questions for surveys or interviews or to transcribe recordings, and surveys, interviews and recordings are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. Large language models also serve in this study as blinded annotators of model outputs; this measurement instrument, its rubric and its agreement statistics are described in Section 3.3 and Appendix B. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and no personal data. All questions come from the public Natural Questions open-domain split, all passages from the public BEIR release of its Wikipedia corpus, and all generators and open-weight judges are publicly released models; the synthetic placeholder passages contain no information about any person. The released material consists of these public questions and passages, model outputs and annotation labels. The study measures when retrieved context removes answers that a model would otherwise give and how measurement choices change that reading; we see no direct harmful application, and the findings are intended to make the evaluation of retrieval-augmented systems more accurate. The model outputs may contain factual errors inherited from the models and are released only as evaluation records. We declare no conflicts of interest or sponsorship that bear on this work.

REPRODUCIBILITY STATEMENT

The population, recall ladder, prompt contracts, decoding settings and metrics are specified in Section 3, with the exact prompt strings and answer extraction in Appendix G, the refusal rules in Appendix H and the placeholder construction in Appendix I. The answer-removal, masking, permissive-contract and larger-budget designs, their registered endpoints and decision rules are described in Section 4.4 and Appendix A, the semantic annotation protocol in Appendix B, the accounting identity with its proof in Appendix J, and all registered analyses with their results in Appendix K. The anonymous supplementary material, described in Appendix A, contains the questions with their retrieval rankings, the paired design with all passages, the extracted outputs of every run behind the reported results, the complete generations of the larger-budget run, the semantic labels with the reason for each missing label, the annotation rubrics, the refusal detectors and four scripts that recompute the results listed in its README without model calls.

REFERENCES

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258, 2025. doi: 10.18653/v1/2025.acl-long.892.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024. doi: 10.1609/aaai.v38i16.29728.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, 2024. doi: 10.1145/3626772.3657834.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23281–23298, 2025. doi: 10.18653/v1/2025.findings-emnlp.1264.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Lovisa Hagström, Youna Kim, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. CUB: Benchmarking context utilisation techniques for language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25101–25133, 2026. doi: 10.18653/v1/2026.acl-long.1151.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *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, 2024. doi: 10.18653/v1/2024.naacl-long.389.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5591–5606, 2023. doi: 10.18653/v1/2023.acl-long.307.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. Retrieval helps or hurts? A deeper dive into the efficacy of retrieval augmentation to language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5506–5521, 2024. doi: 10.18653/v1/2024.naacl-long.308.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546.

Saahil Mathur, Ryan David Rittner, Vedant Ajit Thakur, Daniel Stuart Schiff, and Tunazzina Islam. Retrieval improvements do not guarantee better answers: A study of RAG for AI policy QA. *arXiv preprint arXiv:2603.24580*, 2026.

Michał Mazuryk, Fleur Dolmans, Louis Gehringer, Ina Klaric, Jia-Huei Ju, and Mohammad Aliannejadi. The powerless noise: How experimental settings shape the reported power of noise. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2973–2982, 2026. doi: 10.1145/3805712.3808556.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. Adaptive retrieval without self-knowledge? Bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6355–6384, 2025. doi: 10.18653/v1/2025.acl-long.319.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, 2023. doi: 10.18653/v1/2023.acl-long.559.

Seong-Il Park and Jay-Yoon Lee. Toward robust RALMs: Revealing the impact of imperfect retrieval on retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 12:1686–1702, 2024. doi: 10.1162/tacl_a_00724.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3697–3715, 2025.

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.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, pp. 25055–25083, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 31210–31227, 2023.

Yufei Tao, Adam Hiatt, Erik Haake, Antonie J. Jetter, and Ameeta Agrawal. When context leads but parametric memory follows in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4034–4058, 2024. doi: 10.18653/v1/2024.emnlp-main.234.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10303–10315, 2023. doi: 10.18653/v1/2023.findings-emnlp.691.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. How easily do irrelevant inputs skew the responses of large language models? In *First Conference on Language Modeling*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024. doi: 10.18653/v1/2024.emnlp-main.486.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. How is LLM reasoning distracted by irrelevant context? An analysis using a controlled benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13329–13347, 2025. doi: 10.18653/v1/2025.emnlp-main.674.

Shiping Yang, Jie Wu, Wenbiao Ding, Ning Wu, Shining Liang, Ming Gong, Hongzhi Li, Hengyuan Zhang, Angel X. Chang, and Dongmei Zhang. Quantifying and improving the robustness of retrieval-augmented language models against spurious features in grounding data. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33479–33499, 2026. doi: 10.18653/v1/2026.acl-long.1545.

Samuel Yeh and Sharon Li. How retrieved context shapes internal representations in RAG. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15446–15468, 2026. doi: 10.18653/v1/2026.findings-acl.758.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024. URL https://iclr.cc/virtual/2024/poster/18397.

Youchao Zhou, Heyan Huang, Yicheng Liu, Rui Dai, Xinglin Wang, Xingchen Zhang, Shumin Shi, and Yang Deng. Do retrieval augmented language models know when they don't know? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35158–35166, 2026. doi: 10.1609/aaai.v40i41.40822.

A Removal control: registered replication and answer presence

Registered within-question replication. The protocol, the primary endpoint, which is the paired containment-tax difference on the closed-book-correct questions, the decision rule, which requires a difference above 0, exact McNemar $p < 0.01$ on the discordant pairs and a 95% question-bootstrap interval excluding 0, the secondary analyses and the fresh allocation seed 20260923 were registered, and the paired design was frozen, before generation. The registered analysis was run once on the completed outputs. Every one of the 360 eligible questions was run twice with the generation settings of the removal control, once with each document set, under P2 with greedy decoding, seed 42, a 256-token reasoning budget and 128 answer tokens, each run regenerating the closed-book cell. The run differs from the original control in the input limit, 4,096 tokens instead of 16,384, which does not bind by the bounds of Appendix G; each comparison in this paper is between arms generated with the same settings. All 720 with-document records match the document sets of the frozen design. The two closed-book cells are identical for 360/360 questions. Regenerating each question's original arm reproduces the original alias-EM label for 347/360 questions. Under strict EM the paired tax is $6/17 = 0.353$ against $14/17 = 0.824$, a difference of $+0.471\,[0.176, 0.765]$. The reasoning budget is reached by 187 keep-set and 265 removal-set outputs with documents. A second 180/180 allocation with the registered seed, read from the same runs, gives $12/73 = 0.164$ against $38/62 = 0.613$, a difference of $+0.449\,[0.297, 0.595]$. The four closed-book/with-document cells, correct/correct, correct/incorrect, incorrect/correct and incorrect/incorrect, number $117/18/149/76$ with the original passages and $55/80/12/213$ with the replacements.

Closed-book stability, a post hoc analysis. A single greedy closed-book answer may be correct by chance. On the 119 questions that Qwen3-8B answers correctly closed-book under both P2 and P3, the paired difference is $+0.454\,[0.361, 0.546]$ under P2, with 56 against 2 discordant and $p = 1.2 \times 10^{-14}$, and $+0.294\,[0.202, 0.387]$ under P3, with 38 against 3. On the questions answered correctly closed-book by at least three of the four paired-run models it is $+0.421\,[0.318, 0.514]$ for Qwen3-8B with $n = 107$, $+0.376\,[0.272, 0.472]$ for Llama-3.1-8B with 125, $+0.303\,[0.221, 0.393]$ for Mistral-7B with 122 and $+0.341\,[0.260, 0.423]$ for Qwen3-32B with 123; on the 72 answered by all four it is $+0.375$, $+0.292$, $+0.250$ and $+0.264$, each interval excluding 0. These subsets were defined after the registered analyses and are exploratory.

Further models. The registration applies the same design, endpoint and rule, without pooling, to Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3 and Qwen3-32B, each with the generation path of its ladder run. For Llama-3.1-8B and Mistral-7B this is the generic chain-of-thought path of Appendix G. The Mistral-7B run uses the model's current official chat template with unchanged weights, which, like the template of its ladder run, accepts the system prompt. Settings and input limit are otherwise those of the Qwen3-8B paired run, and all 720 with-document records of each run match the document sets of the frozen design. In each model the two closed-book cells are identical for 360/360 questions. Under strict EM the paired tax is $21/120 = 0.175$ against $74/120 = 0.617$ for Llama-3.1-8B, $+0.442\,[0.333, 0.550]$ with 59 against 6 discordant, and $12/38 = 0.316$ against $26/38 = 0.684$ for Mistral-7B, $+0.368\,[0.211, 0.526]$ with 15 against 1 and $p = 5 \times 10^{-4}$; the registered-seed 180/180 split gives $16/85 = 0.188$ against $55/89 = 0.618$, $+0.430\,[0.300, 0.561]$, and $12/74 = 0.162$ against $41/79 = 0.519$, $+0.357\,[0.219, 0.493]$. The reasoning budget is reached by at most 4 outputs with documents in any of these four runs. Qwen3-32B uses the generic chain-of-thought path of its ladder run; its two closed-book cells are identical for 360/360 questions. Under strict EM its paired tax is $33/76 = 0.434$ against $63/76 = 0.829$, $+0.395\,[0.250, 0.526]$ with 35 against 5 and $p = 1 \times 10^{-6}$, and the registered-seed split gives $7/77 = 0.091$ against $38/75 = 0.507$, $+0.416\,[0.284, 0.545]$. Its reasoning budget is reached by 213 keep-set and 278 removal-set outputs with documents.

