BEYOND EVIDENCE AVAILABILITY: INTERFERENCE AND REASONING DIAGNOSTICS IN LONG-CONTEXT QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Longer contexts let models access more information, including potentially irrelevant or misleading text. We examine how background content affects question answering across context lengths and reasoning settings. Six models at two reasoning settings answer questions requiring information from several documents, with all annotated supporting passages retained in inputs of about 25,000 or 112,000 tokens. Stage 1 compares background sampled from other questions, related passages filtered to exclude answer mentions, and similar related background with a short passage supporting a designated wrong answer. The passage averages less than 0.2% of the input. We first measure the accuracy loss when sampled background is replaced by related text. We then measure the additional loss under similar background containing the misleading passage. This second loss is larger in nine of 12 model–reasoning combinations, with 95% confidence intervals for the difference above zero. Stage 2 examines cases that the same model answers incorrectly with little or no reasoning but correctly with higher reasoning. Keeping the full input, we provide the lower setting with document numbers from the successful higher-setting answer, numbers of the annotated supporting documents, or the first set plus guidance on using the evidence. On selected cases passing guidance checks with valid responses to all three prompts, the guidance-containing prompt exceeds the prompt using the same document numbers alone by 25.1–39.3 percentage points in five of six models, with positive 95% confidence intervals.

1 INTRODUCTION

Longer context windows let models use more retrieved passages and tool observations (Lewis et al., 2020; Yao et al., 2023), supporting diverse tasks and deeper reasoning (Yen et al., 2025; Bai et al., 2025). These inputs mix useful evidence with irrelevant or misleading text. We ask: *how does background content affect accuracy across context lengths when all supporting evidence is retained?*

Stage 1 uses MuSiQue questions that require combining information across documents. We keep the question, correct answer, and supporting passages fixed and compare three backgrounds: passages sampled from other questions; related passages filtered to exclude answer mentions; and similar related text with one short passage supporting a designated wrong answer. We test six models at two reasoning settings and two input lengths, about 25,000 and 112,000 tokens. Prior work highlights the role of informative inputs in learning from noisy data (Liu & Jin, 2026). We extend this line of inquiry to evidence use in long contexts at inference time.

Stage 2 asks: *what does higher reasoning offer on these tasks?* We revisit cases where the same model is wrong with little or no reasoning but correct with higher reasoning. Keeping the full input, we give the lower setting either document numbers from the successful higher-setting answer, the numbers of the annotated supporting documents, or the first set plus guidance generated by the same model under higher reasoning. The two stages provide the following contributions:

- **Background interference in long contexts.** We compare two accuracy losses: the loss when sampled background is replaced by related text, and the additional loss under similar background containing a misleading passage. The second loss is larger in nine of 12 com-

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.

binations of model and reasoning setting, with 95% confidence intervals for the difference above zero. We also track whether answers move toward the designated wrong answer, including changes from one wrong answer to another.

- **Probing reasoning gains through evidence guidance.** On selected cases from all three backgrounds, the prompt with guidance exceeds the prompt using the same document numbers alone by 25.1–39.3 percentage points in five models, with positive 95% confidence intervals. Follow-up analyses also find gains when the supplied document numbers already match the annotations exactly.

Primary results use examples passing two model-based passage checks. Stage 2 compares the same cases passing guidance checks, with valid responses under all three prompts. A smaller 2WikiMultiHopQA study (Ho et al., 2020) checks the misleading-passage effect on another dataset.

2 Related Work

Context scale and distractor difficulty. Long-context studies probe evidence position (Liu et al., 2024), task diversity (Bai et al., 2024), and retrieval or reasoning complexity (Hsieh et al., 2024; Roberts et al., 2025). Cuconasu et al. (2024) find that non-answering retrieved passages can hurt RAG accuracy while random documents can improve it. Jin et al. (2025) compare random and retrieved hard negatives while retaining a gold passage in a controlled RAG experiment. Other controls isolate document count (Levy et al., 2025) and hard-distractor proportion (Gao et al., 2026) at fixed length, revealing nonlinear degradation for the latter. At 25K and 112K tokens, we compare the accuracy cost of replacing sampled background with related passages against the additional cost of a short misleading passage, retaining all annotated support.

Targeted distractors and conflicting context. Knowledge-conflict studies examine parametric–contextual disagreement (Longpre et al., 2021; Neeman et al., 2023). Adversarial QA adds misleading text (Jia & Liang, 2017); QACC studies conflicting retrieved sources (Liu et al., 2025), and CRiT-QA combines counterfactual chains with distractor traps (Yun et al., 2026). Bhuiya et al. (2024) retain gold HotpotQA paragraphs while replacing non-gold content with alternative multi-hop chains, studying chain structure and chain-of-thought prompting. PoisonedRAG targets designated answers through retrieval poisoning (Zou et al., 2025); RobustRAG provides robustness certificates under bounded retrieval corruption using isolated passage groups and response aggregation (Xiang et al., 2026). Stage 1 retains the original question, answer, and complete support to compare the accuracy costs of related background and a short wrong-answer passage. Paired answer transitions further describe changes in error direction.

Reasoning and evidence guidance. NoLiMa tests retrieval beyond literal matching (Modarressi et al., 2025); GSM-IC and GSM-DC test arithmetic distraction (Shi et al., 2023; Yang et al., 2025). NoisyBench studies distractor types and reasoning behavior, and evaluates rationale-aware training (Lee et al., 2026). Mitigations include mixed-context training (Yoran et al., 2024), adaptive retrieval and self-critique (Asai et al., 2024), context compression (Xu et al., 2024), and document-level reading notes (Yu et al., 2024). Stage 2 uses evidence prompts to probe same-model reasoning gains. For cases answered correctly only at the higher setting, we test how far document pointers, alone or with guidance on using the evidence, help the lower setting answer with the full input retained. Table 3 in Appendix A details the experimental differences.

3 Experimental Design

3.1 Support-Preserving Context Construction

For question q, canonical answer y^*, complete annotated support G, length L, and non-gold target t, we construct neutral (N), topically related (hard; H), and targeted-distractor (D) contexts:

$$C_N(L) = G \cup N_L, \tag{1}$$

$$C_H(L) = G \cup H_L, \tag{2}$$

$$C_D(L) = G \cup \widetilde{H}_L \cup \{F_t\}. \tag{3}$$

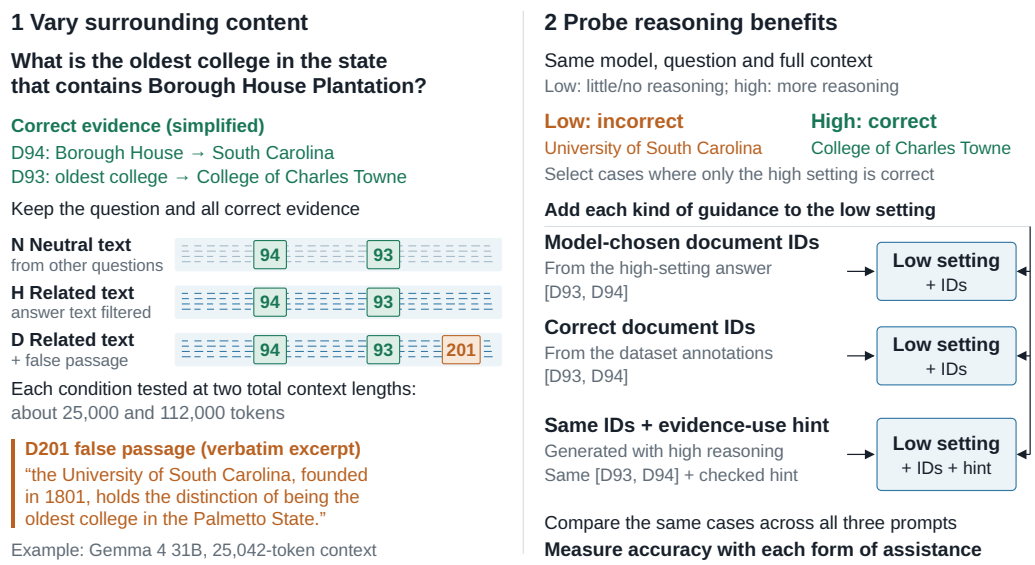


Figure 1: Two-stage study with a MuSiQue example. Left: each background is tested at both lengths with all supporting passages retained. The related backgrounds without (H) and with (D) a misleading passage are nearly matched. Evidence chains and context strips are schematic; the false-passage excerpt is verbatim. Right: cases answered correctly only at the higher setting receive three separate prompts at the lower setting, retaining the full input. Document numbers come from the successful answer or benchmark annotations. A separate call at the higher setting generates the hint with the model-chosen numbers fixed.

N_L contains filtered passages sampled from other MuSiQue examples. H_L contains BM25-retrieved passages (Robertson & Zaragoza, 2009) topically related to the question and gold titles. $\widetilde{H}_L$ denotes the related background used in D; it follows the same retrieval policy as H_L, with small differences in the selected passages after context packing. We exclude background passages from the question's source example or with string matches to the canonical answer, recorded aliases, and gold titles; Appendix B specifies their short-string exceptions. F_t is one short synthetic target-supporting distractor designed to support t at the final answer-bearing relation. The question, canonical answer, and every annotated gold paragraph remain fixed across conditions. The source examples and annotated support chains come from MuSiQue (Trivedi et al., 2022).

We pack contexts offline with `cl100k_base`, the common tokenizer used to report input length across models. The shorter regime averages 25.08K tokens; the longer averages 112.09K. The complete-population target-supporting distractor averages 45.49 construction-tokenizer tokens and occupies 0.1814% of the shorter input and 0.0406% of the longer input. The gold-position category is balanced across front, middle, end, and random assignments and held fixed across variants of an example. Within the random category, variant-specific seeds determine the realized gold locations. We pack the two scales separately.

A frozen composition audit of H_L and $\widetilde{H}_L$ finds mean filler-set Jaccard overlap of 0.9982 at 25K and 0.9996 at 112K, with the order of common fillers preserved in all pairs; exact set equality is 67.6% and 68.0%, respectively. The contrast estimates the effect of the targeted-distractor construction as a whole, including variation in randomized gold placement. Figure 1 summarizes the diagnostic interventions and the three constructions; Appendix B gives exact composition and audit details.

3.2 Datasets, Models, and Reasoning Settings

Our main sample contains 500 answerable examples from the MuSiQue-Ans development split, selected with seed 20260516: 200 two-hop, 200 three-hop, and 100 four-hop questions. We cap each normalized answer at two selected examples. Every example has three conditions at both scales, producing 3,000 frozen contexts. The six primary models are Gemma 31B, DeepSeek V4 Flash,

Nemotron Super, Nemotron Nano, Qwen 27B, and Qwen 35B-A3B. Gemma 26B-A4B is retained in the appendix because the provider trace does not cleanly establish execution of the requested high-thinking mode.

We evaluate each model at temperature zero under disabled-or-minimal reasoning and provider-exposed high reasoning. Provider controls are heterogeneous: Gemma exposes minimal versus high thinking, DeepSeek exposes disabled versus enabled/max reasoning, Nemotron exposes a verified zero-budget setting versus a 4096-token reasoning budget, and Qwen exposes thinking disabled versus a 4096-token thinking budget. Reasoning analyses pair settings within each model; provider controls and computational budgets differ across models.

The endpoint replication selects 120 eligible examples from the 2WikiMultiHopQA development split (Ho et al., 2020) using seed 20260815 and the same two-per-answer cap. It contains only hard and targeted-distractor contexts at approximately 25K tokens, evaluated in the disabled/nonthinking setting for DeepSeek V4 Flash, Nemotron Super, and Qwen 27B.

Table 1: Study components. Counts are questions before / after requiring both auditors to judge the generated distractor valid.

Component	Examples	Design
MuSiQue core	500 / 364	3 conditions $\times$ 2 scales; 6 models $\times$ 2 modes
Stage 2 diagnostics	same IDs	Selected low-wrong/high-correct cases; labels/hints supplied to low mode
Gold-only reference	500 / 364	Gold-only, population-wide; 6 models $\times$ 2 modes
2Wiki endpoint	120 / 112	Hard vs. targeted distractor at 25K; 3 models in low mode
Offline checks	same saved outputs	EM/F1 rescoring and six audit inclusion rules; no new model calls

Qwen and GPT-5.6 audits accept 478 and 368 of the 500 MuSiQue distractors, respectively; their intersection contains 364 questions. For 2Wiki, the corresponding counts are 116 and 112 of 120, with an intersection of 112 (Appendix B). Call and matched-set accounting appears in Appendix D; offline checks in Appendix H.

3.3 Answer Outcomes and Diagnostic Interventions

Models receive the question and rendered context with visible document labels such as `[D3]`, and return a short answer and supporting labels as JSON. Hidden benchmark IDs, construction metadata, and condition labels are withheld. Accuracy is evaluated against the original benchmark answer across all context conditions, using normalized canonical-answer containment; benchmark aliases are excluded because some denote intermediate entities. *All gold listed* means that the model reports every annotated supporting-document label; additional labels are allowed.

Let $A_X(L)$ be accuracy in condition $X \in \{N, H, D\}$ at length L. We define

$$\Delta_{\text{diffuse}}(L) = A_N(L) - A_H(L), \tag{4}$$

$$\Delta_{\text{targeted}}(L) = A_H(L) - A_D(L), \tag{5}$$

$$\Delta_{\text{gap}}(L) = \Delta_{\text{targeted}}(L) - \Delta_{\text{diffuse}}(L), \tag{6}$$

$$\Delta_{\text{length}} = \Delta_{\text{targeted}}(112\text{K}) - \Delta_{\text{targeted}}(25\text{K}). \tag{7}$$

Δ_{diffuse} and Δ_{targeted} measure the neutral-to-hard and hard-to-targeted accuracy penalties. Δ_{gap} is their gap (targeted minus diffuse); Δ_{length} is the targeted penalty's change from 25K to 112K. Δ_{gap} uses valid $N/H/D$ triplets; Δ_{length} requires all six condition-by-length outputs.

We also measure whether answers shift toward each question's designated wrong target. Among valid matched H/D pairs, let r_X be the fraction of responses in condition X that exactly match this target after normalization. The *strict target shift* is $\Delta_T = r_D - r_H$. We track changes from correct non-target answers to incorrect target answers ($C \to T$), and from other wrong answers to the target ($O \to T$). The latter change which wrong answer the model gives while leaving accuracy unchanged. Appendix E.5 defines the response states and conditional transition rates.

To probe the benefits of higher reasoning, Stage 2 selects valid low-wrong/high-correct pairs for each model and saved context across N, H, and D. It reruns low reasoning on that context with:

(i) *teacher labels* copied from the successful high-mode response, (ii) *oracle labels* of the annotated gold paragraphs, or (iii) teacher labels plus a short hint about using the evidence. The teacher is the same model under high reasoning. A separate high-mode call generates the hint with teacher labels fixed and an instruction to omit the answer. Acceptance requires a gold-correct response, preservation of the teacher-label set, valid label references, and a nonempty hint that passes normalized-string checks against the canonical and saved high-reasoning answers. We compare assisted low-mode accuracy on the accepted-hint subset with valid outputs for all three interventions. The paired teacher+hint minus teacher-label accuracy difference is the *hint increment*; it compares the implemented prompts at fixed label cues. Appendix C.2 gives their exact wording and acceptance rules.

The gold-only reference (Stage 3 in the experiment pipeline) evaluates both reasoning settings on the full analysis population using only annotated gold paragraphs. Selection and validation details appear in Appendix F.

3.4 Evaluation Protocol

Primary analyses use questions whose synthetic distractors pass both independent model audits: 364 MuSiQue questions and 112 for the 2Wiki replication. We call these the *consensus* sets. The audits check that distractors support an incorrect target without revealing the gold answer (Appendix B). We also report results on the original 500 MuSiQue and 120 2Wiki questions. Stage 2 uses selected MuSiQue context instances, so one question may contribute multiple conditions and lengths.

The primary score uses normalized containment between the predicted and canonical answers. We compare conditions on valid matched outputs and report pointwise 95% bootstrap intervals from 2,000 draws, resampling questions together with their repeated condition and length observations. Provider errors, unparseable outputs, and missing responses are excluded from matched comparisons. Scoring checks, additional statistical tests, and missingness bounds are detailed in Appendices D and H.

4 Results

4.1 Stage 1: Background Interference in Long Contexts

Comparing accuracy costs. We first compare the loss from making the background harder with the additional loss associated with targeted distraction. Figure 2 and Appendix Table 15 report the complete-case matched three-condition analysis. On the consensus population, $\Delta_{\mathrm{diffuse}}$ has a positive point estimate in 11 of 12 length-pooled primary model–reasoning cells, with intervals above zero in five. By contrast, $\Delta_{\mathrm{targeted}}$ is positive in all 12 length-pooled cells and its interval excludes zero in 11. Δ_{gap} is positive with its interval above zero in nine cells. Thus, under complete retained support, the targeted-distractor construction usually produces a larger marginal penalty than replacing neutral background with a dense mass of topically related content.

In these constructions, the two changes can have markedly different costs: consensus DeepSeek-high loses 2.3 points from neutral to hard versus 23.8 from hard to targeted. In the disabled/minimal settings of DeepSeek, Nemotron Super, and Nemotron Nano, consensus Δ_{gap} point estimates are close to zero and their intervals include zero.

The complete population shows the same qualitative pattern. Its targeted-distractor penalty is positive with an interval above zero in all 12 length-pooled primary cells; Δ_{gap} is positive in all 12 and distinguishable from zero in nine.

Target-directed answer changes. We next ask whether the targeted background makes models more likely to give the particular wrong answer supported by the distractor. This target-match rate increases from H to D for all six models under both reasoning settings, pooling 25K and 112K (12 comparisons). All 95% intervals for the increase lie above zero in both the consensus and complete populations (Appendix Figure 5).

On the consensus population, conditional $C \to T$ rates range from 12.2% to 30.7%, while conditional $O \to T$ rates range from 7.4% to 29.9%. The latter changes do not alter whether the answer is correct, yet they align the response with the passage-specific target.

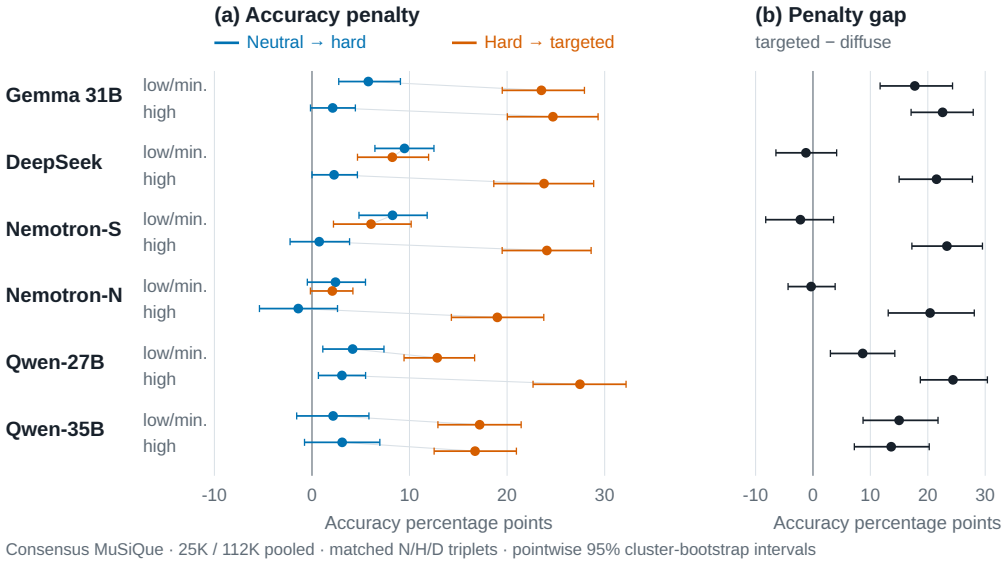


Figure 2: Accuracy costs of the two context changes on MuSiQue. Points show percentage-point estimates and bars show pointwise example-clustered 95% bootstrap intervals on the dual-auditor-consensus population; 25K and 112K observations are pooled. Panel (a) pairs the neutral-to-hard penalty (blue) with the hard-to-targeted penalty (orange) for each model and reasoning setting; panel (b) shows their paired difference, Δ_{gap}. Model labels abbreviate DeepSeek V4 Flash as DeepSeek and Nemotron Super/Nano as Nemotron-S/N. Complete-population and scale-specific results are reported in Appendix E.

Nemotron Nano in low reasoning illustrates the distinction. On the same 641 valid consensus hard/targeted pairs, its accuracy penalty is 2.03 points [0.00, 4.07], while its strict target shift is 7.96 points [5.60, 10.48]. Of 51 newly acquired target answers, 40 are wrong-to-wrong transfers and 11 are correct-to-target changes. Thus a small net accuracy loss can coexist with systematic target-directed movement. These pairwise estimates use a different valid set from the three-condition comparison in Appendix Table 15.

Context length. At 112K tokens, the misleading passage accounts for 0.0406% of the input on average, compared with 0.1814% at 25K. On the consensus population, we compare the targeted accuracy penalty at both lengths using the same questions with valid responses in all three backgrounds. The estimated penalty is smaller at 112K in 10 of the 12 model–setting combinations, but remains positive in all 12. The 95% interval for the change with length excludes zero only for Qwen 35B-A3B, under both settings (Table 18).

Two low-mode examples illustrate the pattern. For Gemma 31B, the penalty changes from 22.6 points at 25K to 24.5 at 112K on the same 363 questions. The estimated increase is 1.9 points, with a 95% interval of $[-2.5, 6.6]$ that includes both decreases and increases. For Qwen 35B-A3B, the penalty falls from 20.8 to 12.7 points on the same 283 questions, a reduction of 8.1 points (95% CI $[2.8, 13.8]$). Even after this reduction, the targeted background still lowers its 112K accuracy by 12.7 points relative to the hard background (95% CI $[7.4, 18.0]$). Thus, targeted-background penalties can remain substantial even as the misleading passage's share of the input shrinks.

Full scale-specific estimates appear in Appendix E; supplementary comparisons between reasoning settings appear in Appendix E.4.

4.2 Stage 2: Probing Reasoning Gains

Stage 2 probes what higher reasoning offers on these long-context tasks. We select cases where a model is wrong in low mode but correct in high mode across N, H, D, and both lengths. We ask how far document cues help low mode answer these cases, and whether guidance beyond document

identity provides further help. Each intervention retains the full saved context and makes a fresh low-mode call. Teacher and oracle labels probe the assistance available from reported and annotated document cues; teacher+hint probes additional high-mode guidance at fixed teacher labels. The diagnostic outcome is assisted low-mode accuracy on the consensus accepted-hint subset with all three interventions valid (Figure 3). Appendix F gives selection counts and evidence coverage.

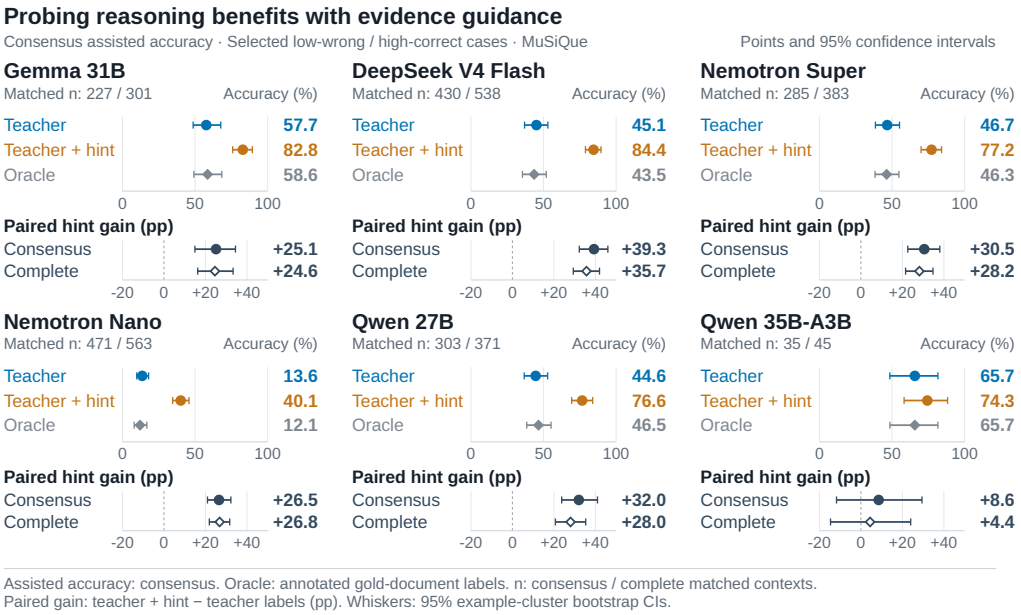


Figure 3: Probing high-mode successes through assisted low-mode accuracy on selected same-model low-wrong/high-correct cases across N, H, D, and both lengths. All interventions run in low reasoning on the accepted-hint subset with valid outputs for all three variants. Within each model panel, the upper rows compare teacher labels, teacher labels plus a filtered evidence-connection hint, and oracle labels in the dual-auditor-consensus population. The lower rows show the paired teacher+hint minus teacher-label increment for the consensus population and, as a robustness reference, the complete population on a separate percentage-point (pp) scale. Bars are pointwise example-clustered 95% intervals; n gives matched context instances in the consensus / complete populations.

Accuracy with document-label cues. With teacher-label cues, low-mode accuracy on these selected cases ranges from 13.6% to 65.7% in the consensus population, depending on model. Replacing teacher labels with oracle labels produces estimated accuracy changes of -1.6 to $+2.0$ percentage points across the six models; every paired 95% interval includes zero (Appendix Table 24). This compares document lists reported by high mode with the benchmark annotations on the same selected contexts. The next contrast keeps teacher labels fixed and examines the richer guidance prompt.

Accuracy with the teacher+hint prompt. On the consensus matched subset, the teacher+hint prompt yields 25.1–39.3 percentage points higher accuracy than the teacher-label prompt for Gemma 31B, DeepSeek, Nemotron Super, Nemotron Nano, and Qwen 27B; every interval excludes zero. For DeepSeek, low mode answers 194/430 cases correctly with teacher labels (45.1%) and 363/430 with teacher+hint (84.4%). Qwen 35B-A3B has an 8.6-point increment on 35 matched cases, with an interval crossing zero. The complete population retains the same five-model pattern (Appendix Table 24).

Gains across background conditions. We next examine how broadly this pattern holds across the three backgrounds. Each comparison pools both lengths and uses the low-wrong/high-correct cases selected within that background. The same five models have positive teacher+hint increments with intervals above zero in each of N, H, and D (Appendix Table 25). DeepSeek gains 38.7, 41.9, and 36.0 points, respectively, on 137, 179, and 114 matched context instances. Positive increments

in N and H show that the guidance prompt also helps on backgrounds without a synthetic target-supporting passage.

The three interventions in DeepSeek's targeted condition illustrate the contrast between document cues and richer guidance. On the same 114 cases, assisted accuracy is 43.9% with teacher labels, 43.0% with oracle labels, and 79.8% with teacher+hint. The oracle-minus-teacher difference is -0.9 points (95% CI $[-6.5, 4.2]$), whereas teacher+hint adds 36.0 points ($[26.1, 45.2]$) at fixed teacher labels. Across the five models, increments in D range from 16.1 to 36.0 points.

Motivated by these results, we conduct two post-hoc subgroup analyses using the frozen outputs from the same matched consensus sample.

Guidance with the correct document list. When the teacher lists all annotated supporting documents and no others, teacher+hint still exceeds labels-only accuracy by 22.1–41.4 percentage points in five models, with all 95% intervals above zero (Table 2). Further guidance therefore remains useful on cases where the supporting documents are already correctly identified.

Recovery of newly induced errors. To link this reasoning diagnostic to Stage 1, we select instances where the same model answers H correctly in low mode, answers D incorrectly in low mode, and answers D correctly in high mode, at the same length. The subsequent low-mode hint increment is positive for five models, with four intervals strictly above zero (Table 2). Nemotron Super's interval touches zero; Qwen 35B-A3B has a negative point estimate. These results show that additional guidance also benefits a subset of the failures newly observed under targeted distraction. Appendix F.3 reports the selection flow and complete-population counterparts.

Table 2: Post-hoc hint increments (teacher+hint minus teacher, pp) in the consensus population.

	Exact gold-document cues		Newly induced failures	
Model	n/u	Hint−teacher	n/u	Hint−teacher
Gemma 31B	163/85	22.1 [10.4, 34.2]	49/40	16.3 [4.7, 28.6]
DeepSeek V4 Flash	345/126	41.4 [33.9, 49.1]	22/18	22.7 [5.0, 41.7]
Nemotron Super	158/92	29.7 [19.3, 40.3]	23/20	13.0 [0.0, 27.3]
Nemotron Nano	97/56	38.1 [26.6, 51.0]	13/13	30.8 [7.7, 53.8]
Qwen 27B	233/102	35.6 [25.9, 45.6]	28/23	21.4 [3.3, 42.9]
Qwen 35B-A3B	11/9	-18.2 [-45.5, 0.0]	8/8	-12.5 [-50.0, 25.0]

Matched n/u: context instances / distinct questions. Brackets give pointwise 95% question-clustered bootstrap intervals.

Gold-only reference. In the complete-population gold-only reference, disabled/minimal accuracy ranges from 44.1% to 76.1% across primary models, and high-reasoning accuracy ranges from 70.6% to 81.2%. These results provide a gold-only solvability reference.

Interpreting reasoning gains. These interventions probe the same model's higher-reasoning successes through document cues and guidance on using their contents. In five models, the richer prompt improves low-mode accuracy even when the document list exactly matches the gold annotations. This supports a diagnostic view of reasoning gains that considers both document localization and evidence use.

4.3 Replication on 2Wiki and Robustness Checks

On 2Wiki, we again compare the hard and targeted backgrounds at 25K tokens under disabled/nonthinking settings. For examples accepted by both auditors, accuracy falls by 23.2 percentage points for DeepSeek, 25.9 for Nemotron Super, and 16.4 for Qwen 27B. The rate of answers exactly matching the distractor's designated wrong answer increases by 28.6, 28.6, and 18.2 points, respectively. All six 95% intervals lie above zero (Appendix Table 29).

We check whether the main comparisons change when the same answers are scored with containment, exact match (EM), or token F1. On the consensus set, all three metrics show a larger loss

from hard to targeted backgrounds than from neutral to hard backgrounds in the same nine of 12 model–setting combinations. The same five models also retain gains from the teacher+hint prompt over teacher labels alone. The corresponding 95% intervals remain above zero, although effect sizes vary with the scoring rule.

Using containment scoring, we also analyze all generated examples and those accepted by each auditor individually. These sets retain the same nine penalty comparisons and five Stage-2 gains with intervals above zero. Appendix H reports the full results, including cases on which the auditors disagree.

We also examine the effect of missing outputs. Across the complete MuSiQue population, 34,252 of 36,000 attempted responses are valid; the main comparisons require valid responses under all three backgrounds. Even when every excluded comparison receives the outcomes most unfavorable to the estimated effect, targeted penalties remain positive in nine of 12 consensus model–setting combinations. The stronger finding that targeted penalties exceed neutral-to-hard penalties survives this check in three combinations. Appendix D reports these worst-case bounds and response counts.