Alias-span masking. The mask design and its analysis were registered and frozen before any masked output existed. For each of the 360 questions it takes the keep top-5, with the same passages, order and identifiers, and replaces every case-insensitive occurrence of any gold alias, widened to the enclosing alphanumeric run and longest alias first, by three underscores in all five passages: 1,201 spans with 13,471 characters, 1.30% of the passage text. Widening makes 62 of these spans, in 23 questions, longer than the matched alias, for example the alias "one" inside "Executioner" or "7" inside "2007"; without these questions the tax is $17/126$ against $46/126$, $+0.230\,[0.151, 0.310]$ with 31 against 2 and $p = 1.3 \times 10^{-7}$, and removal gives $+0.460$ on the same questions. The masking claim concerns the registered containment endpoint; its registered strict-EM secondary is reported in Table 17. No mask set contains a gold alias as a raw substring; 8 contain one as a scoring-normalized token span, from aliases with articles or brackets. The run used the generation settings and weights of the Qwen3-8B paired run; all 360 records are present and the closed-book outputs equal those of the paired run for 360/360 questions. Excluding the 8 questions leaves $16/129$ against $48/129$, $+0.248\,[0.171, 0.326]$ with 33 against 1. Rescue is $149/225$ in keep, $26/225$ in mask and $12/225$ in removal, and accuracy with documents 0.739, 0.311 and 0.186; replacing the masked passages by the next non-answer passages raises the tax by a further $+0.230\,[0.141, 0.319]$. The reasoning budget is reached by 303 masked outputs with documents. On Llama-3.1-8B, run with the generic chain-of-thought path as in its paired run, all 720 masked records are present and the closed-book outputs equal those of the paired run for 360/360 questions. Excluding the 8 token-span questions leaves $24/169$ against $68/169$, $+0.260\,[0.183, 0.337]$ with 51 against 7 and $p = 2 \times 10^{-9}$; under strict EM the tax is $21/120$ against $54/120$, $+0.275\,[0.183, 0.367]$ with 38 against 5 and $p = 2 \times 10^{-7}$. Rescue is $92/186$ in keep, $27/186$ in mask and $13/186$ in removal, and accuracy with documents 0.667, 0.369 and 0.231; replacing the masked passages raises the tax by a further $+0.207\,[0.121, 0.293]$, with 52 against 16. The reasoning budget is reached by 11 closed-book and 4 masked with-document outputs.

Matched non-answer mask. This run and its analysis were registered, and its design frozen, before any of its outputs existed. For every span of the answer mask it replaces by three underscores one run of 1–8 words in the same passage, of similar length, within $\max(2, 0.25L)$ characters where possible, and nearest in relative position, that overlaps no gold-alias occurrence or the two words around it and contains no question content word, meaning a word of the question outside a fixed stopword list, and no alias token. This gives 1,199 spans, with one span unmatched in each of two questions, and 12,658 masked characters, against 1,201 spans and 13,471 characters of the answer mask; the mean length difference is 1.5 characters and the mean relative-position difference 0.064, and every sham set still contains a gold alias. The run used the generation settings and weights of the Qwen3-8B paired run; all 720 records are present and its closed-book outputs equal those of the paired run for 360/360 questions. Its tax equals the keep tax at $18/135$, with five questions changing in each direction and a difference of $0.000\,[-0.044, 0.044]$; accuracy with documents is 0.742 against 0.739 for keep, and rescue $150/225$. Answer masking exceeds the matched mask by $+0.471\,[0.235, 0.706]$ under strict EM, $3/17$ against $11/17$ with $p = 0.008$, and by $+0.231\,[0.157, 0.313]$ without the two questions with an unmatched span; the ratio of this contrast to the answer-mask effect is

$1.00\,[0.80, 1.24]$, a ratio of two interventions rather than an attribution. The matching is approximate and concerns form, not content: a non-answer span can still carry evidence, as "written and" before "recorded by James Brown" does for a question about the song's writer. The reasoning budget is reached by 191 closed-book and 215 with-document outputs of this run.

Permissive contract. The P3 run uses the frozen keep and removal document sets, the Qwen3-8B generation settings and weights of the P2 paired run, and the P3 system prompt of Appendix G in every cell, closed book included. All 1,440 records are present and the two closed-book cells are identical for 360/360 questions; closed-book alias-EM is $139/360 = 0.386$. Rescue is $152/221$ in keep and $20/221$ in removal, and accuracy with documents 0.764 and 0.269. The registered-seed 180/180 split gives $9/76 = 0.118$ against $31/63 = 0.492$, $+0.374\,[0.227, 0.514]$. The reasoning budget is reached by 213 keep-set and 291 removal-set outputs with documents. P3 answers are long, so only 4 closed-book answers are strict-EM correct; the registered strict-EM secondary is listed in Table 17.

A larger reasoning budget. This run, its endpoints, decision rule and analysis were registered before any of its outputs existed. It takes the 180 questions at even positions of the sorted 360 paired-design identifiers and runs Qwen3-8B with the unchanged generator, P2, greedy decoding at seed 42 and 128 answer tokens in four cells: closed book and the frozen keep, removal and answer-mask document sets. The only change is the reasoning budget, 1,024 tokens instead of 256, and both decoded stages are kept in full: the reasoning stage and the answer stage before first-line extraction. At the original budget the same generator reproduces the earlier outputs of the questions checked. All 720 records are present. The closed-book cell defines the closed-book-correct set, 67 questions, and the two primary contrasts, removal minus keep and mask minus keep, each use the registered rule of the paired runs. Removal gives $12/67 = 0.179$ against $45/67 = 0.672$, $+0.493\,[0.343, 0.627]$ with 36 against 3 discordant and $p = 4{\times}10^{-8}$; the answer mask gives $25/67 = 0.373$, $+0.194\,[0.090, 0.299]$ with 15 against 2 and $p = 0.002$; both meet the rule. On the same questions at 256 tokens the contrasts are $+0.448\,[0.313, 0.582]$ and $+0.224\,[0.104, 0.343]$. Scored on the full answer stage instead of its first line, where the closed-book-correct set has 68 questions, they are $+0.485\,[0.338, 0.618]$ and $+0.191\,[0.088, 0.309]$ with the same discordant counts; at most 2.2% of answer stages in any cell span more than one line. The budget is reached by 18, 22, 35 and 62 of the 180 closed-book, keep, removal and mask outputs, against 93, 90, 127 and 144 at 256 tokens, and the mean reasoning length is 397, 393, 522 and 660 tokens, against 224 to 246. Accuracy is 0.372 closed book and 0.739, 0.161 and 0.333 with the keep, removal and masked sets, against 0.761, 0.183 and 0.306 at 256 tokens. The strict-EM secondaries rest on 13 closed-book-correct questions and are listed in Table 17.

Released evidence. The supplementary material contains the 468 questions with their gold aliases, the oracle passage of each and the top-20 ranking with identifiers, scores and passages, from which every removal design is drawn; for every question of the paired design, the keep, removal, masked and matched-mask passages with their identifiers and the second allocation; every closed-book and with-document extracted answer with its reasoning-token count for the thirteen paired, masked and permissive-contract runs, the seeded allocation over all 468 questions, including the natural-miss and ineligible ones, the four ladder configurations and the factorial at $T \in \{0, 1, 5\}$; the semantic labels of the paired runs, the allocation, the factorial P2 cells and all 3,744 P2 closed-book and placeholder-tier ladder outputs, each with the labels of both annotation runs, the agreed label and the reason for every missing label, the annotation rubric and the exact instructions and decoding settings of the open-weight judges; the two-run labels of the answer-mask and matched-mask outputs; the question-level adjudication, with the status and accepted answers of every question, both runs, the second adjudicator and the adjudicated label of every paired-run answer; the document-support labels of the keep, answer-mask and matched-mask sets, with their rubrics; the per-question prompt token counts of every paired-run model; the refusal detectors; and two scripts that recompute without model calls the containment and strict-EM paired endpoints of every paired, masked and permissive-contract run, the strata of Table 5, the closed-book-stability subsets, the semantic complete-case, merge-rule, identification-bound and fixed-target analyses, the ladder tables, also without the token-span exceptions of Appendix I, the detector, anchor, factorial and allocation tables, the semantic refusal readings, namely the Sem. column of Table 2, Table 10 and the agreement of each detector with the semantic labels, and the adjudicated, masked-arm and document-support estimates of Appendix B. A third script rebuilds the matched non-answer mask from the keep passages and checks it against the released design. The released answers of these runs are the extracted first lines; for the larger-

budget run the material holds both decoded stages of all 720 generations in full, with a fourth script that recomputes its registered contrasts, budget counts and full-answer scoring.