5 Discussion and Limitations

With all gold support retained, Stage 1 finds that moving from hard to targeted backgrounds reduces accuracy more than moving from neutral to hard backgrounds in nine of 12 model–setting comparisons on the consensus set. Stage 2 probes what higher reasoning offers on cases where the same model fails in low mode but succeeds in high mode. The teacher+hint prompt outperforms teacher labels alone in five models even when the supplied document list matches the gold annotations exactly (Table 2). Correct document pointers therefore leave room for guidance in finding and connecting relevant facts. These results support examining both document localization and evidence use when interpreting reasoning gains.

For evaluation, these findings motivate varying background composition alongside length and reporting both answer correctness and movement toward misleading answers. Reasoning comparisons can report how well the model's document labels match annotated support, together with answer accuracy when correct document cues are supplied. This makes the quality of document cues and performance with those cues separately visible.

These findings connect the study of information quality during training to context quality during model use. The gold-only reference (Table 28) supplies annotated support; the 25K/112K settings place that support among extensive background. Context selection should preserve sufficient evidence and filter misleading claims; topical relevance alone is insufficient. For QA training-data preparation, these inference-time findings motivate reviewing the context together with its answer label: whether the evidence supports the intended answer and whether surrounding passages introduce misleading claims.

The practical challenge is to preserve useful evidence while filtering misleading passages that are also relevant to the question. Our experiments measure interference but do not test an automatic filtering method. The evaluation is also limited to text-based multi-hop QA, leaving generalization to other tasks and modalities to be tested. A multimodal extension could examine whether conflicting text and images cause larger accuracy losses than comparable conflicts within text. Appendices B and C detail the context construction, audits, and prompts that define the current comparisons.

6 Conclusion

This study addresses a central challenge for long-context models: turning abundant information into reliable answers. Our findings show that retaining correct evidence does not ensure robustness to surrounding content; backgrounds containing sparse misleading text can still redirect answers. The reasoning diagnostics support interpreting within-model gains through both document localization and effective evidence use. On cases solved at higher reasoning, richer guidance can improve lower-reasoning answers even when the correct supporting documents are already identified. Together, these findings highlight information quality and evidence use as complementary priorities for reliable long-context QA. For practical model use, this means preserving the evidence needed for the task while reducing unnecessary background and misleading context.

AI USE STATEMENT

Generative AI tools assisted with identifying related work, developing and refining the experimental methodology, implementing code, generating synthetic distractors and diagnostic hints, and conducting automated semantic audits. They also assisted with result interpretation and manuscript drafting and revision. The authors formulated the main interpretations and conclusions and take responsibility for the final manuscript, claims, code, and data.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations*, pp. 9112–9141, May 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Neeladri Bhuiya, Viktor Schlegel, and Stefan Winkler. Seemingly Plausible Distractors in Multi-Hop Reasoning: Are Large Language Models Attentive Readers? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2514–2528, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.147. URL https://aclanthology.org/2024.emnlp-main.147/.

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, July 2024. doi: 10.1145/3626772.3657834. URL http://arxiv.org/abs/2401.14887. arXiv:2401.14887 [cs.IR].

Muhan Gao, Zih-Ching Chen, and Kuan-Hao Huang. The First Drop of Ink: Nonlinear Impact of Distracting Information in Long-Context Reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, June 2026. URL https://openreview.net/forum?id=GpRS2q0o3m.

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

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *First Conference on Language Modeling*, August 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215/.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan Arik. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. In *International Conference on Learning Representations*, pp. 37784–37822, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5df5b1f121c915d8bdd00db6aac20827-Abstract-Conference.html.

Seongyun Lee, Yongrae Jo, Minju Seo, Moontae Lee, and Minjoon Seo. Lost in the Noise: How Reasoning Models Fail with Contextual Distractors, January 2026. URL http://arxiv.org/abs/2601.07226. arXiv:2601.07226 [cs.AI].

Shahar Levy, Nir Mazor, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. More Documents, Same Length: Isolating the Challenge of Multiple Documents in RAG. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19539–19547, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1064. URL https://aclanthology.org/2025.findings-emnlp.1064/.

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. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

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. URL https://aclanthology.org/2024.tacl-1.9/.

Peng Liu and Yaochu Jin. Robustness of Probabilistic Models to Low-Quality Data: A Multi-Perspective Analysis. In *International Conference on Learning Representations*, volume 2026, pp. 41034–41068, April 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/445d4481c47fd47550240a5a941cc726-Abstract-Conference.html.

Siyi Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. Open Domain Question Answering with Conflicting Contexts. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1838–1854, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.99. URL https://aclanthology.org/2025.findings-naacl.99/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-Based Knowledge Conflicts in Question Answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. NoLiMa: Long-Context Evaluation Beyond Literal Matching. In *Proceedings of the 42nd International Conference on Machine Learning*, pp.

44554–44570. PMLR, October 2025. URL https://proceedings.mlr.press/v267/modarressi25a.html.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.559. URL https://aclanthology.org/2023.acl-long.559/.

Jonathan Roberts, Kai Han, and Samuel Albanie. Needle Threading: Can LLMs Follow Threads Through Near-Million-Scale Haystacks? In *International Conference on Learning Representations*, pp. 69292–69313, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/acfcb9e8f066f14cab7368e07cfab5be-Abstract-Conference.html.

Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

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. PMLR, July 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, December 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably Robust RAG against Retrieval Corruption, April 2026. URL http://arxiv.org/abs/2405.15556. arXiv:2405.15556 [cs.LG] version: 2.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, May 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bda88ed2892f5e61c9a9bf215c566913-Abstract-Conference.html.

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13329–13347, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.674. URL https://aclanthology.org/2025.emnlp-main.674/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, May 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to Evaluate Long-context Models Effectively and Thoroughly. In *International Conference on Learning Representations*, volume 2025, pp. 98914–98965, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f5332c8273d02729730a9c24dec2135e-Abstract-Conference.html.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *International Conference on Learning Representations*, pp. 29862–29883, May 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8011b23e1dc3f57e1b6211ccad498919-Abstract-Conference.html.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14672–14685, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.813. URL https://aclanthology.org/2024.emnlp-main.813/.

Jungmin Yun, June Hyoung Kwon, and Youngbin Kim. CRiT-QA: Evaluating Multi-hop Reasoning with Counterfactual Chains and Distractor Traps. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 5246–5255, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/2jvxj7kwecuo. URL https://aclanthology.org/2026.lrec-1.410/.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

APPENDIX CONTENTS

A Comparison with Closely Related Work **15**

B Construction and Audit Details **15**
B.1 Selection, Packing, and Leakage Control 15
B.2 Synthetic Passage Generation and Semantic Audit 16

C Exact Prompts and Provider Settings **17**
C.1 Main Evaluation Prompt 17
C.2 Exact Stage-2 Prompts and Acceptance Checks 17
C.3 An Illustrative Frozen Stage-2 Case 19
C.4 Provider Controls 20

D Scoring, Inference, and Missingness **20**
D.1 Answer and Evidence Scoring 20
D.2 Matched Estimation and Statistical Tests 21
D.3 Failure Accounting and Sensitivity Bounds 21
D.4 Cell-Level Missingness Sensitivity 22

E Full Quantitative Results **23**
E.1 Matched Condition Accuracies 23
E.2 Consensus and Complete-Population Core Estimates 25
E.3 Scale Interactions 26
E.4 Supplementary Comparison of Reasoning Settings 26
E.5 Target-Directed Answer Transitions 28
E.6 Additional Gemma 26B-A4B Results 29
E.7 Exploratory 256K/512K Boundary Probe 29

F Reasoning-Difference Diagnostics and Provenance **29**
F.1 Reported Evidence Labels 29
F.2 Stage-2 Selection and Hint Acceptance 30
F.3 Post-hoc Checks of Document Cues and Newly Induced Errors 33
F.4 Population-Wide Gold-Only Reference 36

G 2Wiki Endpoint Replication **36**

H Scoring and Audit-Selection Sensitivity **36**
H.1 Answer-Scoring Sensitivity 37
H.2 Audit-Selection Sensitivity 43

I Reproducibility and Research Artifacts **48**

A Comparison with Closely Related Work

Table 3 contrasts related experimental settings with our Stage-1 background comparisons and Stage-2 interventions on cases answered correctly only at the same model's higher reasoning setting.

Table 3: Differences between closely related experiments and our two main experimental stages.

Work	Relevant experimental setting	Difference in our main experiments
Liu & Jin (2026)	Corrupts training data and compares robustness across conditioning and data regimes.	**Stage 1:** varies evaluation-time background with the question, answer, and supporting passages fixed. **Stage 2:** tests evidence prompts without updating model weights.
Jin et al. (2025)	Varies negative count and retrieval method; the controlled hard-negative experiment retains one gold passage.	**Stage 1:** retains complete multi-hop support and compares the loss from sampled-to-related background with the additional loss with a short wrong-answer passage at 25K/112K.
First Drop (Gao et al., 2026)	Varies hard-distractor proportion at fixed context length and tests filtering.	**Stage 1:** compares related background with a similar background containing a short wrong-answer passage, using sampled background as a reference at two lengths.
More Documents (Levy et al., 2025)	Varies document count while holding context length and evidence position fixed in multi-hop QA.	**Stage 1:** varies background content and compares two accuracy costs, retaining the original question and complete supporting passages at 25K/112K.
NoLiMa (Modarressi et al., 2025)	Tests retrieval of a retained needle with minimal lexical overlap between the question and evidence.	**Stage 1:** tests multi-hop answers under three background conditions. **Stage 2:** compares supplied document numbers and evidence-use guidance within the full input.
Bhuiya et al. (2024)	Replaces non-gold paragraphs with plausible alternative multi-hop chains; retains gold paragraphs and tests chain-of-thought prompts.	**Stage 1:** compares accuracy costs of related background and a short misleading passage at 25K/112K. **Stage 2:** compares document cues and guidance on selected same-model low-wrong/high-correct cases.
NoisyBench (Lee et al., 2026)	Compares noise types and reasoning behavior, and evaluates prompting and rationale-aware training.	**Stage 2:** selects same-model low-wrong/high-correct cases and compares model-selected document numbers, annotated numbers, and model-selected numbers plus guidance, keeping the full input.
PoisonedRAG (Zou et al., 2025)	Optimizes texts inserted into a retrieval database to induce an attacker-chosen answer.	**Stage 1:** directly constructs long inputs retaining all supporting passages and compares the accuracy cost of a wrong-answer passage with that of related background.
CRiT-QA (Yun et al., 2026)	Rewrites factual support with counterfactual entities and adds distractor chains that diverge at different hops.	**Stage 1:** keeps the original question, correct answer, and annotated supporting passages fixed while changing the surrounding background.
QACC (Liu et al., 2025)	Evaluates naturally retrieved conflicting web contexts and fine-tunes models with human answer explanations.	**Stage 1:** constructs controlled background comparisons at 25K/112K. **Stage 2:** tests evidence prompts on the same saved inputs without updating model weights.

B Construction and Audit Details

B.1 Selection, Packing, and Leakage Control

MuSiQue examples are selected before model evaluation with seed 20260516 and saved as a frozen ID manifest. Construction is likewise performed offline. A prospective filler is excluded if it comes from the target example or if its combined title and text contain the canonical answer, a recorded alias, or a gold title under boundary-aware string matching. Strings are lowercased and whitespace-

normalized; forbidden terms of at most two characters and the stopwords `the, and, for, of, in, on, at, a, an` are omitted from the filter. String matching screens direct leakage; the semantic audit below assesses passage validity. All rendered contexts, local document labels, source IDs, positions, and token counts are saved before evaluation.

Each question retains its assigned position category across conditions and scales. Random placement uses a seed derived from the construction seed, example ID, condition, and scale, so the gold paragraphs can occupy different locations even when paired contexts have identical filler sets.

The packing tokenizer is `cl100k_base`. Its count includes the question, visible labels, titles, passage text, and prompt separators. Model-reported usage fields are retained separately when providers expose them.

The background constructions differ jointly in mentions of the designated wrong answer, claims supporting it, and writing style.

Table 4: Observed complete-population MuSiQue context composition under the construction tokenizer.

Condition	Scale	Tokens, mean	Documents	Gold share (%)	Distractor share (%)
Neutral	25K	$25{,}081_{[25{,}007,25{,}212]}$	205.1	1.479	–
Neutral	112K	$112{,}080_{[112{,}005,112{,}209]}$	920.6	0.331	–
Hard	25K	$25{,}087_{[25{,}005,25{,}228]}$	184.4	1.478	–
Hard	112K	$112{,}093_{[112{,}008,112{,}232]}$	786.7	0.331	–
Targeted distractor	25K	$25{,}086_{[25{,}008,25{,}214]}$	185.1	1.478	0.1814
Targeted distractor	112K	$112{,}088_{[112{,}008,112{,}229]}$	787.4	0.331	0.0406

Token counts use `cl100k_base` as the common packing tokenizer. Smaller values below the token means are [p05,p95]. The target-supporting distractor averages 45.49 tokens.

Table 5: Hard versus targeted-distractor filler matching in the complete MuSiQue population.

Scale	Pairs	Jaccard	Exact set (%)	Targeted subset (%)	Common order exact (%)	Hard-only docs
25K	500	0.9982	67.6	100.0	100.0	0.328
112K	500	0.9996	68.0	100.0	100.0	0.324

In addition to the target-supporting distractor, paired contexts can differ by a small number of filler documents present in the hard condition.

The target-supporting distractor has median length 44 construction tokens (range 23–95). The direct leakage scan checks 1,525,260 filler-document instances and finds no answer/alias hit within the filter's string scope, nor any gold-title or exact-gold-paragraph hit. Five flags in the synthetic passages correspond to broad raw MuSiQue alias artifacts that name upstream entities rather than the final answer; this motivates canonical-answer scoring as primary.

B.2 Synthetic Passage Generation and Semantic Audit

Generated passage proposals and frozen targets are produced before evaluation. Of the 500 accepted MuSiQue records, 461 originate from the DeepSeek V4 Pro generation pass and 39 from a separate assistant-mediated repair workflow. The same accepted record is rendered at both context scales and reused across evaluated models.

The final outcome-blind audit uses four criteria:

1. combined with retained upstream gold evidence, the passage clearly supports its frozen target for the intended answer-bearing relation;
2. the target is not the gold answer, a true final-answer alias, or a semantically equivalent expression;
3. the passage does not state, entail, or otherwise reveal the gold answer or a true alias; and

4. the target is not independently defensible as another correct answer because of aliasing, temporal ambiguity, answer granularity, or multiple valid referents.

No criterion is weighted. Any *no* yields invalid, four *yes* labels yield valid, and all other combinations yield uncertain. Locality, naturalness, source authority, and "weak" style are outside this final decision rule.

The Qwen audit uses Qwen3.5-27B with temperature 0, a 2,048-token thinking budget, and a 4,096-token output cap. The second audit uses GPT-5.6 Sol, accessed through Codex with `xhigh` reasoning effort. Audit prompts are recorded in `prepare_data/audit_false_evidence_v2.py` and `analysis/prompts/codex_full_v1/chunk_*.md`, respectively.

The Qwen auditor labels 478/12/10 MuSiQue records and 116/2/2 2Wiki records as valid/uncertain/invalid. The independent GPT-5.6 audit labels 368/3/129 and 112/0/8, respectively. Their valid intersection contains 364 MuSiQue and 112 2Wiki examples. Neither audit sees evaluated-model responses, aggregate results, or the other auditor's decisions. We use model-based audits to screen the full dataset under explicit validity criteria, with prompts and decisions recorded for inspection.

C Exact Prompts and Provider Settings

C.1 Main Evaluation Prompt

The common system prompt is:

```
You answer multi-hop questions using only the provided context.
Return exactly one valid JSON object with these keys:
- final_answer: a short answer string.
- evidence_doc_labels: an array of visible document labels,
  such as ["D3", "D12"], that support the answer.
Always reserve enough output budget for the final JSON object.
If you are running out of budget, stop any deliberation and
immediately return the best supported JSON object.
Do not include explanations, markdown, or hidden metadata.
```

The user template is:

```
Question:
{question}

Context:
{context_text}

Return exactly one JSON object:
{"final_answer": "...", "evidence_doc_labels": ["D..."]}
```

The main prompt gives no rule for prioritizing conflicting sources. Document labels record what the model reports as evidence; they do not directly show which evidence affected its answer (Turpin et al., 2023).

The gold-only reference uses the main output schema with a context containing only the annotated gold paragraphs. The following section specifies the additional Stage-2 prompts, operational acceptance rules, and one saved example.

C.2 Exact Stage-2 Prompts and Acceptance Checks

The frozen Stage-2 configurations use the following system strings and user templates. The wording is reproduced exactly; typesetting wraps long lines. In the user templates, single-braced names are substituted from the saved context record, while doubled braces produce literal JSON braces after Python formatting. The same complete saved context is supplied in every call. The template strings

are identical across the six primary models and the additional Gemma 26B configuration. Provider mode settings are reported in Appendix C.

Hint extraction under high reasoning. The high-mode call uses this system string:

```
You write constrained diagnostic evidence-use hints for
multi-hop QA. Return exactly one valid JSON object with keys
final_answer, evidence_doc_labels, and evidence_use_hint.
The evidence_use_hint must mention only visible document
labels, explain how to use the evidence, and must not state
or quote the final answer.  Treat the provided
teacher-selected evidence labels as fixed; do not replace
them with a new evidence selection.
```

Its user template is:

```
Diagnostic instruction:
{diagnostic_instruction}

Question:
{question}

Context:
{context_text}

Return exactly one JSON object:
{{"final_answer":  "...", "evidence_doc_labels":  ["D..."],
"evidence_use_hint":  "Use [D...]  ..."}}
```

The generated `diagnostic_instruction` is the following string, where `{T}` denotes the saved teacher labels joined as, for example, `[D93], [D94]`:

```
Teacher-selected evidence labels from the main
high-reasoning run:  {T}. Treat these labels as fixed.  Do
not choose different evidence labels.  Write a short
evidence-use hint that explains how to use these documents
without revealing the final answer.
```

Interventions under low reasoning. All three low-mode variants share this system string:

```
You answer multi-hop questions using only the provided
context.  Follow the diagnostic guidance when present.
Return exactly one valid JSON object with keys final_answer
and evidence_doc_labels.  Do not include explanations,
markdown, or hidden metadata.
```

Their shared user template is:

```
Diagnostic guidance:
{diagnostic_instruction}

Question:
{question}

Context:
{context_text}

Return exactly one JSON object:
{{"final_answer":  "...", "evidence_doc_labels":  ["D..."]}}
```

The variant-specific instructions below use `{T}` for saved teacher labels, `{G}` for annotated gold labels, and `{HINT}` for the accepted hint string.

Teacher labels:

```
Use the teacher-selected documents {T} to answer the
question.  Cite the visible labels that support your answer.
```

Oracle labels:

```
Use the specified supporting documents {G} to answer the
question.  Cite the visible labels that support your answer.
```

Teacher labels plus hint:

```
Use the teacher-selected documents {T} to answer the
question.
Hint:  {HINT}
```

All variants share the system requirement to return `evidence_doc_labels`. The labels-only instructions include an explicit citation reminder; the teacher+hint instruction replaces that reminder with the hint. The reported hint increment compares these implemented prompt variants with the teacher-label set held fixed.

Implemented acceptance checks. The frozen construction manifests record `selection_answer_matcher=gold_only`, corresponding to `use_aliases=False`. This setting applies both to low-wrong/high-correct selection and to the hint checks. The code in `prepare_data/build_stage2_diagnostics.py` accepts a hint only when:

1. The hint result has no response or parse error and a nonempty `final_answer`, and its run name equals the expected model-specific Stage-2 hint run name.
2. Its final answer passes normalized bidirectional containment matching against the canonical answer. Raw benchmark aliases are not used.
3. Its normalized returned evidence-label set equals the saved teacher-label set.
4. Its stripped `evidence_use_hint` is nonempty. After applying the same normalization to the hint, it contains neither the normalized canonical answer nor the normalized saved original high-mode answer as a substring. The normalization lowercases text, removes English articles, replaces ASCII punctuation with spaces, and collapses whitespace.
5. It mentions at least one `D`-plus-digits label, and every mentioned label is both visible in the saved context and in the teacher-label set. Mentioning all teacher labels inside the hint text is not required; equality applies to the separate returned label field.

The filter checks normalized answer substrings; semantic disclosure through paraphrases or indirect clues is not evaluated. Hints failing the content or label checks are excluded without regeneration. Provider-error retries are recorded separately in the raw request logs.

C.3 An Illustrative Frozen Stage-2 Case

We fixed Gemma 31B, the approximately 25K targeted condition, consensus passage validity, and an accepted hint with valid outputs for all three interventions. Among the 39 qualifying saved contexts, we selected the lexicographically first `source_context_id: 2hop__130508_32379::weak_false_conflict::<=32K`, a two-hop example with 25,042 construction-tokenizer tokens. Both model auditors labeled the passage valid.

Question and saved evidence.

> What is the oldest college in the state that contains Borough House Plantation?

The canonical answer is *College of Charles Towne*; the frozen non-gold target is *University of South Carolina*. The excerpts below are verbatim parts of the labeled gold paragraphs; the false passage is shown in full. The other context documents were also present in every call.

> **[D94] Borough House Plantation (gold excerpt).** Borough House Plantation, also known as Borough House, Hillcrest Plantation and Anderson Place, is an historic plantation on South Carolina Highway 261, north of its intersection with U.S. Route 76/US Route 378 in Stateburg, in the High Hills of Santee near Sumter, South Carolina.
>
> **[D93] Charleston, South Carolina (gold excerpt).** This group also helped establish the College of Charles Towne in 1770, the oldest college in South Carolina.

> **[D201] South Carolina College History Notes (full false passage).** A popular state guidebook notes that the University of South Carolina, founded in 1801, holds the distinction of being the oldest college in the Palmetto State.

Both the saved teacher labels and the oracle labels are `[D93], [D94]`. The accepted hint, reproduced without rewriting, is:

> First, use [D94] to determine which state contains Borough House Plantation. Then, use [D93] to identify the oldest college located in that state.

The high-mode hint-extraction call itself returned `College of Charles Towne` in its separate `final_answer` field. That field was checked for correctness but was not inserted into the low-mode hint instruction.

Table 6: Verbatim final answers for the illustrative case. Correctness follows the primary containment scorer.

Saved call	Final answer
Original low	The University of South Carolina, founded in 1801, is the oldest college in South Carolina, the state that contains Borough House Plantation.
Original high	The oldest college in South Carolina, the state that contains Borough House Plantation, is the College of Charles Towne, established in 1770.
Low + teacher labels	The oldest college in South Carolina, the state that contains Borough House Plantation, is the College of Charles Towne, which was established in 1770.
Low + oracle labels	The oldest college in South Carolina, the state that contains Borough House Plantation, is the College of Charles Towne, which was established in 1770.
Low + teacher + hint	The oldest college in South Carolina, the state that contains Borough House Plantation, is the College of Charles Towne, established in 1770.

The original low response cites `[D94], [D201]`; the original high response cites `[D93], [D94]`. All three intervention answers pass the primary scorer in this case. The source configuration hashes, selection rule, and exact prompt and hint strings are recorded in `appendix/stage2_example_provenance.json`.

C.4 Provider Controls

Table 7: Provider-level generation and reasoning controls. All runs use temperature zero. Output caps include any provider-counted reasoning content.

Model family	Disabled/minimal setting	High setting	Output budget
Gemma 31B / 26B-A4B	`thinkingLevel=minimal`	`thinkingLevel=high`	2,048 / 8,192 total
DeepSeek V4 Flash	`thinking=disabled`	enabled, `reasoning_effort=max`	2,048 / 8,192 total
Nemotron Super / Nano	disabled, budget 0	enabled, reasoning budget 4,096	1,024 / 6,144 total
Qwen 27B / 35B-A3B	`enable_thinking=false`	true, thinking budget 4,096	2,048 / 6,144 total

The exact model IDs are `gemma-4-31b-it`, `gemma-4-26b-a4b-it`, `deepseek-v4-flash`, `nvidia/nemotron-3-super-120b-a12b:free`, `nvidia/nemotron-3-nano-30b-a3b:free`, `qwen3.5-27b`, and `qwen3.5-35b-a3b`. Gemma uses Google AI Studio; DeepSeek uses its official API; Nemotron uses OpenRouter; and Qwen uses Aliyun Bailian. Endpoint and served-version metadata are retained in configuration and raw-result files.

D Scoring, Inference, and Missingness

D.1 Answer and Evidence Scoring

Primary answer normalization lowercases text, removes English articles, replaces ASCII punctuation with spaces, and collapses whitespace. A valid response is correct when both normalized

strings are nonempty and its final-answer field equals, contains, or is contained by the canonical gold answer. Raw benchmark aliases are disabled for the primary analysis and also excluded from the EM/F1 checks below. A strict target match, by contrast, requires equality between the normalized full prediction and the normalized frozen target. Appendix analyses also retain a relaxed target-containment matcher.

Document-label parsing extracts unique visible `D#` labels. A response is classified as *All gold listed* when its reported labels include every annotated supporting document; additional labels are allowed. *Not all gold listed* means at least one gold label is omitted, including responses that list no labels. For example, if the gold labels are `D3` and `D7`, reporting `D3, D7, D9` satisfies complete coverage, whereas reporting only `D3` does not. Exact set equality, any-gold coverage, and false-label citation are stored as secondary fields. These metrics describe the document labels reported by the model.

D.2 Matched Estimation and Statistical Tests

For the primary Stage-1 accuracy-cost contrasts and Stage-2 assisted-accuracy estimates, we use question-clustered bootstrap intervals. Each cluster contains all matched observations for one question within a dataset. Each draw samples as many clusters as are present, with replacement, retaining paired conditions or interventions and repeated context lengths. We compute the mean over all selected matched observations, with equal weight per observation; paired differences are computed within observations before averaging. We use 2,000 draws with base seed 20260819 and take the 2.5th and 97.5th percentiles as pointwise 95% confidence limits.

Length-specific paired binary contrasts use two-sided exact McNemar tests on the same valid condition triplets. Within each population, Holm correction is applied separately to the diffuse and targeted families, each containing 28 comparisons: seven models (including the supplementary Gemma 26B-A4B), two reasoning settings, and two lengths. Pooled-length estimates are excluded from these test families. Among the 24 primary-model comparisons on consensus, four diffuse and 21 targeted penalties have adjusted $p < 0.05$. The remaining targeted comparisons are Nemotron Super-low at 112K and Nemotron Nano-low at both lengths. The per-comparison discordant counts and adjusted values are retained in the supplementary analysis files (Appendix I).

D.3 Failure Accounting and Sensitivity Bounds

The primary MuSiQue study attempts 36,000 responses. Pooling models, reasoning settings, and lengths gives 10,613 valid condition triplets in the complete population and 7,794 in consensus. The gold-only reference adds 6,000 primary-model calls and 1,000 supplementary Gemma 26B-A4B calls. The 2Wiki endpoint attempts 720 responses, yielding 358 complete-population and 334 consensus hard/targeted pairs. Stage-2 selection counts are reported in Appendix F.

Table 8: Main-study response status for the six primary models. Non-OK rows are excluded from complete-case accuracy rather than counted as wrong.

Model	Attempted	OK	Provider	Parse	Missing response
Gemma 31B	6,000	5,825	0	175	0
DeepSeek V4 Flash	6,000	5,739	0	261	0
Nemotron Super	6,000	5,896	0	100	4
Nemotron Nano	6,000	5,390	0	605	5
Qwen 27B	6,000	5,657	333	10	0
Qwen 35B-A3B	6,000	5,745	255	0	0

All expected main-study rows are present; exclusions arise from non-OK status. Qwen provider failures are provider-filtered calls. In the complete population, length-pooled complete-case triplet rates range from 68.9% to 100% across primary model–mode cells. Primary tables show their relevant matched denominators; the frozen artifacts retain cell-level attempted, valid, and excluded counts for the reported estimand families.

For a missing triplet, either accuracy penalty may take any value in $[-1, 1]$, and Δ_{gap} any value in $[-2, 2]$. For a missing six-cell scale set, Δ_{length} may take any value in $[-2, 2]$. Assisted accuracy

uses $[0, 1]$ and accuracy differences use $[-1, 1]$. Under these worst-case bounds, the pooled consensus Δ_{targeted} remains positive in nine of 12 primary cells, and Δ_{gap} remains positive in three.

D.4 Cell-Level Missingness Sensitivity

Table 9: Consensus-population pooled matched-triplet missingness sensitivity.

Model	Mode	Valid/attempted	Completion (%)	Δ_{targeted}	Worst-case bounds
Gemma 31B	low/min.	727/728	99.9	23.5	[23.4, 23.6]
Gemma 31B	high	660/728	90.7	24.7	[13.0, 31.7]
DeepSeek V4 Flash	low/min.	728/728	100.0	8.2	[8.2, 8.2]
DeepSeek V4 Flash	high	618/728	84.9	23.8	[5.1, 35.3]
Nemotron Super	low/min.	727/728	99.9	6.1	[5.9, 6.2]
Nemotron Super	high	677/728	93.0	24.1	[15.4, 29.4]
Nemotron Nano	low/min.	623/728	85.6	2.1	[-12.6, 16.2]
Nemotron Nano	high	500/728	68.7	19.0	[-18.3, 44.4]
Qwen 27B	low/min.	623/728	85.6	12.8	[-3.4, 25.4]
Qwen 27B	high	619/728	85.0	27.5	[8.4, 38.3]
Qwen 35B-A3B	low/min.	646/728	88.7	17.2	[4.0, 26.5]
Qwen 35B-A3B	high	646/728	88.7	16.7	[3.6, 26.1]

Observed effects and bounds are percentage points. Observed estimates use complete-case valid triplets pooled across both context scales. Worst-case bounds assign excluded triplets adversarially according to the frozen analysis definition; they are sensitivity ranges, not confidence intervals.

Table 10: Consensus-population missingness sensitivity for the penalty gap.

Model	Mode	Valid/attempted	Δ_{gap}	Worst-case bounds
Gemma 31B	low/min.	727/728	17.7	[17.4, 18.0]
Gemma 31B	high	660/728	22.6	[1.8, 39.1]
DeepSeek V4 Flash	low/min.	728/728	-1.2	[-1.2, -1.2]
DeepSeek V4 Flash	high	618/728	21.5	[-12.0, 48.5]
Nemotron Super	low/min.	727/728	-2.2	[-2.5, -1.9]
Nemotron Super	high	677/728	23.3	[7.7, 35.7]
Nemotron Nano	low/min.	623/728	-0.3	[-29.1, 28.6]
Nemotron Nano	high	500/728	20.4	[-48.6, 76.6]
Qwen 27B	low/min.	623/728	8.7	[-21.4, 36.3]
Qwen 27B	high	619/728	24.4	[-9.2, 50.7]
Qwen 35B-A3B	low/min.	646/728	15.0	[-9.2, 35.9]
Qwen 35B-A3B	high	646/728	13.6	[-10.4, 34.6]

Observed Δ_{gap} and bounds are percentage points on the same pooled matched-triplet sets. Bounds assign excluded triplets adversarially and are not confidence intervals.

E Full Quantitative Results

E.1 Matched Condition Accuracies

Table 11: Matched-triplet condition accuracy: complete, 25K.