Decomposing the replacement, an exploratory analysis. Table 5 splits each paired run's closed-book-correct questions by properties of its replacement: the number of answer-bearing top-5 positions replaced, tertiles of the change in prompt length under the model's tokenizer, removal minus keep, and tertiles of the change in mean dense retrieval score of the five passages. The same endpoint, test and bootstrap as the registered analysis are applied within each stratum; the unstratified rows reproduce the registered results exactly. The analysis was registered as exploratory, without a decision rule, before it was run. The effect is positive, with an interval excluding 0, in every stratum of every model, including the length-matched subset whose removal prompt is within 10% of the keep prompt.

Table 5: Paired containment-tax difference, removal − keep, within strata of the closed-book-correct questions, with 95% question-bootstrap interval and n. *Length*: tertiles of the prompt-length change, removal − keep, where low is most shortened; *score*: tertiles of the drop in mean retrieval score, keep − removal, where low is the smallest drop. *Matched*: $|\Delta\text{length}| \leq 10\%$ of the keep prompt.

Stratum	Qwen3-8B	Llama-3.1-8B	Mistral-7B	Qwen3-32B
All	$+0.459\,[.37, .55]$ 135	$+0.448\,[.37, .53]$ 174	$+0.353\,[.27, .43]$ 153	$+0.375\,[.29, .46]$ 152
1 replaced	$+0.345\,[.14, .55]$ 29	$+0.477\,[.30, .64]$ 44	$+0.294\,[.15, .44]$ 34	$+0.382\,[.21, .56]$ 34
2 replaced	$+0.449\,[.31, .59]$ 49	$+0.371\,[.24, .50]$ 62	$+0.355\,[.23, .48]$ 62	$+0.379\,[.24, .52]$ 58
≥ 3 replaced	$+0.526\,[.39, .65]$ 57	$+0.500\,[.37, .62]$ 68	$+0.386\,[.26, .51]$ 57	$+0.367\,[.25, .48]$ 60
Length, low	$+0.489\,[.33, .64]$ 45	$+0.500\,[.36, .64]$ 58	$+0.314\,[.20, .45]$ 51	$+0.400\,[.26, .54]$ 50
Length, mid	$+0.444\,[.29, .60]$ 45	$+0.379\,[.24, .52]$ 58	$+0.314\,[.18, .45]$ 51	$+0.353\,[.22, .49]$ 51
Length, high	$+0.444\,[.31, .60]$ 45	$+0.466\,[.33, .60]$ 58	$+0.431\,[.27, .57]$ 51	$+0.373\,[.24, .51]$ 51
Score, low	$+0.422\,[.29, .56]$ 45	$+0.414\,[.28, .55]$ 58	$+0.314\,[.20, .45]$ 51	$+0.260\,[.12, .40]$ 50
Score, mid	$+0.378\,[.22, .53]$ 45	$+0.431\,[.29, .57]$ 58	$+0.333\,[.20, .47]$ 51	$+0.431\,[.29, .57]$ 51
Score, high	$+0.578\,[.42, .73]$ 45	$+0.500\,[.36, .64]$ 58	$+0.412\,[.25, .55]$ 51	$+0.431\,[.29, .57]$ 51
Matched	$+0.481\,[.35, .62]$ 52	$+0.400\,[.27, .53]$ 70	$+0.362\,[.22, .50]$ 58	$+0.357\,[.23, .50]$ 56

Answer presence and the seeded allocation. A gold alias occurs as a raw substring in the five supplied documents for every keep set and no removal or mask set, and as a scoring-normalized token span in 336/360 keep, 4/360 removal and 8/360 mask sets, and in 169/180 keep and 2/180 removal sets of the allocation. The token-span disagreements arise from articles, punctuation and footnote markers inside aliases; for example, an alias that loses its article becomes a common word. In the allocation, excluding the 13 such questions gives a tax of $8/67 = 0.119$ against $41/65 = 0.631$, a difference of $+0.511\,[0.366, 0.651]$; in the paired runs, excluding the 28 gives $16/129$ against $77/129$ for Qwen3-8B, $+0.473\,[0.380, 0.566]$ with 63 against 2 discordant questions, $25/168$ against $100/168$ for Llama-3.1-8B, $+0.446\,[0.363, 0.530]$ with 80 against 5, $21/148$ against $74/148$ for Mistral-7B, $+0.358\,[0.277, 0.439]$ with 55 against 2, and $14/148$ against $71/148$ for Qwen3-32B, $+0.385\,[0.304, 0.466]$ with 59 against 2. The bootstrap mean of the allocation's alias-EM tax difference is 0.4951 against the observed 0.4945; under strict EM the observed difference is $5/6 - 3/11 = 0.5606$, with 9,974 of 10,000 resamples valid because some contain no closed-book-correct question in an arm. Of the 468 questions, 360 enter the comparison, 17 lack enough replacement passages and 91 have no answer-bearing passage in the original top-5. The same run gives the 91 natural-miss questions their unmodified top-5: the tax is $8/11 = 0.727$ with 95% Wilson interval $[0.434, 0.903]$ and the rescue $3/80$, against $42/67$ and $7/113$ in the removal arm; this comparison was registered as exploratory before it was computed.

B BLINDED MEASUREMENT VALIDATION

We stratified all $5{,}616$ Qwen3-8B ladder records on the cross of the alias-EM and CCR labels and sampled 80, 58, 60, 60 from the four cells, whose populations are $2{,}591$, 58, 843 and $2{,}124$, so the containment-with-refusal cell is a census. The judge, a proprietary LLM at its maximum reasoning setting, saw only the question, the gold aliases and the response, never either label, the tier or the contract, and decided whether the response *asserts* a correct answer, a gold string inside a denial or an aside not counting, and whether its primary act is declining. Records were presented in random order

in blocks of twenty. Ladder intervals are stratified *record*-bootstrap percentiles over $4{,}000$ resamples; the 258 records come from 188 questions, so they understate question-level uncertainty. The factorial intervals resample questions with their paired records. Reweighted to the population, alias-EM has precision $0.905\,[0.843, 0.957]$, recall $0.801\,[0.738, 0.870]$ and $\kappa = 0.700\ [0.593, 0.799]$. CCR has $\kappa = 0.788$ under P2 and 0.877 in a 132-record rule-based adjudication. Re-judging 22 items blind returned the same verdict for 22/22. Containment errors concentrate in rate but not in mass where containment and refusal co-occur, with 58.6% correct against 91.3% without refusal: of an expected 251 false positives about 24 fall in that 58-record cell and 227 where containment occurs alone. On the factorial's $T = 0$ and $T = 5$ P2 records the same procedure over 230 records gives unweighted $\kappa = 0.662\ [0.565, 0.756]$ and population-reweighted $\kappa = 0.620\ [0.450, 0.769]$. The full annotation below replaces these samples as the semantic reading of the factorial P2 cells.

Full semantic annotation of the removal control and the P2 factorial. All 1,690 unique question, reference and response items behind the 936 removal-control records, closed book and with documents for all 468 questions, and the 936 P2 factorial $T \in \{0, 5\}$ records were labelled by two independent runs of one proprietary LLM annotator, a later model than the judge above, at a high reasoning setting and with a different batch composition, each seeing only question, reference aliases and response, with a rubric and a merge rule registered before labelling. An item is labelled when both runs agree on correctness and on answer status, such as ANSWER, REFUSAL or MIXED; a response judged correct must quote its answer span exactly; an item whose reference either run flags as inadequate stays unknown. The two runs agree on 98.7% of items both label, $\kappa = 0.974$, a measure of the repeatability of one annotator. Agreement with alias-EM is $\kappa = 0.60$ on the removal control and 0.61 on the factorial. Table 6 gives the removal control; on the factorial P2 cells the semantic tax at $T = 5$ is $25/199 = 0.126$ and rescue $102/146 = 0.699$, with 123 pairs without an agreed label.

Table 6: Removal control under blinded semantic labels, the agreed labels of two annotation runs. Denominators are the closed-book-correct questions for the tax and the closed-book-incorrect questions for the rescue, both under the semantic labels.

	Keep arm	Removal arm
Tax	$3/72 = 0.042$	$25/67 = 0.373$
Rescue	$60/67 = 0.896$	$11/59 = 0.186$
Pairs without an agreed label	41	54
Containment losses: semantic loss / unchanged / unknown	$1/5/3$	$20/10/12$

The tax difference is $+0.331$ with question-bootstrap interval over the agreed labels $[0.204, 0.461]$; each run alone gives $+0.330$ and $+0.376$. The agreed-label denominators exclude questions whose closed-book answer is agreed correct but whose with-document answer has no agreed label: 75 closed-book answers are agreed correct in each arm, of which 3 in keep and 8 in removal lack an agreed with-document label. The same annotation of the five supplied passages finds a passage that supports the answer for 156 of 180 keep sets and 26 of 180 removal sets, with 5 and 9 unknown. The rubric and merge rule were registered before labelling began; Table 17 lists the registered analyses.