Model	Mode	Matched n	Neutral	Hard	Targeted distractor
Gemma 31B	low/min.	500	73.4	68.2	43.4
Gemma 31B	high	451	84.3	82.0	55.0
DeepSeek V4 Flash	low/min.	500	60.2	50.8	37.6
DeepSeek V4 Flash	high	416	77.4	77.4	49.0
Nemotron Super	low/min.	500	57.4	50.6	39.8
Nemotron Super	high	488	73.6	72.5	44.3
Nemotron Nano	low/min.	479	19.4	17.7	13.8
Nemotron Nano	high	412	61.2	57.5	37.9
Qwen 27B	low/min.	473	66.0	61.1	44.0
Qwen 27B	high	471	79.2	76.2	45.9
Qwen 35B-A3B	low/min.	477	57.2	54.5	32.3
Qwen 35B-A3B	high	478	55.9	53.3	32.0

Accuracies are percentages on the shared valid-triplet denominator. Hard denotes related background filtered to exclude answer mentions. Paired-contrast confidence intervals and missingness bounds are provided in the analysis outputs listed in Appendix I.

Table 12: Matched-triplet condition accuracy: complete, 112K.

Model	Mode	Matched n	Neutral	Hard	Targeted distractor
Gemma 31B	low/min.	498	68.9	63.9	38.4
Gemma 31B	high	441	81.4	79.1	53.5
DeepSeek V4 Flash	low/min.	500	54.2	48.4	37.8
DeepSeek V4 Flash	high	418	77.3	72.7	45.5
Nemotron Super	low/min.	499	52.9	43.7	37.1
Nemotron Super	high	428	68.9	68.2	45.1
Nemotron Nano	low/min.	363	14.0	12.1	11.6
Nemotron Nano	high	277	44.0	49.8	38.3
Qwen 27B	low/min.	374	60.7	56.4	44.9
Qwen 27B	high	369	77.0	71.8	46.1
Qwen 35B-A3B	low/min.	400	46.0	45.8	33.5
Qwen 35B-A3B	high	401	47.6	47.6	34.4

Accuracies are percentages on the shared valid-triplet denominator. Hard denotes related background filtered to exclude answer mentions. Paired-contrast confidence intervals and missingness bounds are provided in the analysis outputs listed in Appendix I.

Table 13: Matched-triplet condition accuracy: consensus, 25K.

Model	Mode	Matched n	Neutral	Hard	Targeted distractor
Gemma 31B	low/min.	364	76.9	70.6	48.1
Gemma 31B	high	332	86.7	83.7	59.6
DeepSeek V4 Flash	low/min.	364	62.6	51.4	41.5
DeepSeek V4 Flash	high	308	78.6	78.9	54.2
Nemotron Super	low/min.	364	60.4	53.3	45.6
Nemotron Super	high	359	76.6	75.8	49.9
Nemotron Nano	low/min.	351	19.4	16.8	13.7
Nemotron Nano	high	303	63.7	61.4	38.0
Qwen 27B	low/min.	343	67.1	63.6	48.1
Qwen 27B	high	342	83.3	81.0	51.2
Qwen 35B-A3B	low/min.	346	59.0	55.2	34.4
Qwen 35B-A3B	high	346	58.1	53.2	34.1

Accuracies are percentages on the shared valid-triplet denominator. Hard denotes related background filtered to exclude answer mentions. Paired-contrast confidence intervals and missingness bounds are provided in the analysis outputs listed in Appendix I.

Table 14: Matched-triplet condition accuracy: consensus, 112K.

Model	Mode	Matched n	Neutral	Hard	Targeted distractor
Gemma 31B	low/min.	363	71.9	66.7	42.1
Gemma 31B	high	328	82.9	81.7	56.4
DeepSeek V4 Flash	low/min.	364	56.3	48.6	42.0
DeepSeek V4 Flash	high	310	79.4	74.5	51.6
Nemotron Super	low/min.	363	56.2	46.8	42.4
Nemotron Super	high	318	70.4	69.8	47.8
Nemotron Nano	low/min.	272	13.6	11.4	10.7
Nemotron Nano	high	197	45.7	52.8	40.6
Qwen 27B	low/min.	280	62.1	57.1	47.5
Qwen 27B	high	277	80.1	76.2	51.6
Qwen 35B-A3B	low/min.	300	47.7	47.3	34.3
Qwen 35B-A3B	high	300	50.0	49.0	35.0

Accuracies are percentages on the shared valid-triplet denominator. Hard denotes related background filtered to exclude answer mentions. Paired-contrast confidence intervals and missingness bounds are provided in the analysis outputs listed in Appendix I.

E.2 Consensus and Complete-Population Core Estimates

Table 15: Core estimates on the dual-auditor-consensus MuSiQue population. These are complete-case matched estimates in percentage points, with pointwise example-clustered percentile 95% bootstrap intervals.

Model	Mode	n_{all}	Δ_{diffuse}	Δ_{targeted}	Δ_{gap}
Gemma 31B	low/min.	727	5.8 [2.7, 9.1]	23.5 [19.5, 27.9]	17.7 [11.7, 24.3]
Gemma 31B	high	660	2.1 [-0.2, 4.4]	24.7 [20.0, 29.3]	22.6 [17.1, 27.9]
DeepSeek V4 Flash	low/min.	728	9.5 [6.5, 12.5]	8.2 [4.7, 12.0]	-1.2 [-6.5, 4.1]
DeepSeek V4 Flash	high	618	2.3 [0.0, 4.7]	23.8 [18.6, 28.9]	21.5 [15.0, 27.8]
Nemotron Super	low/min.	727	8.3 [4.8, 11.8]	6.1 [2.2, 10.2]	-2.2 [-8.2, 3.6]
Nemotron Super	high	677	0.7 [-2.2, 3.9]	24.1 [19.5, 28.6]	23.3 [17.2, 29.5]
Nemotron Nano	low/min.	623	2.4 [-0.5, 5.5]	2.1 [-0.2, 4.2]	-0.3 [-4.3, 3.9]
Nemotron Nano	high	500	-1.4 [-5.4, 2.6]	19.0 [14.3, 23.8]	20.4 [13.1, 28.1]
Qwen 27B	low/min.	623	4.2 [1.1, 7.4]	12.8 [9.4, 16.7]	8.7 [3.0, 14.3]
Qwen 27B	high	619	3.1 [0.7, 5.5]	27.5 [22.7, 32.2]	24.4 [18.7, 30.4]
Qwen 35B-A3B	low/min.	646	2.2 [-1.6, 5.8]	17.2 [12.9, 21.4]	15.0 [8.7, 21.8]
Qwen 35B-A3B	high	646	3.1 [-0.8, 7.0]	16.7 [12.5, 21.0]	13.6 [7.2, 20.2]

n_{all} counts matched $N/H/D$ triplets (example–scale observations) pooled across 25K and 112K; repeated scales remain clustered within example for bootstrap inference. Intervals are shown in brackets. Scale-specific penalties and strict target shifts are reported separately in the appendix.

Table 16: Core estimates on the complete MuSiQue population. These are complete-case matched estimates in percentage points, with pointwise example-clustered percentile 95% bootstrap intervals.

Model	Mode	n_{all}	Δ_{diffuse}	Δ_{targeted}	Δ_{gap}
Gemma 31B	low/min.	998	5.1 [2.4, 7.6]	25.2 [21.5, 28.9]	20.0 [14.8, 24.8]
Gemma 31B	high	892	2.2 [0.3, 4.4]	26.3 [22.2, 30.3]	24.1 [19.2, 28.9]
DeepSeek V4 Flash	low/min.	1,000	7.6 [5.1, 10.3]	11.9 [8.7, 15.1]	4.3 [-0.4, 9.2]
DeepSeek V4 Flash	high	834	2.3 [0.2, 4.4]	27.8 [23.5, 32.5]	25.5 [20.0, 31.2]
Nemotron Super	low/min.	999	8.0 [5.1, 11.1]	8.7 [5.4, 12.1]	0.7 [-4.5, 5.7]
Nemotron Super	high	916	0.9 [-1.6, 3.4]	25.9 [21.8, 30.0]	25.0 [19.8, 30.2]
Nemotron Nano	low/min.	842	1.8 [-0.7, 4.4]	2.5 [0.6, 4.4]	0.7 [-2.9, 4.2]
Nemotron Nano	high	689	-0.1 [-3.7, 3.3]	16.4 [12.6, 20.4]	16.5 [10.2, 22.7]
Qwen 27B	low/min.	847	4.6 [1.5, 7.6]	14.6 [11.5, 18.1]	10.0 [5.1, 15.1]
Qwen 27B	high	840	3.9 [1.7, 6.3]	28.3 [24.1, 32.6]	24.4 [19.3, 29.8]
Qwen 35B-A3B	low/min.	877	1.6 [-1.7, 4.9]	17.7 [14.1, 21.5]	16.1 [10.4, 21.8]
Qwen 35B-A3B	high	879	1.4 [-2.1, 4.7]	17.6 [14.0, 21.4]	16.3 [10.8, 22.2]

n_{all} is the pooled matched-triplet denominator, counting matched observations at 25K and 112K. Intervals are shown in brackets. Scale-specific penalties and strict target shifts are reported separately in the appendix.

E.3 Scale Interactions

Table 17: Targeted-distractor penalty by scale: complete population.

Model	Mode	n_6	$\Delta_{\text{targeted}}(25\text{K})$	$\Delta_{\text{targeted}}(112\text{K})$	Δ_{length}
Gemma 31B	low/min.	498	24.7 [20.7, 29.1]	25.5 [21.1, 29.9]	0.8 [-3.0, 4.6]
Gemma 31B	high	426	26.8 [22.1, 31.2]	25.6 [21.1, 30.1]	-1.2 [-5.2, 2.6]
DeepSeek V4 Flash	low/min.	500	13.2 [9.4, 17.2]	10.6 [7.0, 14.2]	-2.6 [-6.0, 1.0]
DeepSeek V4 Flash	high	390	27.7 [22.8, 32.6]	27.2 [22.1, 32.1]	-0.5 [-4.4, 3.3]
Nemotron Super	low/min.	499	10.8 [7.0, 14.8]	6.6 [2.8, 10.4]	-4.2 [-8.0, -0.2]
Nemotron Super	high	420	26.4 [21.4, 31.0]	22.6 [17.9, 27.6]	-3.8 [-8.6, 1.0]
Nemotron Nano	low/min.	355	2.8 [-0.3, 5.9]	0.6 [-1.4, 2.8]	-2.3 [-6.2, 1.4]
Nemotron Nano	high	244	15.6 [9.4, 21.7]	11.9 [6.6, 17.6]	-3.7 [-10.7, 3.3]
Qwen 27B	low/min.	357	15.1 [10.6, 19.6]	10.9 [6.7, 14.8]	-4.2 [-8.7, 0.0]
Qwen 27B	high	350	30.6 [25.4, 35.7]	25.4 [20.3, 30.6]	-5.1 [-9.7, -0.9]
Qwen 35B-A3B	low/min.	380	21.6 [16.8, 26.1]	12.1 [7.6, 16.8]	-9.5 [-14.7, -4.5]
Qwen 35B-A3B	high	381	21.3 [16.5, 26.2]	13.4 [8.9, 17.8]	-7.9 [-12.9, -3.1]

$\Delta_{\text{length}} = \Delta_{\text{targeted}}(112\text{K}) - \Delta_{\text{targeted}}(25\text{K})$. n_6 requires all six condition-by-scale outputs. Effects are percentage points with pointwise 95% bootstrap intervals.

Table 18: Targeted-distractor penalty by scale: consensus population.

Model	Mode	n_6	$\Delta_{\text{targeted}}(25\text{K})$	$\Delta_{\text{targeted}}(112\text{K})$	Δ_{length}
Gemma 31B	low/min.	363	22.6 [17.9, 27.5]	24.5 [19.8, 29.5]	1.9 [-2.5, 6.6]
Gemma 31B	high	316	23.7 [18.4, 29.1]	25.6 [20.6, 30.7]	1.9 [-2.2, 6.0]
DeepSeek V4 Flash	low/min.	364	9.9 [5.5, 14.3]	6.6 [2.2, 10.7]	-3.3 [-7.4, 1.1]
DeepSeek V4 Flash	high	289	23.5 [18.3, 29.1]	22.1 [15.9, 27.3]	-1.4 [-5.5, 3.1]
Nemotron Super	low/min.	363	7.7 [3.3, 12.4]	4.4 [-0.3, 9.1]	-3.3 [-8.0, 1.7]
Nemotron Super	high	314	23.6 [18.5, 29.0]	21.3 [15.9, 27.4]	-2.2 [-7.6, 3.5]
Nemotron Nano	low/min.	267	2.6 [-1.1, 6.0]	0.7 [-1.5, 3.0]	-1.9 [-6.4, 2.2]
Nemotron Nano	high	175	16.6 [9.7, 23.4]	13.1 [6.9, 19.4]	-3.4 [-12.0, 4.6]
Qwen 27B	low/min.	267	13.1 [8.2, 18.0]	9.0 [4.5, 13.5]	-4.1 [-9.4, 1.1]
Qwen 27B	high	262	29.4 [24.0, 35.1]	24.4 [18.7, 30.5]	-5.0 [-10.3, 0.4]
Qwen 35B-A3B	low/min.	283	20.8 [14.8, 26.9]	12.7 [7.4, 18.0]	-8.1 [-13.8, -2.8]
Qwen 35B-A3B	high	283	20.1 [14.8, 25.8]	14.1 [9.2, 19.1]	-6.0 [-11.3, -0.7]

$\Delta_{\text{length}} = \Delta_{\text{targeted}}(112\text{K}) - \Delta_{\text{targeted}}(25\text{K})$. n_6 requires all six condition-by-scale outputs. Effects are percentage points with pointwise 95% bootstrap intervals.

E.4 Supplementary Comparison of Reasoning Settings

As a supplementary analysis of the Stage-1 evaluation settings, we compare the same model's low-to-high accuracy gains under hard and targeted backgrounds:

$$I_R = (A_D^{\text{high}} - A_D^{\text{low}}) - (A_H^{\text{high}} - A_H^{\text{low}}).$$

On common valid hard/targeted-by-mode quadruples, the primary containment analysis gives negative 95% intervals for I_R in DeepSeek, Nemotron Super, Nemotron Nano, and Qwen 27B in both populations; Gemma 31B and Qwen 35B-A3B have intervals containing zero (Table 19). On consensus, the four supported differences correspond to 13.8–18.8 percentage points less gain in the targeted condition. On the same 627 valid consensus instances, DeepSeek's accuracy rises from 51.52% to 76.40% in hard backgrounds, versus 43.06% to 52.31% in targeted backgrounds: gains of 24.88 and 9.25 points, respectively. NoisyBench links longer sampled reasoning trajectories to lower accuracy under hard-negative distraction (Lee et al., 2026). Here, changing the reasoning setting improves all four models in both backgrounds, but smaller gains under targeted distraction widen the hard-to-targeted accuracy gap (Figure 4).

Table 19 reports I_R in percentage points with pointwise 95% example-clustered bootstrap intervals for both populations; n counts matched quadruples pooled across lengths.

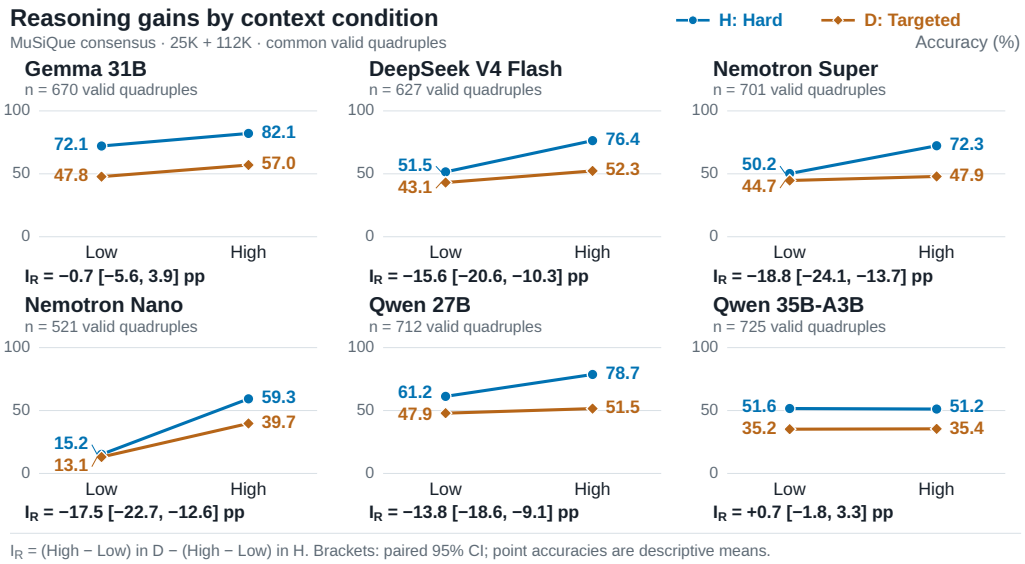


Figure 4: Supplementary comparison of within-model reasoning gains under hard and targeted backgrounds. Each panel uses the same valid hard/targeted-by-mode quadruples in the consensus population, pooling both lengths. Lines connect descriptive low/minimal and high accuracy means on a common scale. The paired interaction I_R and its pointwise example-clustered 95% bootstrap interval appear below each panel; negative values indicate smaller gains in the targeted condition. n counts matched context instances.

Table 19: Within-model reasoning interaction for the hard-to-targeted-distractor contrast.

Model	Population	n	I_R (targeted-minus-hard reasoning gain)
Gemma 31B	complete	909	-0.2 [-4.2, 3.7]
Gemma 31B	consensus	670	-0.7 [-5.6, 3.9]
DeepSeek V4 Flash	complete	847	-15.8 [-20.7, -11.0]
DeepSeek V4 Flash	consensus	627	-15.6 [-20.6, -10.3]
Nemotron Super	complete	954	-17.5 [-21.8, -13.1]
Nemotron Super	consensus	701	-18.8 [-24.1, -13.7]
Nemotron Nano	complete	709	-14.5 [-19.0, -10.2]
Nemotron Nano	consensus	521	-17.5 [-22.7, -12.6]
Qwen 27B	complete	976	-13.3 [-17.5, -9.3]
Qwen 27B	consensus	712	-13.8 [-18.6, -9.1]
Qwen 35B-A3B	complete	993	0.0 [-2.1, 2.3]
Qwen 35B-A3B	consensus	725	0.7 [-1.8, 3.3]

Negative I_R indicates a smaller high-minus-low accuracy gain in targeted-distractor contexts than in hard contexts. Provider controls are heterogeneous and not compute matched.

E.5 TARGET-DIRECTED ANSWER TRANSITIONS

Each response receives two binary scores: gold-answer correctness and exact normalized target matching. We label it C if only the gold check passes, CT if both pass, T if only the target check passes, and O if neither passes. Thus the target-match rate r_X includes both CT and T responses. The same valid H/D pairs are used for both condition-specific rates and all transition counts.

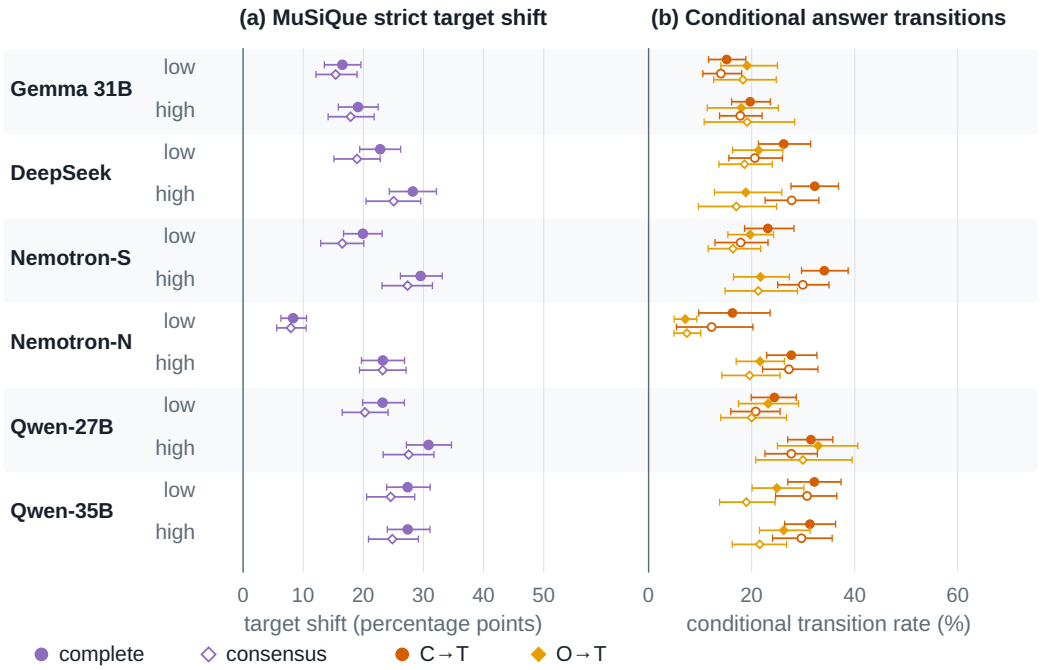


Figure 5: Target-directed answer movement on MuSiQue, pooling both lengths. Panel (a) reports the increase in exact target matches from H to D. Panel (b) reports $C \to T$ and $O \to T$ rates, conditioned on a gold-correct hard-context response or an O-state hard-context response, respectively. Bars show pointwise example-clustered 95% bootstrap intervals. Filled and open markers denote complete and dual-auditor-consensus populations, respectively.

Table 20 reports the corresponding answer-transition counts on the consensus population.

In all reported strict-match results, no hard-context response occupies state CT. Induced-targeted rates therefore divide $C \to T$ counts by the number of hard-context C responses, equivalently all gold-correct hard-context responses. Wrong-to-target rates divide $O \to T$ counts by the number of hard-context O responses.

Table 20: Consensus-population strict-match pooled directionality accounting.

Model	Mode	Valid	$C \to T$	$O \to T$	Rescues	Target acq.	Target drop
Gemma 31B	low/min.	727	70	41	17	113	1
Gemma 31B	high	671	98	22	12	120	0
DeepSeek V4 Flash	low/min.	728	75	64	35	139	1
DeepSeek V4 Flash	high	627	133	25	20	158	1
Nemotron Super	low/min.	727	65	56	52	121	1
Nemotron Super	high	702	152	40	25	192	0
Nemotron Nano	low/min.	641	11	40	15	51	0
Nemotron Nano	high	573	89	46	29	135	2
Qwen 27B	low/min.	716	91	54	16	145	0
Qwen 27B	high	712	155	44	17	199	3
Qwen 35B-A3B	low/min.	725	115	63	35	178	0
Qwen 35B-A3B	high	725	110	72	32	182	2

Counts use valid hard/targeted-distractor pairs pooled across context scales under strict normalized target matching; each row has 728 attempted pairs. $C \to T$ is an induced targeted failure; $O \to T$ is an accuracy-neutral target transfer; rescues are wrong-to-correct transitions.

E.6 Additional Gemma 26B-A4B Results

Gemma 26B-A4B completed the full matrix but is excluded from primary figures because provider output does not cleanly establish execution of the requested high-thinking mode. On the consensus population, its disabled/minimal Δ_{diffuse}, Δ_{targeted}, and Δ_{gap} are 5.7 [2.2, 9.2], 16.6 [12.3, 20.9], and 11.0 [4.8, 17.1] points; the requested-high estimates are 0.5 [-3.2, 4.2], 19.7 [15.4, 24.1], and 19.2 [12.5, 25.5].

E.7 Exploratory 256K/512K Boundary Probe

Table 21: DeepSeek V4 Flash long-context probe at 256K and 512K.

Length	Mode	Valid n	Hard	Targeted	Penalty [95% CI]	Holm p
256K	low/min.	80	53.8	40.0	13.8 [2.5, 26.2]	0.069
256K	high	71	78.9	53.5	25.4 [14.1, 36.6]	3.6×10^{-4}
512K	low/min.	80	43.8	35.0	8.8 [0.0, 18.8]	0.143
512K	high	72	72.2	44.4	27.8 [18.1, 37.5]	7.6×10^{-6}

Length	Valid quadruples	Targeted-minus-hard reasoning gain [95% CI]
256K	71	-11.3 [-25.4, 4.2]
512K	72	-19.4 [-31.9, -5.6]

Accuracies are percentages; hard-minus-targeted penalties and reasoning interactions are percentage points. Paired-bootstrap intervals are shown in brackets. The accuracy rows compare the same hard and targeted-distractor endpoints used by the primary analysis; the second panel defines the interaction as $(D_{\text{high}} - D_{\text{low}}) - (H_{\text{high}} - H_{\text{low}})$. Holm p values are the canonical length-specific exact McNemar adjustments.

This single-model boundary diagnostic uses a separate frozen 80-example subset and evaluates the hard-to-targeted-distractor endpoint. Its estimates are reported separately from the main scale interaction.

F Reasoning-Difference Diagnostics and Provenance

F.1 Reported Evidence Labels

Listing all gold document labels and answering correctly do not always coincide: a model may answer correctly without listing every gold label, or list every gold label while answering incorrectly. This table uses valid responses from the original 500-question MuSiQue population, pooling all three

Table 22: Answer correctness and whether the model lists all gold document labels. Results use the original 500-question MuSiQue population, pooled across the three background conditions and both context lengths.

Model	Mode	OK n	Correct answer		Wrong answer		Conditional rates	
			All gold listed	Not all gold listed	All gold listed	Not all gold listed	Accuracy when all gold listed	All gold listed when correct
Gemma 31B	low/min.	2,998	40.5	18.8	8.2	32.4	83.1	68.3
Gemma 31B	high	2,827	57.4	13.5	8.0	21.2	87.8	81.0
DeepSeek V4 Flash	low/min.	3,000	17.1	31.1	8.9	43.0	65.9	35.5
DeepSeek V4 Flash	high	2,739	49.2	16.5	12.2	22.1	80.1	74.8
Nemotron Super	low/min.	2,999	13.7	33.2	6.8	46.2	66.9	29.3
Nemotron Super	high	2,897	31.1	29.8	5.9	33.2	84.0	51.1
Nemotron Nano	low/min.	2,777	0.9	13.5	5.3	80.3	14.0	6.0
Nemotron Nano	high	2,613	7.6	37.5	2.3	52.5	76.5	16.9
Qwen 27B	low/min.	2,832	33.2	22.3	9.0	35.5	78.6	59.8
Qwen 27B	high	2,825	54.1	11.4	9.8	24.7	84.7	82.5
Qwen 35B-A3B	low/min.	2,872	15.9	29.4	7.8	46.9	67.1	35.0
Qwen 35B-A3B	high	2,873	15.5	29.9	7.9	46.7	66.3	34.2

Gold documents are the benchmark's annotated supporting documents; all remain in the input. All gold listed means the model reports every gold document label, with additional labels allowed. Not all gold listed means at least one gold label is omitted. The four Correct/Wrong columns give percentages of all valid (OK) responses. The final two columns give the percentage correct among responses listing all gold labels and the percentage listing all gold labels among correct responses, respectively. All rates are descriptive percentages based on reported labels.

background conditions and both context lengths. Across primary model–mode cells, 6.2%–65.4% of responses list every gold label. Among these responses, answer accuracy ranges from 14.0% to 87.8%. Conversely, among correct responses, 6.0%–82.5% list every gold label. The interventions below directly test the assistance offered by label cues and additional evidence-connection guidance on selected cases.

F.2 Stage-2 Selection and Hint Acceptance

Stage 2 diagnoses selected disagreements between the low and high reasoning settings of the same model, across N, H, and D. Each intervention re-evaluates low reasoning on the saved context with additional guidance. Across the six primary models, Stage 2 identifies 2,975 eligible context instances among 16,593 valid low/high pairs in the complete population and constructs 2,970 diagnostic cases (17.9% of valid pairs). In the consensus population, it identifies 2,307 eligible instances among 12,195 valid pairs and constructs 2,303 cases (18.9%). The five complete-population and four consensus-population differences are Nemotron Nano cases without the required saved labels. These are context instances, not unique questions. The original low accuracy is 0% and the paired high accuracy 100% by construction.

Table 23 reports per-model selection and hint-acceptance counts.

Teacher- and oracle-label calls cover all constructed cases. Reported comparisons use accepted-hint cases with valid teacher, oracle, and teacher+hint outputs. Appendix C.2 specifies hint acceptance and retry handling.

Figure 3 presents consensus assisted accuracy and paired increments for both populations. Table 24 reports assisted accuracy and paired contrasts for both populations; Table 25 separates the hint increments by background condition.

Table 23: Stage-2 selection and hint acceptance by model and population.

Model	Valid pairs	Eligible	Selected	Accepted	Matched	Accept (%)
Consensus population						
Gemma 31B	2,079	274	274	228	227	83.2
DeepSeek V4 Flash	2,027	482	482	430	430	89.2
Nemotron Super	2,125	431	431	310	285	71.9
Nemotron Nano	1,801	677	673	486	471	72.2
Qwen 27B	2,063	373	373	310	303	83.1
Qwen 35B-A3B	2,100	70	70	35	35	50.0
Complete population						
Gemma 31B	2,825	372	372	304	301	81.7
DeepSeek V4 Flash	2,739	611	611	538	538	88.1
Nemotron Super	2,896	585	585	421	383	72.0
Nemotron Nano	2,440	849	844	589	563	69.8
Qwen 27B	2,822	461	461	379	371	82.2
Qwen 35B-A3B	2,871	97	97	45	45	46.4

Counts are context instances pooled across the three background conditions and both lengths. Valid pairs have valid low and high responses; eligible pairs are low-wrong/high-correct. Selected cases additionally require teacher and oracle labels; the missing-label exclusions are four consensus and five complete Nemotron Nano cases. Accepted hints pass the checks in Appendix C.2. Matched cases have valid outputs for all three interventions. Accept (%) is $100 \times \mathrm{Accepted}/\mathrm{Selected}$.

Table 24: Stage-2 assisted low-mode accuracy and paired contrasts on the same matched cases.

Model	Valid/ accepted	Assisted accuracy (%)			Paired increment (pp)	
		Teacher	Oracle	Teacher+hint	Hint	Oracle
Consensus population						
		57.7	58.6	82.8	25.1	0.9
Gemma 31B	227/228	[48.7, 67.7]	[49.2, 68.4]	[75.9, 89.4]	[14.9, 34.5]	[-3.8, 5.8]
		45.1	43.5	84.4	39.3	-1.6
DeepSeek V4 Flash	430/430	[36.9, 52.9]	[35.5, 51.9]	[78.8, 89.6]	[32.2, 46.0]	[-5.8, 2.2]
		46.7	46.3	77.2	30.5	-0.4
Nemotron Super	285/310	[38.5, 55.1]	[38.3, 54.7]	[70.0, 84.1]	[22.6, 38.0]	[-5.0, 4.0]
		13.6	12.1	40.1	26.5	-1.5
Nemotron Nano	471/486	[9.8, 18.0]	[8.1, 16.8]	[34.6, 45.8]	[21.0, 32.3]	[-4.0, 1.0]
		44.6	46.5	76.6	32.0	2.0
Qwen 27B	303/310	[36.7, 52.8]	[38.4, 55.2]	[69.3, 83.8]	[23.7, 41.0]	[-2.0, 6.2]
		65.7	65.7	74.3	8.6	0.0
Qwen 35B-A3B	35/35	[48.5, 81.6]	[48.6, 81.6]	[58.3, 88.2]	[-11.8, 29.4]	[-17.6, 17.6]
Complete population						
		59.1	58.8	83.7	24.6	-0.3
Gemma 31B	301/304	[51.3, 66.9]	[50.5, 66.7]	[77.7, 89.1]	[16.3, 33.3]	[-5.5, 4.7]
		47.6	45.0	83.3	35.7	-2.6
DeepSeek V4 Flash	538/538	[40.8, 55.1]	[38.4, 51.7]	[78.3, 88.0]	[29.3, 42.0]	[-6.4, 1.2]
		49.3	48.3	77.5	28.2	-1.0
Nemotron Super	383/421	[42.7, 56.7]	[41.4, 55.6]	[71.7, 82.9]	[21.5, 34.7]	[-4.8, 2.7]
		15.6	13.1	42.5	26.8	-2.5
Nemotron Nano	563/589	[11.9, 19.7]	[9.6, 17.4]	[37.1, 48.0]	[21.9, 31.7]	[-4.9, 0.0]
		49.3	49.6	77.4	28.0	0.3
Qwen 27B	371/379	[41.8, 56.7]	[41.9, 57.4]	[70.2, 83.9]	[20.7, 35.3]	[-3.4, 3.9]
		68.9	68.9	73.3	4.4	0.0
Qwen 35B-A3B	45/45	[54.3, 82.6]	[54.3, 83.0]	[59.5, 86.7]	[-14.6, 24.0]	[-14.0, 14.0]

Each estimate is followed by its pointwise 95% example-clustered bootstrap interval. All rates use the accepted-hint cases with valid outputs for all three interventions; Valid/accepted gives this common denominator over accepted-hint attempts. Hint and oracle increments are teacher+hint minus teacher and oracle minus teacher, respectively, in percentage points.