The paired runs. The same two-run annotation, with the same annotator, rubric and merge rule and new shuffles, was registered for all 2,880 records of the two paired runs; items identical to an already labelled item keep that label, and the other 1,085 unique items were labelled. The runs agree at $\kappa = 0.964$ on the Qwen3-8B and 0.967 on the Llama-3.1-8B outputs, and the agreed labels agree with alias-EM at $\kappa = 0.61$ and 0.54. Of the 150 Qwen3-8B and 202 Llama-3.1-8B closed-book answers agreed correct, 133 and 187 have agreed labels in both with-document cells and form the complete cases of Section 4.4; 73 and 89 closed-book labels are unknown. Table 17 lists the registered identification bounds.

A judge from a second model family, an exploratory analysis. Two runs of one annotator measure its repeatability, not its validity. The registration also provides open-weight judges, of which Qwen3-32B has labelled the 1,690 items above and the 1,440 records of the Qwen3-8B paired run, blind to arm and cell, with the same rubric; a label is usable when it is a boolean with a verbatim quote for a correct answer and no reference flag. Against the agreed labels of the two-run annotation it agrees on 92.3% of the 769 items both label, $\kappa = 0.635$, and on 93.8% of the 471 paired-run

items both label, $\kappa = 0.695$; it leaves 800 of the 1,690 and 473 of the 1,028 unique paired-run items without a usable label, so these figures describe the labelled overlap. On its own labels the complete-case paired tax is $1/84 = 0.012$ against $15/84 = 0.179$, $+0.167\,[0.095, 0.250]$ with 14 against 0 discordant and $p = 1.2 \times 10^{-4}$. On the cross-family consensus, where both annotators are known and equal, 125 closed-book answers are consensus-correct and 218 unknown, and the tax is $0/63$ against $10/63 = 0.159$, $+0.159\,[0.079, 0.254]$ with 10 against 0 and $p = 0.002$. These analyses were registered as exploratory before they were run.

Two open-weight judges. The registered two-judge analysis adds Llama-3.3-70B-Instruct, run once over the same items with the same rubric and usability rule. It leaves 169 of the 1,690 items and 120 of the 1,028 unique paired-run items without a usable label, and agrees with the two-run annotation on 90.9% of the 1,191 items both label, $\kappa = 0.811$, and on 90.7% of 710 paired-run items, $\kappa = 0.800$; the two judges agree at $\kappa = 0.62$ and with alias-EM at $\kappa = 0.30$ for Qwen3-32B and 0.61 for Llama-3.3-70B. On the Qwen3-8B paired run the registered merge rules are given in Section 4.4; 75 questions are complete, and 277 of 360 questions lack a label that the merge rules need. The two merge rules are computed on these complete questions; they are not identification bounds over missing labels, which Table 17 lists separately. On its own labels Llama-3.3-70B gives $12/140$ against $55/140$, $+0.307\,[0.229, 0.386]$ with 45 against 2 and $p = 2 \times 10^{-11}$, and the consensus of all three annotators gives $0/55$ against $8/55 = 0.145$, $+0.145\,[0.055, 0.236]$ with 8 against 0 and $p = 0.008$, an exploratory analysis registered before it was run. The per-judge and two-judge readings of the seeded allocation are listed in Table 17.

The three annotators on one question set, an exploratory analysis. The estimates above cover different questions. On the 60 questions that all three annotators label in all three cells and label correct closed-book, the tax rises from $0/60$ to $9/60$ under the two-run annotation, $+0.150\,[0.067, 0.250]$ with 9 against 0 and $p = 0.004$, from $1/60$ to $10/60$ under Qwen3-32B, $+0.150\,[0.067, 0.250]$, and from $1/60$ to $11/60$ under Llama-3.3-70B, $+0.167\,[0.083, 0.267]$ with 10 against 0 and $p = 0.002$, against $+0.217\,[0.100, 0.333]$ under containment on the same questions. On the common sets of each pair of annotators, of 66, 100 and 75 questions, the two estimates differ by at most 0.010: $+0.167$ and $+0.167$ for the two-run annotation and Qwen3-32B, $+0.290\,[0.200, 0.380]$ and $+0.300\,[0.200, 0.400]$ for the two-run annotation and Llama-3.3-70B, and $+0.147$ and $+0.147$ for the two judges. These overlaps are selected by label availability, so the agreement describes the annotators on them only. The analysis was registered as exploratory before it was run.

Missing labels and a fixed target set, an exploratory analysis. Table 9 gives the reason for every missing label of the two-run annotation on the two paired runs. Most arise because at least one run flags the reference answer as inadequate for judging the response; the judges leave labels missing mainly on removal-set outputs, where Qwen3-32B flags 175 of their references and Llama-3.3-70B returns no usable label for 73. We fix a target set by the closed-book label alone, the questions whose closed-book answer the two-run annotation labels correct, 150 for Qwen3-8B and 202 for Llama-3.1-8B, and estimate the paired semantic tax difference on all of it. With the 20 Qwen3-8B and 24 Llama-3.1-8B missing with-document labels set to the containment label of the same output, the difference is $9/150$ against $60/150$, $+0.340\,[0.260, 0.420]$ with 52 against 1 discordant, and $22/202$ against $93/202$, $+0.351\,[0.282, 0.421]$ with 74 against 3. Filling the Qwen3-8B labels from Llama-3.3-70B and then Qwen3-32B leaves one question unlabelled and gives $9/149$ against $56/149$, $+0.315\,[0.242, 0.396]$ with 48 against 1. Within the target set, the containment effect is $+0.391\,[0.293, 0.481]$ on the 133 complete-case questions and $+0.471\,[0.235, 0.706]$ on the 17 others, and for Llama-3.1-8B $+0.412$ on 187 and $+0.333$ on 15: for Qwen3-8B the complete cases show a smaller lexical effect than the other target questions, for Llama-3.1-8B a larger one. On the 91 questions of the original allocation whose natural top-5 holds no answer-bearing passage the semantic tax is $8/20$, against $3/72$ in the keep arm and $25/67$ in the removal arm. These analyses were registered as exploratory before they were run.

Beyond the complete cases, an exploratory analysis. Each missing label of the two-run annotation is filled from Llama-3.3-70B and then Qwen3-32B, now also in the closed-book cell, so the target set is no longer restricted to questions with a known two-run closed-book label. For Qwen3-8B, 173 questions are labelled correct closed-book, 23 of them through a judge label, and 4 of the 360 closed-book answers remain unlabelled; on the 170 with both with-document labels the tax rises from $12/170$ to $66/170$, $+0.318\,[0.247, 0.394]$ with 55 against 1 and $p = 1.6 \times 10^{-15}$. On the 91 questions for which a run of the two-run annotation flags the reference in at least one cell, Llama-3.3-70B alone gives $2/23$ against $9/23$, $+0.304\,[0.130, 0.522]$ with 7 against 0 and $p = 0.016$, and

containment gives $3/22$ against $16/22$, $+0.591\,[0.364, 0.773]$. On the seeded allocation the same cascade gives $5/87 = 0.057\,[0.025, 0.128]$ in the keep arm and $35/83 = 0.422\,[0.321, 0.529]$ in the removal arm, a difference of $+0.364\,[0.244, 0.484]$ under a within-arm bootstrap, with 8 and 9 questions left unlabelled. These analyses were registered as exploratory before they were run. The cascade uses each judge's label as produced, without adjudicating the reference, so individual labels may rest on different readings of a time-dependent reference: one accepted closed-book answer lists a monarch with wrong reign dates, and the three answers to a question about a current office holder were judged against different dates. The intervals are conditional on the labels. The question-level adjudication below fixes the reading of every reference once.

Registered question-level adjudication of the reference. The same annotator, in two runs at a high reasoning setting with blind batches and the key withheld, received each of the 360 paired-design questions once, with its reference aliases and every distinct answer produced for it by Qwen3-8B in the closed-book, keep, removal, answer-mask and matched-mask cells and by Llama-3.1-8B in the closed-book, keep, removal and answer-mask cells: 2,434 answers in a per-question shuffled order, without model or arm. A written policy first fixes the accepted answers of the question, once. The question is read as of the reference; for a time-dependent question, such as a current office holder, a record or a latest event, the answer correct today is accepted as well; a reference that is wrong for the question as of that time is replaced by the correct answer; a question without a single defensible reading is marked ambiguous and all its answers stay unresolved. An answer is then correct if and only if it asserts an accepted answer and no conflicting one; a false qualifier attached to the answer, such as a wrong date, number, place or role, makes it incorrect. A label is known when both runs return valid, agreeing labels. The runs agree on the status of 338 questions, 237 adequate, 51 time-dependent, 34 with a replaced reference and 16 ambiguous; 28 questions are unresolved, and 2,142 of the 2,434 answers have a known label. The two runs agree at $\kappa = 0.957$, the adjudicated labels agree with the earlier two-run labels at $\kappa = 0.935$ and with alias-EM at $\kappa = 0.48$. Table 7 gives the registered estimates; the identification bounds are in Table 17. On the questions whose keep, removal and answer-mask labels are all known, 146 for Qwen3-8B and 227 for Llama-3.1-8B, the answer-mask effect is $0.37\,[0.18, 0.57]$ and $0.36\,[0.22, 0.51]$ of the removal effect under the adjudicated labels, against 0.62 and 0.65 under containment on the same questions; this exploratory ratio was registered before it was computed.

A registered second adjudicator. To check the adjudication independently of the annotator that produced it, a different proprietary LLM, the judge of the ladder validation above at its maximum reasoning setting, adjudicated the same blind items once, under the same policy and batch composition. It labelled 340 of the 360 questions; the labels of the other 20 are unknown and enter the identification bounds adversarially. It agrees with the two-run adjudication at $\kappa = 0.919$ on the 1,975 answers both label, and gives the same question status for 295 of the 319 questions that have an agreed two-run status and a status of its own. On its own labels the Qwen3-8B tax rises from $9/153$ to $58/153$, $+0.320\,[0.235, 0.405]$ with 54 against 5 and $p = 2 \times 10^{-11}$, with identification bounds $[0.036, 0.464]$, which meets the registered rule; for Llama-3.1-8B it gives $+0.330\,[0.260, 0.401]$ with bounds $[0.131, 0.433]$. On the answers both adjudications label alike, the Qwen3-8B difference is $+0.323\,[0.228, 0.417]$ on 127 questions, and the answer mask exceeds the matched mask by $+0.091\,[0.026, 0.156]$ under the second adjudicator and by $+0.117\,[0.058, 0.182]$ under the consensus.

Registered document support and semantic labels of the masked arms. With the passage rubric of the allocation run, two runs labelled whether each of the 1,080 five-passage sets of the keep, answer-mask and matched-mask arms supports the answer; the mask placeholder is visible and the arm is not stated. A support label is known when both runs agree and every supporting quote is verbatim. Support is $316/343 = 0.921$ for keep, $312/342 = 0.912$ for the matched mask and $126/335 = 0.376$ for the answer mask; on the sets known in both arms the matched-mask difference is $-0.006\,[-0.015, 0.000]$ with 2 against 0, inside the registered margin of ± 0.05, and the answer-mask difference is $-0.541\,[-0.596, -0.486]$ with 178 against 0 and $p = 5 \times 10^{-54}$. On the 150 questions of the two-run closed-book target the rates are 0.959, 0.953 and 0.503. The two-run answer annotation extended to all 2,160 answer-mask and matched-mask outputs, 612 of them new items, with $\kappa = 0.972$–0.975 between runs, gives on its complete cases an answer-mask minus matched-mask difference of $+0.109\,[0.051, 0.168]$ on 137 questions, with 17 against 2 and $p = 7 \times 10^{-4}$, answer mask minus keep of $+0.133\,[0.081, 0.193]$ for Qwen3-8B and $+0.153\,[0.100, 0.211]$ for Llama-3.1-8B, and matched mask minus keep of $+0.007\,[-0.021, 0.035]$.

Table 7: Paired tax differences under question-level adjudication of the reference, on the complete cases among the questions whose closed-book answer is adjudicated correct, 157 for Qwen3-8B and 234 for Llama-3.1-8B, with 95% question-bootstrap intervals. K, S, M, R: keep, matched mask, answer mask, removal. *Taxes*: second-named arm → first-named arm; *disc.*: questions lost only in the first-named arm against questions lost only in the other. *Adequate*: questions whose reference both runs mark adequate.

Run	Contrast	n	Taxes	Difference	Disc.
Qwen3-8B	R − K	147	$10/147 \to 56/147$	$+0.313\,[0.224, 0.408]$	53 : 7
Qwen3-8B	M − S	155	$10/155 \to 28/155$	$+0.116\,[0.052, 0.181]$	22 : 4
Qwen3-8B	M − K	154	$11/154 \to 28/154$	$+0.110\,[0.052, 0.169]$	21 : 4
Qwen3-8B	S − K	155	$11/155 \to 10/155$	$-0.006\,[-0.039, 0.026]$	3 : 4
Llama-3.1-8B	R − K	229	$20/229 \to 98/229$	$+0.341\,[0.275, 0.410]$	83 : 5
Llama-3.1-8B	M − K	228	$20/228 \to 48/228$	$+0.123\,[0.070, 0.175]$	34 : 6
Qwen3-8B, adequate	R − K	112	$4/112 \to 42/112$	$+0.339\,[0.250, 0.438]$	40 : 2
Qwen3-8B, adequate	M − S	114	$6/114 \to 18/114$	$+0.105\,[0.044, 0.175]$	14 : 2

Table 8 collects the semantic estimates of the removal effect in the two paired runs with the target set and label source behind each.

Table 8: Semantic estimates of the removal effect in the paired runs, as the removal − keep tax difference with 95% question-bootstrap interval, by the label that defines the closed-book-correct target and the labels used with documents. *Two-run*: agreed label of the two-run annotation; *cascade*: two-run label, else Llama-3.3-70B, else Qwen3-32B; *reg.*: a registered analysis, whose criterion and outcome are in Table 17; *expl.*: exploratory. The seeded allocation, whose semantic readings are between-arm differences, is reported in the text and in Table 17, with the identification bounds.

Run	Closed-book target	With-document labels	n	Difference	Status
Qwen3-8B	two-run	two-run, complete cases	133	$+0.331\,[0.248, 0.414]$	reg.
Llama-3.1-8B	two-run	two-run, complete cases	187	$+0.342\,[0.273, 0.417]$	reg.
Qwen3-8B	both judges	both judges, conservative merge	75	$+0.107\,[0.040, 0.173]$	reg.
Qwen3-8B	both judges	both judges, liberal merge	75	$+0.187\,[0.107, 0.280]$	reg.
Qwen3-8B	Llama-3.3-70B	Llama-3.3-70B	140	$+0.307\,[0.229, 0.386]$	expl.
Qwen3-8B	Qwen3-32B	Qwen3-32B	84	$+0.167\,[0.095, 0.250]$	expl.
Qwen3-8B	two-run = Qwen3-32B	two-run = Qwen3-32B	63	$+0.159\,[0.079, 0.254]$	expl.
Qwen3-8B	all three agree	all three agree	55	$+0.145\,[0.055, 0.236]$	expl.
Qwen3-8B	all three label	each annotator	60	$+0.150$ to $+0.167$	expl.
Qwen3-8B	two-run	two-run, else containment	150	$+0.340\,[0.260, 0.420]$	expl.
Llama-3.1-8B	two-run	two-run, else containment	202	$+0.351\,[0.282, 0.421]$	expl.
Qwen3-8B	two-run	cascade	149	$+0.315\,[0.242, 0.396]$	expl.
Qwen3-8B	cascade	cascade	170	$+0.318\,[0.247, 0.394]$	expl.
Qwen3-8B	flagged reference; Llama-3.3-70B	Llama-3.3-70B	23	$+0.304\,[0.130, 0.522]$	expl.
Qwen3-8B	adjudicated	adjudicated, complete cases	147	$+0.313\,[0.224, 0.408]$	reg.
Llama-3.1-8B	adjudicated	adjudicated, complete cases	229	$+0.341\,[0.275, 0.410]$	reg.
Qwen3-8B	adjudicated, adequate reference	adjudicated, complete cases	112	$+0.339\,[0.250, 0.438]$	reg.
Qwen3-8B	second adjudicator	second adjudicator	153	$+0.320\,[0.235, 0.405]$	reg.
Llama-3.1-8B	second adjudicator	second adjudicator	227	$+0.330\,[0.260, 0.401]$	reg.
Qwen3-8B	both adjudications agree	both adjudications agree	127	$+0.323\,[0.228, 0.417]$	reg.

Refusal labels for the four ladder configurations. The same two-run annotation was applied, blind to configuration and tier, to all 3,744 P2 closed-book and placeholder outputs of the four configurations, 3,391 unique items. One item without a label from the first run and three answer quotes of

Table 9: Reasons for missing labels of the two-run annotation on the paired runs, with 360 questions per cell. *Reference*: at least one run flags the reference answer; *invalid*: a run returns no valid label or a non-verbatim answer quote; *disagree*: valid labels that differ in correctness or answer status.

Run	Cell	Known	Reference	Invalid	Disagree
Qwen3-8B	Closed book	287	67	0	6
	Keep	286	68	2	4
	Removal	269	79	1	11
Llama-3.1-8B	Closed book	271	86	0	3
	Keep	283	71	1	5
	Removal	268	85	0	7

the second run that fail exact-quote validation stay unknown, leaving 3,344 items with an agreed answer status, against 3,347 before quote validation. A refusal label is used when both runs give REFUSAL; it does not depend on the reference, so reference flags affect only the correctness labels. Table 10 gives the results; intervals are question-bootstrap percentiles. By the registered rule listed in Table 17, an ordering of two configurations is stated only when their intervals do not overlap.