Table 25: Condition-specific Stage-2 hint increments in the MuSiQue consensus population.

Model	Condition	Valid/accepted	Hint − teacher [95% CI]
Gemma 31B	neutral	75/75	24.0 [7.9, 40.0]
Gemma 31B	hard	70/70	27.1 [11.9, 42.5]
Gemma 31B	targeted distractor	82/83	24.4 [12.8, 36.0]
DeepSeek V4 Flash	neutral	137/137	38.7 [29.2, 47.3]
DeepSeek V4 Flash	hard	179/179	41.9 [33.1, 50.3]
DeepSeek V4 Flash	targeted distractor	114/114	36.0 [26.1, 45.2]
Nemotron Super	neutral	91/98	41.8 [30.2, 53.8]
Nemotron Super	hard	125/137	28.0 [19.3, 37.2]
Nemotron Super	targeted distractor	69/75	20.3 [7.5, 34.2]
Nemotron Nano	neutral	158/160	31.6 [22.4, 40.5]
Nemotron Nano	hard	189/197	29.1 [22.3, 35.9]
Nemotron Nano	targeted distractor	124/129	16.1 [8.5, 24.2]
Qwen 27B	neutral	97/101	32.0 [17.8, 45.5]
Qwen 27B	hard	118/119	33.9 [23.7, 44.1]
Qwen 27B	targeted distractor	88/90	29.5 [18.2, 40.7]
Qwen 35B-A3B	neutral	13/13	7.7 [-23.1, 38.5]
Qwen 35B-A3B	hard	11/11	27.3 [0.0, 54.5]
Qwen 35B-A3B	targeted distractor	11/11	-9.1 [-40.0, 20.0]

The increment is teacher+hint assisted accuracy minus teacher-label assisted accuracy in percentage points. Valid/accepted is the exact accepted-hint subset valid for all three interventions over accepted-hint attempts. Condition-specific assisted accuracy levels, selected/valid low–high counts, complete-population estimates, and condition-by-length rows remain in `stage2_selection_counts.csv` and `stage2_matched_hint_contrasts.csv`.

F.3 Post-hoc Checks of Document Cues and Newly Induced Errors

We use two subsets of the existing Stage-2 results to examine when the intervention prompts help. These are post-hoc analyses: the subset rules were specified after the main experiments and before calculating the subgroup results. We report consensus results as primary and complete-population results as a robustness reference.

When the document list is already correct. We ask whether teacher+hint improves on teacher labels when the supplied document list already matches the gold annotations. We select cases whose teacher-label set is nonempty and contains exactly the annotated gold documents, with no extra documents; label order is ignored. The cases span neutral, hard, and targeted backgrounds at both lengths. This comparison tests the value of further guidance when the document pointers are already correct, while the model must still extract and combine the relevant facts.

Errors that appear under targeted distraction. We ask whether the Stage-2 interventions help with the new failures observed in Stage 1. For the same model, question, and context length, low mode must answer correctly under the hard background (H) but incorrectly under targeted distraction (D), while high mode answers correctly under D. All three responses must be valid; N and H-high responses do not enter this selection. We then examine the interventions on these same D contexts, linking the Stage-2 results to the observed H-to-D failures.

Matched comparisons and confidence intervals. Both subsets use the original Stage-2 scoring, validity, and hint-acceptance rules (Appendix C.2). We retain accepted-hint cases with valid answers for all three interventions, so their accuracies and paired differences use the same cases. We use each model's own saved teacher labels and link intervention answers by model, original context ID, and prompt variant.

We report both the number of context instances (n) and distinct questions (u), since one question can appear with several backgrounds or lengths. To obtain confidence intervals, we resample questions with replacement 2,000 times (seed 20260925), keeping each question's retained context instances together. Each draw is shared across all three interventions and their paired comparisons, with equal weight per context instance. The 2.5th and 97.5th percentiles give pointwise 95% intervals without multiple-comparison adjustment. The two subsets can contain some of the same questions and contexts.

Table 2 summarizes consensus hint increments in the main text. Tables 26 and 27 provide assisted accuracies and selection flows for all six models in both populations. Improved/worsened counts distinguish teacher-label errors corrected by the hint prompt from teacher-label successes lost under that prompt. The accompanying saved analysis files provide the full selection flow, case membership, condition/length breakdowns, and all three pairwise contrasts.

Table 26: Post-hoc assisted low-mode accuracy when teacher and gold document-label sets coincide. Consensus results are primary; the complete population is a robustness reference.

(a) Matched assisted accuracy and paired hint increment

Model	n/u	Teacher	Oracle	Teacher+hint	Hint−teacher	I/W
Consensus (primary)						
Gemma 31B	163/85	61.3 [50.6, 73.5]	57.7 [47.2, 69.2]	83.4 [74.4, 91.8]	22.1 [10.4, 34.2]	45/9
DeepSeek V4 Flash	345/126	41.4 [33.4, 50.6]	40.9 [32.7, 49.5]	82.9 [76.7, 88.7]	41.4 [33.9, 49.1]	145/2
Nemotron Super	158/92	46.2 [35.7, 57.2]	46.8 [37.1, 57.2]	75.9 [66.2, 84.8]	29.7 [19.3, 40.3]	51/4
Nemotron Nano	97/56	12.4 [4.3, 21.7]	11.3 [4.4, 19.6]	50.5 [38.9, 62.8]	38.1 [26.6, 51.0]	38/1
Qwen 27B	233/102	39.9 [31.2, 48.8]	44.2 [34.9, 53.7]	75.5 [66.8, 83.8]	35.6 [25.9, 45.6]	91/8
Qwen 35B-A3B	11/9	100.0 [100.0, 100.0]	81.8 [63.6, 100.0]	81.8 [54.5, 100.0]	-18.2 [-45.5, 0.0]	0/2
Complete (robustness)						
Gemma 31B	224/116	61.2 [51.7, 70.8]	58.0 [49.1, 67.1]	84.4 [77.0, 90.7]	23.2 [13.1, 33.8]	64/12
DeepSeek V4 Flash	431/160	45.0 [37.1, 52.8]	42.9 [35.3, 50.6]	81.2 [75.5, 86.7]	36.2 [29.2, 44.0]	167/11
Nemotron Super	205/120	49.8 [41.2, 58.6]	49.3 [41.1, 57.9]	75.6 [67.5, 82.6]	25.9 [17.0, 34.2]	59/6
Nemotron Nano	104/62	12.5 [4.7, 21.1]	11.5 [4.8, 19.6]	51.0 [38.8, 62.5]	38.5 [26.7, 50.5]	41/1
Qwen 27B	287/133	44.9 [36.7, 53.1]	47.4 [39.0, 56.2]	76.0 [67.7, 83.3]	31.0 [22.2, 39.4]	99/10
Qwen 35B-A3B	15/13	93.3 [76.9, 100.0]	80.0 [61.5, 100.0]	80.0 [53.8, 100.0]	-13.3 [-40.0, 13.4]	1/3

(b) Selection flow: context instances / independent questions

Model	Selected	Exact	Accepted	Matched
Consensus (primary)				
Gemma 31B	274/129	186/92	163/85	163/85
DeepSeek V4 Flash	482/160	382/133	345/126	345/126
Nemotron Super	431/172	224/109	171/95	158/92
Nemotron Nano	673/234	123/65	100/57	97/56
Qwen 27B	373/139	289/114	240/102	233/102
Qwen 35B-A3B	70/62	20/17	11/9	11/9
Complete (robustness)				
Gemma 31B	372/177	259/127	226/116	224/116
DeepSeek V4 Flash	611/208	485/173	431/160	431/160
Nemotron Super	585/236	288/145	223/125	205/120
Nemotron Nano	844/298	137/75	108/63	104/62
Qwen 27B	461/186	357/152	295/133	287/133
Qwen 35B-A3B	97/86	30/27	15/13	15/13

Assisted accuracy is in percent; Hint−teacher is in percentage points. Brackets give pointwise 95% question-clustered bootstrap intervals (2,000 draws; seed 20260925), with context instances weighted equally. n/u denotes matched contexts / independent questions; I/W counts paired improvements / worsenings for teacher+hint versus teacher. Selected denotes the original Stage-2 selected cases; Exact requires a nonempty teacher label set equal to the gold set. Accepted requires an accepted hint; Matched additionally requires valid outputs for all three interventions. All assisted accuracy estimates and contrasts use that same matched subset.

Table 27: Post-hoc assisted low-mode accuracy on Stage-1-induced failures with successful high-mode answers. Consensus results are primary; the complete population is a robustness reference.

(a) Matched assisted accuracy and paired hint increment

Model	n/u	Teacher	Oracle	Teacher+hint	Hint−teacher	I/W
Consensus (primary)						
Gemma 31B	49/40	73.5 [60.4, 86.0]	77.6 [62.2, 89.8]	89.8 [80.9, 97.9]	16.3 [4.7, 28.6]	9/1
DeepSeek V4 Flash	22/18	63.6 [41.7, 87.0]	54.5 [31.8, 78.9]	86.4 [70.0, 100.0]	22.7 [5.0, 41.7]	5/0
Nemotron Super	23/20	73.9 [55.0, 90.9]	73.9 [56.5, 90.5]	87.0 [71.4, 100.0]	13.0 [0.0, 27.3]	3/0
Nemotron Nano	13/13	53.8 [30.8, 76.9]	46.2 [23.1, 76.9]	84.6 [61.5, 100.0]	30.8 [7.7, 53.8]	4/0
Qwen 27B	28/23	67.9 [46.4, 86.2]	78.6 [61.5, 93.3]	89.3 [75.9, 100.0]	21.4 [3.3, 42.9]	7/1
Qwen 35B-A3B	8/8	75.0 [37.5, 100.0]	75.0 [37.5, 100.0]	62.5 [25.0, 87.5]	-12.5 [-50.0, 25.0]	1/2
Complete (robustness)						
Gemma 31B	66/53	72.7 [60.3, 83.9]	77.3 [64.7, 88.1]	87.9 [79.4, 95.3]	15.2 [6.2, 25.4]	11/1
DeepSeek V4 Flash	29/24	65.5 [45.2, 84.4]	62.1 [42.4, 81.5]	86.2 [72.4, 96.7]	20.7 [6.7, 37.5]	6/0
Nemotron Super	31/26	67.7 [51.6, 82.8]	71.0 [55.2, 85.3]	83.9 [69.7, 96.6]	16.1 [3.6, 29.4]	5/0
Nemotron Nano	17/17	47.1 [23.5, 70.6]	47.1 [23.5, 70.6]	70.6 [47.1, 88.2]	23.5 [5.9, 47.1]	4/0
Qwen 27B	39/31	76.9 [60.5, 90.5]	82.1 [68.6, 93.0]	92.3 [82.9, 100.0]	15.4 [2.3, 31.8]	7/1
Qwen 35B-A3B	14/13	78.6 [53.8, 100.0]	78.6 [53.8, 100.0]	64.3 [37.5, 92.3]	-14.3 [-50.0, 23.1]	2/4

(b) Selection flow: context instances / independent questions

Model	H_c/D_w	D_c^{high}	Selected	Accepted	Matched
Consensus (primary)					
Gemma 31B	188/121	66/50	66/50	50/41	49/40
DeepSeek V4 Flash	95/66	24/20	24/20	22/18	22/18
Nemotron Super	96/67	31/25	31/25	25/22	23/20
Nemotron Nano	28/26	17/17	17/17	14/14	13/13
Qwen 27B	110/77	34/27	34/27	29/23	28/23
Qwen 35B-A3B	154/104	16/16	16/16	8/8	8/8
Complete (robustness)					
Gemma 31B	272/177	92/69	92/69	67/54	66/53
DeepSeek V4 Flash	158/104	32/27	32/27	29/24	29/24
Nemotron Super	145/100	42/33	42/33	34/28	31/26
Nemotron Nano	41/38	21/21	21/21	18/18	17/17
Qwen 27B	165/115	50/39	50/39	40/31	39/31
Qwen 35B-A3B	208/146	22/21	22/21	14/13	14/13

Assisted accuracy is in percent; Hint−teacher is in percentage points. Brackets give pointwise 95% question-clustered bootstrap intervals (2,000 draws; seed 20260925), with context instances weighted equally. n/u denotes matched contexts / independent questions; I/W counts paired improvements / worsenings for teacher+hint versus teacher. H_c/D_w counts valid pairs with H-low correct and D-low wrong; D_c^{high} additionally requires a valid, correct D-high answer for the same model, question, and length. Selected retains the original Stage-2 selection. Accepted requires an accepted hint; Matched additionally requires valid outputs for all three interventions. All assisted accuracy estimates and contrasts use that same matched subset.

F.4 POPULATION-WIDE GOLD-ONLY REFERENCE

Table 28: Population-wide gold-only accuracy reference.

Model	Population	Valid/scope	Low/min.	High	High–low
Gemma 31B	complete	490/500	76.1 [72.2, 80.0]	80.8 [77.1, 84.3]	4.7 [1.4, 8.0]
Gemma 31B	consensus	360/364	77.5 [73.1, 81.7]	82.8 [78.9, 86.7]	5.3 [1.4, 9.2]
DeepSeek V4 Flash	complete	476/500	69.1 [64.9, 73.1]	80.7 [77.1, 84.2]	11.6 [7.8, 15.5]
DeepSeek V4 Flash	consensus	354/364	70.9 [66.1, 75.4]	82.8 [78.5, 86.7]	11.9 [7.1, 16.4]
Nemotron Super	complete	498/500	66.9 [62.8, 71.1]	75.7 [71.9, 79.5]	8.8 [4.8, 12.7]
Nemotron Super	consensus	362/364	68.8 [64.1, 73.5]	80.4 [76.2, 84.5]	11.6 [7.5, 16.3]
Nemotron Nano	complete	497/500	44.1 [39.6, 48.3]	71.4 [67.4, 75.3]	27.4 [22.9, 31.8]
Nemotron Nano	consensus	362/364	46.1 [41.2, 51.7]	74.0 [69.6, 78.5]	27.9 [22.7, 33.1]
Qwen 27B	complete	500/500	70.2 [66.2, 74.0]	81.2 [77.6, 84.6]	11.0 [7.6, 14.6]
Qwen 27B	consensus	364/364	72.8 [68.1, 77.2]	84.9 [81.3, 88.5]	12.1 [8.0, 15.9]
Qwen 35B-A3B	complete	500/500	68.0 [63.8, 72.0]	70.6 [66.6, 74.6]	2.6 [0.4, 5.0]
Qwen 35B-A3B	consensus	364/364	68.4 [63.7, 73.1]	70.6 [65.9, 75.0]	2.2 [-0.5, 4.9]

Valid/scope reports complete-case pairs with both outputs valid over the full 500/364 attempted example-pair scope. The gold-only reference is separate from the Stage-2 post-selected subset. Low/min. and high accuracies are percentages; high–low differences are percentage points. Intervals are pointwise 95% bootstrap intervals.