Table 10: The P2 placeholder tier under blinded semantic labels. *Known*: closed-book answer labelled correct by both runs. *Over-refusal*: known questions whose placeholder-tier output both runs label a refusal. κ: agreement of each lexical rule with the semantic refusal label over the placeholder outputs with an agreed status.

Configuration	Known	Over-refusal	Semantic tax	κ phrase	κ clause
Qwen3-8B	178	$58/178 = 0.326\,[0.258, 0.393]$	$73/169 = 0.432$	0.867	0.966
Mistral-7B-v0.3	209	$61/209 = 0.292\,[0.230, 0.354]$	$77/201 = 0.383$	0.504	0.893
Llama-3.1-8B	253	$112/249 = 0.450\,[0.390, 0.510]$	$145/240 = 0.604$	0.059	0.900
Qwen3-32B	223	$77/222 = 0.347\,[0.284, 0.410]$	$87/214 = 0.407$	0.304	0.911

C THE QWEN3-8B LADDER, STRICT-EM TAXES AND THE QWEN SIZE CONTRAST

Table 11: Paired tax, the fraction of closed-book-correct answers lost, and net change in alias-EM relative to the matched closed-book cell, P2 ladder, all 468 questions. Ph: placeholder; R30/R60: recall30/recall60; Or1: oracle1. n_{cb} is the closed-book-correct denominator of the tax. Net-change intervals are shown in Figure 2; semantic readings of the placeholder tier are in Table 10.

		Tax				Net ΔAlias			
Configuration	n_{cb}	Ph	R30	R60	Or1	Ph	R30	R60	Or1
Qwen3-8B	161	0.447	0.385	0.273	0.186	-0.118	-0.017	$+0.128$	$+0.276$
Mistral-7B-v0.3	178	0.427	0.360	0.281	0.157	-0.120	-0.049	$+0.060$	$+0.192$
Llama-3.1-8B	204	0.623	0.500	0.314	0.206	-0.246	-0.132	$+0.002$	$+0.124$
Qwen3-32B	183	0.448	0.366	0.230	0.153	-0.139	-0.045	$+0.077$	$+0.246$

With one oracle passage, P2 and P3 reach similar accuracy in every configuration, $0.620/0.677$ for Qwen3-8B, $0.573/0.562$ for Mistral, $0.560/0.575$ for Llama-3.1-8B and $0.637/0.660$ for Qwen3-32B, and the removal control of Section 4.4 compares retrieved document sets directly. Table 13 compares the four run configurations at the placeholder tier. Across Qwen3-8B and Qwen3-32B the placeholder tax is flat at 0.447 and 0.448, with ΔAlias -0.118 and -0.139, while the context-phrase hit rate falls from 0.534 to 0.141. Under the clause detector the same two runs give 0.481 and 0.391, a $1.2\times$ rather than $3.8\times$ change: the larger run abstains almost as often and cites the context less. With scaffolding varying alongside size, this is an association within Qwen on NQ.

Table 12: Qwen3-8B under P2 on the same 468 NQ questions, relative to the matched closed-book cell. Injected% is the fraction of questions given an answer-bearing passage. Non-answer documents are placeholder passages, so the lower tiers describe uninformative context rather than weak retrieval.

Tier	Injected%	Alias-EM	ΔAlias	CCR
Closed book	0.0	0.344	0.000	0.000
Placeholder	0.0	0.226	−0.118	0.534
Recall30	28.6	0.327	−0.017	0.429
Recall60	60.7	0.472	+0.128	0.342
Oracle1	100.0	0.620	+0.276	0.109
Oracle5	100.0	0.626	+0.282	0.201

Table 13: Run configurations under one P2 ladder, differing in checkpoint and scaffolding together. Tax is the fraction of closed-book-correct answers lost at the placeholder tier. $\mathrm{CCR}_{\mathrm{ph}}$ is the context-phrase detector's hit rate over all 468 outputs, and Table 2 shows what it counts; $\Delta\mathrm{Alias}_{\mathrm{or1}}$ compares oracle1 with the matched closed-book cell.

Configuration	n_{cb} strict/alias	Strict tax	Alias tax	$\mathrm{CCR}_{\mathrm{ph}}$	$\Delta\mathrm{Alias}_{\mathrm{or1}}$
Qwen3-8B	22 / 161	0.773	0.447	0.534	+0.276
Mistral-7B-v0.3	39 / 178	0.641	0.427	0.222	+0.192
Llama-3.1-8B	136 / 204	0.676	0.623	0.041	+0.124
Qwen3-32B	85 / 183	0.600	0.448	0.141	+0.246

D A REAL ANSWER-BEARING PASSAGE AGAINST A PLACEHOLDER BACKGROUND

The intermediate tiers contain a within-tier comparison of the answer-bearing and the other questions under the hash allocation of the ladder, with 936 records per contract. Table 14 reports the observed contrasts. Alias-EM is higher with an answer-bearing passage in all four cells, with intervals excluding zero, and the observed tax is lower in all four. At recall30 the two groups have the same closed-book accuracy, with Δcb near zero, so the recall30 rows compare balanced groups; at recall60 Δcb is $+0.08$. The comparison concerns replacing a placeholder with a real answer-bearing passage; Section 4.4 uses genuine passages on both sides.

Table 14: Answer-bearing minus other questions within one tier of the Qwen3-8B ladder, as observed differences, with 95% question-bootstrap intervals for ΔAlias in brackets. Groups are $134/334$ at recall30 and $284/184$ at recall60. Δcb is the same contrast on the matched closed-book cell.

Cell	ΔAlias	ΔTax	ΔCCR	Δcb
P2/r30	+0.378 [.28, .48]	−0.265	−0.361	−0.001
P2/r60	+0.465 [.39, .54]	−0.313	−0.341	+0.083
P3/r30	+0.323 [.23, .42]	−0.145	+0.021	+0.003
P3/r60	+0.391 [.31, .48]	−0.058	+0.011	+0.079

E CLOSED-BOOK AND RETRIEVAL ANCHORS ON THE SAME OUTPUTS

Table 15 computes retrieval-anchored and closed-book-anchored quantities on the same saved outputs. With a shared correctness label and the closed-book cell as the anchor, the change in accuracy is a gain term minus a loss term by the identity of Appendix J, and a retention of closed-book-correct answers equals $1 - \tau$. A retention anchored at a retrieval cell instead conditions on answers that are already correct with documents, so the anchor decides which losses a per-question metric can count. A RAD-style retention from oracle1 to oracle5 lies between 0.88 and 0.93 in all four configurations, which measures robustness to added documents; the closed-book anchor answers a different question: the same runs have already lost 15–21% of their closed-book-correct answers at

oracle1, and those questions are outside the RAD denominator by construction. The two estimands are complementary. In the factorial the RAD-style retention is 0.90–0.93 for every contract, including P1, whose closed-book cell answers nothing; only the closed-book anchor shows that P1 has no parametric answer to lose. Finally, measuring known-question over-refusal with the context-phrase detector ranks Llama-3.1-8B lowest of the four configurations at 0.020 and Qwen3-8B highest at 0.398; the broader clause rule ranks Llama-3.1-8B highest at 0.446. The two metrics are computed on identical outputs, and Table 2 shows that 88 of Llama-3.1-8B's 127 lost answers are generic declines against none for Qwen3-8B. None of these contrasts requires a new generation run.

Table 15: Retrieval-anchored and closed-book-anchored metrics on the same saved outputs under P2 and alias-EM, 468 questions. *RAD-style*: among questions correct at oracle1, the fraction still correct at oracle5, where four placeholder passages are added, following the construction of Park & Lee (2024) with placeholder rather than adversarial passages. *Lost at oracle1*: closed-book-correct answers already lost at the retrieval anchor, which a retrieval-anchored metric cannot count. *Over-refusal*: among closed-book-correct questions, the fraction flagged as declining at the placeholder tier, as in the known-question over-refusal of Zhou et al. (2026), under the context-phrase detector and the clause detector. Bottom rows: the factorial, where the retrieval anchor is $T=1$.

	RAD-style retention	Lost at oracle1	Over-refusal at placeholder	
Configuration	oracle1→oracle5	closed-book anchor	Phrase	Clause
Qwen3-8B	$271/290 = 0.934$	$30/161 = 0.186$	$64/161 = 0.398$	$54/161 = 0.335$
Mistral-7B-v0.3	$237/268 = 0.884$	$28/178 = 0.157$	$33/178 = 0.185$	$53/178 = 0.298$
Llama-3.1-8B	$235/262 = 0.897$	$42/204 = 0.206$	$4/204 = 0.020$	$91/204 = 0.446$
Qwen3-32B	$263/298 = 0.883$	$28/183 = 0.153$	$24/183 = 0.131$	$62/183 = 0.339$
	RAD-style $T{=}1 \to 5$	Closed-book alias-EM	Tax at $T{=}5$, closed-book anchor	
Factorial P1	$183/203 = 0.901$	0.000	undefined, 0 closed-book correct	
Factorial P2	$210/230 = 0.913$	0.382	$25/179 = 0.140$	
Factorial P3	$233/257 = 0.907$	0.387	$17/181 = 0.094$	
Factorial P4	$223/240 = 0.929$	0.387	$17/181 = 0.094$	

F IGNORE-CONTEXT AND PERMISSIVE-CONTRACT DIAGNOSTICS

Under placeholder context CCR fires on 0.058 of P3 outputs against 0.534 under P2 for Qwen3-8B. Since CCR is validated under P2 and the contracts have different closed-book baselines and answer lengths, we read this as a hit-rate difference. In the factorial the regex fires on a single P4 output at $T = 0$ and on none at $T = 1$ or $T = 5$, while P4's alias-EM climbs from 0.387 to 0.603 with five documents and 9.4% of its closed-book-correct answers are lost. The instruction removes context-phrase refusals and leaves accuracy dependent on the documents. Because the factorial is not stratified by whether the supplied documents contain the answer, this shows that the documents change P4's outputs, not which part of them does so.

G EXACT PROMPT CONTRACTS

Each contract is a system prompt. The message that follows it is "Context:" followed by the documents, each introduced as "[Doc i]", then "Question:" with the question and "Answer:"; with no documents the context block is omitted. In the factorial, the Qwen3-8B removal-control, paired and masking runs and the Qwen3-8B ladder nothing else is added; the model's native thinking block is closed at the 256-token reasoning budget, which binds in at least 96% of factorial records in every cell. In the generic chain-of-thought runs, namely the Qwen3-32B, Llama-3.1-8B and Mistral-7B ladder runs and the Llama-3.1-8B, Mistral-7B and Qwen3-32B paired runs, the system prompt is followed by a space and the sentence "Think step by step briefly."; after at most 256 reasoning tokens the answer is generated from the reasoning followed by two line breaks and the phrase "Therefore, the short answer is:". The path of these three ladder runs is established from their 16,848 saved records, each of which records it. In the factorial, the presence figures of Section 4.5, 81.0% and 77.4%, are computed from the passages supplied in the run, which the released records identify by

their corpus identifiers; an independent exact re-retrieval over the same corpus gives 81.2%. Mean emitted think tokens in the ladder are 225.8 for Qwen3-8B against 230.1, 59.8 and 51.9 for the Qwen3-32B, Llama-3.1-8B and Mistral-7B runs. In all runs the answer is the text after any closing think tag, with a leading "Answer:", "Final Answer:", "The answer is" or "A:" removed, cut at the first line break and stripped of a final period. No extracted answer is empty in any factorial cell at $T \in \{0, 1, 5\}$, but the first-line cut can end an answer at a list preamble, as in two of P1's 38 strict losses. The original removal control was run with a 16,384-token input limit, its registered paired runs and the ladder with 4,096. Tokenized with the Qwen3-8B tokenizer and chat template, system prompt included, the longest first-stage removal-control prompt is 1,596 tokens over the 360 assigned document sets and 1,837 over all 720 sets of the paired design. Rebuilding each paired run's prompts with its own tokenizer, chat template and generation path, for the closed book and the keep, removal and mask sets, and adding the 256-token reasoning budget and the closing string bounds the second stage at 2,096 tokens for Qwen3-8B, 1,981 for Llama-3.1-8B, 2,322 for Mistral-7B and 2,107 for Qwen3-32B, a computed bound rather than a measured length, so no limit binds in any of these runs. In the removal control the reasoning budget binds for 52% of keep-arm and 76% of removal-arm outputs with documents; within either stratum of the with-document output, budget reached or not, the containment tax is far higher in the removal arm, $34/54$ against $6/37$ and $8/13$ against $3/31$, although this split conditions on a post-treatment variable.

- **P1**: "Answer the question based ONLY on the provided documents. If the documents do not contain the answer, say so."
- **P2**: "Answer the question. Give a short, direct answer."
- **P3**: "Answer the question using any knowledge available to you. You may use the provided documents if helpful, but also use your own knowledge. Give a short, direct answer."
- **P4**: "Ignore the documents provided below; they were retrieved automatically and may be irrelevant. Answer the question from your own knowledge only. Give a short, direct answer."

H REFUSAL DETECTION RULES

The context-phrase detector, CCR, fires when any of the following 22 phrases occurs in the generation, ignoring case: "not contain", "does not contain", "do not contain", "cannot answer", "can't answer", "cannot be answered", "can't be answered", "no information", "not enough information", "insufficient information", "does not provide", "do not provide", "not mentioned", "not provided", "no relevant", "unable to determine", "unable to answer", "cannot determine", "can't determine", "not possible to determine", "the provided" followed by document, documents, passage, passages, context, text or texts and then by "do not" or "does not", and "none of the provided" or "none of the given" followed by document, documents, passage or passages.

None of the 22 patterns covers "I don't know". Several of them, such as "cannot answer" and "unable to determine", do not name the context, so CCR is a rule for explicitly worded declines; the context-attributed versus generic split of Table 2 is assigned separately, from the wording of the hedge, as described below. A variant of the detector adds a 23rd pattern, "based on the provided" or "based on the given" followed by document, documents, passage, passages or context, which requires no negation. Excluding it changes no Qwen3-8B P2 ladder label, so the validation of Appendix B carries over unchanged; it changes the placeholder-tier lost-answer rates only for Mistral, from 0.408 to 0.368, and leaves the $31.8\times$ ratio unchanged. In the factorial, where P1 answers often open with "Based on the provided documents", it lowers the P1 hit rate from 0.647 to 0.382 at $T = 1$ and from 0.669 to 0.156 at $T = 5$; every CCR number in this paper uses the 22 patterns. The clause detector splits the generation into clauses, marks those matching a hedge or inability cue, and counts the output a refusal only if no surviving clause asserts an entity, number, date or yes/no answer; a clause introduced by an answer-after-hedge connector such as "however" or "the answer is" survives. The three-way split of Table 2 labels a declining output as *context-attributed* when its hedge refers to the documents, passages or context, *generic* otherwise, and *decline-then-answer* when a surviving clause asserts an answer. Over all 5,616 records of each run, generic declines without an answer number 900 for Llama-3.1-8B against 4 for Qwen3-8B, 15 for Qwen3-32B and 87 for Mistral, while context-attributed declines number 96, 783, 603 and 906. With the 23rd pattern included, the detector reproduces the labels of that variant on all 22,464 ladder records of these four configurations. On the

placeholder-tier lost subsets of Table 2, the phrase rule matches 51 of 52 context-attributed declines for Qwen3-8B, 26 of 47 for Mistral, 21 of 62 for Qwen3-32B and 0 of 2 for Llama-3.1-8B, and 3 of Llama's 88 generic declines. The clause rule is a heuristic: its hedge list omits bare declines such as "Unknown" and "I'm not sure", which it therefore counts as answers, at least seven such outputs on the Mistral and Llama lost subsets.

Table 16: Verbatim P2 placeholder-tier outputs on two questions that Qwen3-8B and Llama-3.1-8B both answer correctly in the closed-book cell. The context-phrase detector fires on the Qwen3-8B outputs only; the clause detector counts all four as refusals.

Question; gold answer	Qwen3-8B	Llama-3.1-8B
who does the voice of nala in the lion king; Moira Kelly, ...	"The provided documents do not contain information about the voice actor for Nala in *The Lion King*"	"I don't know"
how many seasons of vampire diaries r there; eight	"The provided documents do not mention 'Vampire Diaries' or related information. Therefore, the answer cannot be determined from the given context"	"I don't know"

I PLACEHOLDER CONSTRUCTION AND ANSWER-PRESENCE ACCOUNTING

All $10{,}000$ entries match one template, *"This is a synthetic passage about topic i. It contains information relevant to queries about subject j."* For 71 of the 468 questions the gold answer is a bare number that also appears as a topic or subject index, which could either raise containment or draw the answer toward a competing number. Answer-bearing passages from the 500K pool are selected by character-substring containment of a normalized gold alias. We measure document-side presence with a separate token-span rule over lower-cased runs of letters and digits; unlike the answer scorer, which deletes punctuation, it splits possessives and footnote markers from the word. Under this rule 462/468 of the passages contain a gold alias, the six exceptions being artifacts such as *George II* inside *George III*; without these six questions every sign and significance statement of Section 4.2 holds on the remaining 462, for example a Qwen3-8B tax of $72/159 = 0.453$, 0.384, 0.277 and 0.182 across the four tiers and a Mistral-7B recall30 net change of -0.048 with $p = 0.039$. The measured answer-bearing fractions of the intermediate tiers are 28.2% and 60.3% against injection rates of 28.6% and 60.7%. With the 15.2% collision ceiling, answer presence lies in ≤ 0.152, $[0.282, 0.434]$ and $[0.603, 0.754]$ across the placeholder, recall30 and recall60 tiers, so the tier ordering holds under every consistent assignment. Dropping the 71 collision questions leaves the placeholder tax essentially unchanged at $60/135 = 0.444$ for Qwen3-8B, $59/142 = 0.415$ for Mistral, $97/163 = 0.595$ for Llama-3.1-8B and $61/148 = 0.412$ for Qwen3-32B, each with a significantly negative net change at $p < 2 \times 10^{-6}$.

J THE ACCOUNTING IDENTITY

Claim. Let N questions be scored by one correctness label in a closed-book cell and in a with-document cell of the same contract. Let K be the number correct closed-book, L the number of these that are incorrect with documents, and G the number of closed-book-incorrect questions that are correct with documents. With $a = K/N$, with $\tau = L/K$ for $K > 0$ and with $\rho = G/(N - K)$ for $K < N$, the with-document accuracy A_t satisfies

$$A_t = a - a\tau + (1 - a)\rho,$$
$$A_t - a = (1 - a)\rho - a\tau.$$

Proof. Each question is in exactly one of four closed-book/with-document states: $K - L$ questions are correct/correct, L are correct/incorrect, G are incorrect/correct and $N - K - G$ are incor-

rect/incorrect. The with-document correct questions are the first and third groups, so

$$\begin{aligned} A_t &= \frac{(K-L)+G}{N} \\ &= \frac{K}{N} - \frac{K}{N}\cdot\frac{L}{K} + \frac{N-K}{N}\cdot\frac{G}{N-K} \\ &= a - a\tau + (1-a)\rho. \end{aligned}$$

Subtracting a gives the second line. The identity is an exact count over the evaluated questions; it needs no assumption on how losses and gains are related. If $K = 0$ the loss term is 0 and τ is undefined, which is the case of P1 in Table 4; if $K = N$ the gain term is 0 and ρ is undefined. □

K REGISTERED CONFIRMATORY ANALYSES

Table 17 lists the confirmatory analyses registered before the corresponding runs or labelling, with their criteria and results. Each registration is a dated entry in the authors' own records, written before the run or labelling it governs.

Table 17: Registered confirmatory analyses. Rule of the semantic removal-control analyses: the removal effect is described as semantic if the criteria of both rows are met, and otherwise stated for alias-EM with the semantic estimates reported. Identification bounds fill the pairs without an agreed label adversarially.

Analysis	Criterion	Result
Within-question removal, Qwen3-8B paired run: containment tax difference, removal − keep	> 0, McNemar $p < 0.01$, 95% interval excludes 0	$+0.459\,[0.370, 0.548]$, $p = 6 \times 10^{-17}$
Within-question removal, Llama-3.1-8B paired run: the same	the same	$+0.448\,[0.368, 0.529]$, $p = 3 \times 10^{-19}$
Within-question removal, Mistral-7B-v0.3 paired run: the same	the same	$+0.353\,[0.275, 0.431]$, $p = 1 \times 10^{-14}$
Within-question removal, Qwen3-32B paired run: the same	the same	$+0.375\,[0.289, 0.461]$, $p = 2 \times 10^{-15}$
Alias-span masking, Qwen3-8B paired run: containment tax difference, mask − keep	the same	$+0.230\,[0.156, 0.311]$, $p = 4 \times 10^{-8}$
Alias-span masking, Qwen3-8B paired run, secondary: strict-EM tax difference	reported	$+0.294\,[0.000, 0.588]$, $p = 0.125$; $6/17$ vs. $11/17$
Alias-span masking, Llama-3.1-8B paired run: containment tax difference, mask − keep	> 0, McNemar $p < 0.01$, 95% interval excludes 0	$+0.241\,[0.161, 0.322]$, $p = 3 \times 10^{-8}$
Matched non-answer mask, Qwen3-8B paired run: containment tax difference, answer mask − matched mask	the same	$+0.230\,[0.156, 0.311]$, $p = 1 \times 10^{-7}$
Reasoning budget of 1,024 tokens, Qwen3-8B, 180 paired questions: containment tax difference, removal − keep	> 0, McNemar $p < 0.01$, 95% interval excludes 0	$+0.493\,[0.343, 0.627]$, $p = 4 \times 10^{-8}$
Reasoning budget of 1,024 tokens: containment tax difference, mask − keep	the same	$+0.194\,[0.090, 0.299]$, $p = 0.002$
Reasoning budget of 1,024 tokens, secondaries: strict-EM tax differences on 13 closed-book-correct questions, removal − keep; mask − keep	reported	$+0.385\,[0.000, 0.769]$, $p = 0.18$; $+0.231\,[-0.077, 0.538]$, $p = 0.38$
Removal control: semantic tax difference, removal − keep, agreed labels	> 0, 95% interval excludes 0	$+0.331\,[0.204, 0.461]$
Removal control: identification bounds of the same difference	lower bound > 0	$[-0.085, 0.604]$
Paired runs: semantic tax difference, complete cases, Qwen3-8B; Llama-3.1-8B	> 0, McNemar $p < 0.01$, 95% interval excludes 0	$+0.331\,[0.248, 0.414]$; $+0.342\,[0.273, 0.417]$
Paired runs: identification bounds of the same differences	lower bound > 0	$[-0.079, 0.571]$; $[-0.027, 0.568]$

Table 17, continued.

Analysis	Criterion	Result
Two open-weight judges, Qwen3-32B and Llama-3.3-70B, Qwen3-8B paired run: semantic tax difference on questions both judges label, under the conservative and the liberal merge rule	> 0, 95% interval excludes 0, both rules	$+0.107\,[0.040, 0.173]$; $+0.187\,[0.107, 0.280]$
Two open-weight judges, seeded allocation: the same	the same	$+0.063\,[-0.020, 0.162]$; $+0.178\,[0.045, 0.320]$
Each open-weight judge alone, seeded allocation: per-arm tax and difference, Qwen3-32B; Llama-3.3-70B	reported	$3/68$ vs. $6/42$, $+0.099\,[-0.012, 0.222]$; $6/79$ vs. $33/72$, $+0.382\,[0.251, 0.510]$
Within-question removal under P3, Qwen3-8B: containment tax difference	> 0, McNemar $p < 0.01$, 95% interval excludes 0	$+0.331\,[0.245, 0.417]$, $p = 4 \times 10^{-11}$
Within-question removal under P3, Qwen3-8B, secondary: strict-EM tax difference	reported	0.000; $3/4$ vs. $3/4$
Placeholder tier: over-refusal ordering of the six configuration pairs	intervals disjoint	Llama-3.1-8B > Mistral-7B-v0.3; other five pairs overlap
Question-level reference adjudication, Qwen3-8B paired run: tax difference, removal − keep, complete cases, with identification bounds	> 0, McNemar $p < 0.01$, 95% interval excludes 0, lower bound > 0	$+0.313\,[0.224, 0.408]$, $p = 8 \times 10^{-10}$; bounds $[0.074, 0.463]$
Question-level reference adjudication, Qwen3-8B: answer mask − matched mask, the same	the same	$+0.116\,[0.052, 0.181]$, $p = 5 \times 10^{-4}$; bounds $[-0.065, 0.279]$
Question-level reference adjudication, secondaries: Qwen3-8B answer mask − keep; matched mask − keep; Llama-3.1-8B removal − keep; answer mask − keep; bounds, with estimates in Table 7	reported	$[-0.080, 0.270]$; $[-0.176, 0.160]$; $[0.175, 0.434]$; $[-0.015, 0.244]$
Question-level reference adjudication, secondaries: the two primary contrasts on adequate-reference questions; bounds	reported	$[0.294, 0.383]$; $[0.083, 0.150]$
Two-run labels of the answer-mask and matched-mask outputs, Qwen3-8B: answer mask − matched mask, complete cases, with identification bounds	> 0, McNemar $p < 0.01$, 95% interval excludes 0, lower bound > 0	$+0.109\,[0.051, 0.168]$, $p = 7 \times 10^{-4}$; bounds $[-0.194, 0.429]$
Two-run labels, secondaries: answer mask − keep for Qwen3-8B; for Llama-3.1-8B; matched mask − keep; bounds	reported	$[-0.212, 0.433]$; $[-0.169, 0.397]$; $[-0.301, 0.282]$
Second adjudicator, Qwen3-8B paired run: tax difference, removal − keep, with identification bounds	> 0, McNemar $p < 0.01$, 95% interval excludes 0, lower bound > 0	$+0.320\,[0.235, 0.405]$, $p = 2 \times 10^{-11}$; bounds $[0.036, 0.464]$
Second adjudicator, secondaries: consensus removal − keep; answer mask − matched mask under the second adjudicator; under the consensus; Llama-3.1-8B removal − keep; bounds	reported	$[-0.101, 0.561]$; $[-0.138, 0.286]$; $[-0.224, 0.393]$; $[0.131, 0.433]$
Document support: matched mask − keep, sets known in both arms	95% interval inside $[-0.05, 0.05]$	$-0.006\,[-0.015, 0.000]$
Document support: answer mask − keep, the same	< 0, McNemar $p < 0.01$, 95% interval excludes 0	$-0.541\,[-0.596, -0.486]$, $p = 5 \times 10^{-54}$