G 2WIKI ENDPOINT REPLICATION

The 2Wiki adapter requires a non-binary answer and at least two supporting titles that can be matched to context documents. The frozen sample contains 94 two-hop and 26 four-hop questions. Gold position is balanced, with 30 examples assigned to each of front, middle, end, and random. Mean construction-token counts are 25,115 for hard and 25,124 for the targeted-distractor condition.

Table 29: 2WikiMultiHopQA endpoint replication at approximately 25K tokens.

Model	Population	n	Accuracy penalty	Strict target shift
DeepSeek V4 Flash	complete	120	22.5 [15.0, 30.0]	27.5 [20.0, 35.8]
DeepSeek V4 Flash	consensus	112	23.2 [16.1, 31.2]	28.6 [19.6, 37.5]
Nemotron Super	complete	120	26.7 [18.3, 34.2]	29.2 [21.7, 37.5]
Nemotron Super	consensus	112	25.9 [17.9, 33.0]	28.6 [19.6, 36.6]
Qwen 27B	complete	118	16.1 [9.3, 23.7]	18.6 [11.8, 26.3]
Qwen 27B	consensus	110	16.4 [9.1, 23.6]	18.2 [10.9, 26.4]

The replication evaluates the hard to targeted-distractor comparison at approximately 25K tokens under disabled/nonthinking settings.

The three models use the same disabled/nonthinking provider configurations as the MuSiQue study. DeepSeek and Nemotron have all 120 complete and 112 consensus pairs; Qwen has 118 and 110 valid pairs. Missing calls follow the same complete-case and sensitivity policy as the main study.

H SCORING AND AUDIT-SELECTION SENSITIVITY

We examine whether the reported performance comparisons depend on how answers are scored or which questions are retained after passage auditing. Both checks use existing model responses.

For the *scoring check*, we evaluate the same matched answers using containment, normalized exact match (EM), and token F1, on both the complete and consensus sets. For the *audit-selection check*, we keep containment scoring and compare question groups defined by the auditors' existing judgments. Stage-2 case selection and hint acceptance retain their original decisions; the audit check filters these cases by question-group membership.

Each table below identifies the scoring rule, question group, and comparison being checked. The two post-hoc subgroups in Appendix F.3 use the original containment scoring on the complete and consensus sets. Those subgroups, target-match increases, and interactions with context length and reasoning settings are outside these additional checks.

H.1 Answer-Scoring Sensitivity

All three metrics use the normalization in Appendix D and only the canonical answer, without raw benchmark aliases. Containment is the original bidirectional substring rule. Normalized EM gives one only when the nonempty prediction and answer strings are equal. Token F1 uses multiset overlap of whitespace-delimited normalized tokens:

$$\mathrm{F1}(p, g) = \frac{2 \sum_t \min\{c_p(t), c_g(t)\}}{|p| + |g|},$$

where $c_p(t)$ and $c_g(t)$ count occurrences of token t. Empty normalized strings score zero under each metric. F1 measures continuous token overlap. EM can reject an answer embedded in an otherwise correct sentence, while containment can accept a partial answer or a response listing incompatible candidates.

The same valid triplets, replication pairs, and original Stage-2 matched sets are used for every metric; non-OK responses remain missing. We recompute row-weighted estimates and pointwise percentile intervals from 2,000 whole example-cluster draws, using seed 20260919. Draws are shared across metrics within each matched group, and the machine-readable outputs also report paired differences from containment. Stage-2 low-wrong/high-correct selection and accepted hints remain frozen under the original scorer. Its original 0%/100% selection endpoints therefore need not hold under EM or F1. Gold-only evaluation remains a separate population-wide reference.

The scoring check supports the two main findings reported in the text. On the consensus set, all three metrics show that the performance loss from hard to targeted backgrounds exceeds that from neutral to hard backgrounds in the same nine of 12 model–setting comparisons, and that teacher+hint outperforms teacher labels alone in the same five of six models. The 95% intervals for these nine Δ_{gap} contrasts and five overall Stage-2 prompt gains remain wholly above zero. All 12 pooled targeted-penalty point estimates are positive, with intervals above zero in the same 11 comparisons; all three 2Wiki penalties also retain positive intervals. On the complete set, the same nine positive Δ_{gap} intervals and five positive Stage-2 prompt-gain intervals persist, while positive targeted-penalty intervals decrease from 12/12 under containment to 11/12 under EM and F1.

The numerical effects can nevertheless change substantially. For consensus Gemma-low, the targeted penalty is 23.52 accuracy percentage points [19.09, 27.92] under containment, 10.73 accuracy percentage points [7.83, 13.76] under EM, and 13.03 F1 points [10.21, 16.00]. The saved outputs record lexical-scoring disagreements without semantic adjudication.

Gemma's gold-only contrast changes sign across scorers. For consensus Gemma 31B, high-minus-low is +5.28 accuracy percentage points [1.67, 9.17] under containment, -7.22 accuracy percentage points [-11.67, -2.50] under EM, and -7.43 F1 points [-11.17, -3.52]. The complete population has the same reversal. The gold-only high-minus-low difference is therefore scorer-dependent, while the matched Stage-2 hint increments retain positive intervals for the same five models.

The recomputed containment points and denominators reproduce the canonical analyses, but finite-bootstrap endpoints can differ because the draws differ. Two secondary comparisons illustrate the boundary: consensus DeepSeek-high's diffuse lower endpoint changes from zero to +0.16 accuracy percentage points, and complete-population Nemotron Nano's oracle-minus-teacher interval changes from [-4.89, 0.00] to [-5.09, -0.16]. We retain the original canonical intervals in the main result figures and tables and report the recomputed intervals in the sensitivity tables.

Tables 30–36 report pooled point estimates and intervals for both original populations; the source CSVs retain condition- and length-specific estimates and paired metric differences.

Table 30: Scoring sensitivity of the diffuse penalty Δ_{diffuse} on MuSiQue, pooling 25K and 112K contexts.

Model	Mode	n/u	Containment	Normalized EM	Token F1
Complete (500 examples before output filtering)					
Gemma 31B	low	998/500	5.1 [2.5, 7.8]	4.0 [2.0, 6.1]	5.4 [3.6, 7.3]
Gemma 31B	high	892/466	2.2 [0.3, 4.3]	5.5 [3.2, 7.7]	5.1 [3.4, 6.9]
DeepSeek	low	1000/500	7.6 [5.0, 10.3]	4.6 [2.0, 7.0]	5.6 [3.4, 7.9]
DeepSeek	high	834/444	2.3 [0.2, 4.4]	1.4 [−1.0, 3.8]	0.9 [−0.7, 2.5]
Nemotron S	low	999/500	8.0 [5.2, 11.0]	5.5 [3.0, 8.0]	5.4 [3.1, 7.7]
Nemotron S	high	916/496	0.9 [−1.7, 3.5]	2.5 [0.1, 4.9]	1.3 [−0.8, 3.4]
Nemotron N	low	842/487	1.8 [−0.9, 4.3]	2.3 [0.4, 4.2]	0.6 [−1.7, 2.7]
Nemotron N	high	689/445	−0.1 [−3.8, 3.5]	0.6 [−2.6, 3.6]	0.7 [−2.3, 3.8]
Qwen 27B	low	847/490	4.6 [1.6, 7.6]	1.7 [−1.0, 4.4]	3.7 [1.3, 6.3]
Qwen 27B	high	840/490	3.9 [1.7, 6.3]	1.3 [−1.1, 3.8]	2.7 [0.8, 4.7]
Qwen 35B	low	877/497	1.6 [−1.5, 4.7]	0.9 [−1.9, 3.7]	1.1 [−1.7, 3.9]
Qwen 35B	high	879/498	1.4 [−2.0, 4.6]	1.0 [−1.9, 3.9]	1.4 [−1.5, 4.2]
Both auditors valid (364 examples before output filtering)					
Gemma 31B	low	727/364	5.8 [2.5, 9.1]	5.1 [2.7, 7.6]	6.2 [4.0, 8.3]
Gemma 31B	high	660/344	2.1 [−0.2, 4.4]	6.4 [3.3, 9.3]	5.5 [3.2, 7.8]
DeepSeek	low	728/364	9.5 [6.2, 12.8]	5.6 [2.6, 8.7]	7.4 [4.4, 10.1]
DeepSeek	high	618/329	2.3 [0.2, 4.8]	0.8 [−2.1, 3.6]	0.6 [−1.2, 2.4]
Nemotron S	low	727/364	8.3 [4.7, 12.0]	5.2 [2.2, 8.4]	5.7 [2.8, 8.6]
Nemotron S	high	677/363	0.7 [−2.4, 3.6]	1.2 [−1.6, 4.2]	0.4 [−2.0, 2.7]
Nemotron N	low	623/356	2.4 [−0.5, 5.3]	3.7 [1.6, 6.0]	1.5 [−1.0, 4.0]
Nemotron N	high	500/325	−1.4 [−5.5, 2.6]	0.0 [−3.8, 3.7]	−0.3 [−3.7, 3.0]
Qwen 27B	low	623/356	4.2 [1.1, 7.3]	1.4 [−1.6, 4.3]	3.4 [0.6, 6.1]
Qwen 27B	high	619/357	3.1 [0.7, 5.5]	1.5 [−1.3, 3.9]	2.0 [−0.1, 4.0]
Qwen 35B	low	646/363	2.2 [−1.6, 6.0]	1.4 [−2.0, 4.6]	1.6 [−1.6, 4.8]
Qwen 35B	high	646/363	3.1 [−0.8, 7.0]	2.5 [−0.8, 5.8]	3.0 [−0.1, 6.2]

Cells give estimates [pointwise 95% cluster-bootstrap intervals]. Containment and EM differences are accuracy percentage points; token-F1 differences are F1 points, with scores multiplied by 100. n/u denotes valid matched sets / distinct examples; the same sets are used for all three metrics. Low denotes the disabled-or-minimal setting. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 31: Scoring sensitivity of the targeted penalty Δ_{targeted} on MuSiQue, pooling 25K and 112K contexts.

Model	Mode	n/u	Containment	Normalized EM	Token F1
Complete (500 examples before output filtering)					
Gemma 31B	low	998/500	25.2 [21.3, 28.8]	11.8 [9.3, 14.3]	13.8 [11.2, 16.3]
Gemma 31B	high	892/466	26.3 [22.5, 30.4]	21.3 [17.6, 25.2]	23.7 [20.4, 27.0]
DeepSeek	low	1000/500	11.9 [8.5, 15.3]	8.0 [5.4, 10.7]	8.5 [5.7, 11.3]
DeepSeek	high	834/444	27.8 [23.4, 32.1]	18.9 [14.9, 23.0]	21.5 [17.9, 25.0]
Nemotron S	low	999/500	8.7 [5.5, 12.1]	6.1 [3.3, 9.0]	5.9 [3.2, 8.6]
Nemotron S	high	916/496	25.9 [22.1, 30.0]	18.1 [14.5, 22.2]	19.7 [16.5, 23.0]
Nemotron N	low	842/487	2.5 [0.6, 4.4]	1.2 [−0.2, 2.6]	0.5 [−1.2, 2.2]
Nemotron N	high	689/445	16.4 [12.2, 20.5]	11.2 [7.6, 14.7]	12.8 [9.4, 16.4]
Qwen 27B	low	847/490	14.6 [11.4, 17.9]	10.7 [8.0, 13.6]	10.2 [7.6, 12.9]
Qwen 27B	high	840/490	28.3 [24.5, 32.5]	22.4 [18.9, 26.5]	22.3 [19.2, 25.9]
Qwen 35B	low	877/497	17.7 [14.0, 21.3]	12.0 [8.7, 15.1]	12.8 [9.8, 15.7]
Qwen 35B	high	879/498	17.6 [14.0, 21.5]	11.1 [8.1, 14.3]	12.4 [9.4, 15.5]
Both auditors valid (364 examples before output filtering)					
Gemma 31B	low	727/364	23.5 [19.1, 27.9]	10.7 [7.8, 13.8]	13.0 [10.2, 16.0]
Gemma 31B	high	660/344	24.7 [20.2, 29.1]	20.0 [15.4, 24.5]	23.0 [19.2, 26.9]
DeepSeek	low	728/364	8.2 [4.8, 12.1]	6.5 [3.6, 9.8]	6.6 [3.6, 9.9]
DeepSeek	high	618/329	23.8 [18.9, 28.7]	16.2 [11.7, 20.8]	19.0 [15.0, 23.1]
Nemotron S	low	727/364	6.1 [2.1, 10.0]	4.7 [1.2, 8.1]	4.5 [1.3, 7.7]
Nemotron S	high	677/363	24.1 [19.8, 28.4]	18.0 [13.6, 22.4]	19.8 [16.2, 23.8]
Nemotron N	low	623/356	2.1 [0.0, 4.2]	1.1 [−0.5, 2.9]	0.4 [−1.5, 2.4]
Nemotron N	high	500/325	19.0 [14.4, 23.8]	14.2 [10.1, 18.9]	15.8 [11.6, 20.2]
Qwen 27B	low	623/356	12.8 [9.4, 16.6]	9.1 [6.1, 12.4]	9.2 [6.4, 12.3]
Qwen 27B	high	619/357	27.5 [22.6, 32.0]	20.8 [16.3, 25.2]	22.0 [18.1, 26.1]
Qwen 35B	low	646/363	17.2 [12.7, 21.6]	11.5 [7.6, 15.4]	12.8 [9.0, 16.4]
Qwen 35B	high	646/363	16.7 [12.4, 21.2]	10.7 [7.0, 14.4]	12.1 [8.7, 15.6]

Cells give estimates [pointwise 95% cluster-bootstrap intervals]. Containment and EM differences are accuracy percentage points; token-F1 differences are F1 points, with scores multiplied by 100. n/u denotes valid matched sets / distinct examples; the same sets are used for all three metrics. Low denotes the disabled-or-minimal setting. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 32: Scoring sensitivity of the penalty gap Δ_{gap} on MuSiQue, pooling 25K and 112K contexts.

Model	Mode	n/u	Containment	Normalized EM	Token F1
Complete (500 examples before output filtering)					
Gemma 31B	low	998/500	20.0 [15.0, 25.0]	7.8 [4.2, 11.4]	8.4 [5.0, 11.7]
Gemma 31B	high	892/466	24.1 [19.5, 29.1]	15.8 [11.0, 20.7]	18.5 [14.4, 22.7]
DeepSeek	low	1000/500	4.3 [−0.7, 9.0]	3.4 [−0.5, 7.4]	2.9 [−1.2, 7.0]
DeepSeek	high	834/444	25.5 [19.9, 30.8]	17.5 [12.2, 22.7]	20.6 [16.4, 24.8]
Nemotron S	low	999/500	0.7 [−4.3, 5.7]	0.6 [−3.7, 4.8]	0.5 [−3.3, 4.4]
Nemotron S	high	916/496	25.0 [19.6, 30.5]	15.6 [10.5, 20.7]	18.4 [14.0, 22.9]
Nemotron N	low	842/487	0.7 [−3.0, 4.5]	−1.1 [−3.9, 1.7]	−0.1 [−3.2, 3.0]
Nemotron N	high	689/445	16.5 [10.1, 23.0]	10.6 [5.0, 16.2]	12.1 [6.6, 17.5]
Qwen 27B	low	847/490	10.0 [5.3, 15.0]	9.1 [4.7, 13.5]	6.5 [2.5, 10.7]
Qwen 27B	high	840/490	24.4 [19.5, 29.8]	21.1 [16.3, 26.3]	19.6 [15.5, 24.0]
Qwen 35B	low	877/497	16.1 [10.4, 21.6]	11.1 [5.9, 16.0]	11.8 [6.9, 16.3]
Qwen 35B	high	879/498	16.3 [10.6, 22.2]	10.1 [5.2, 15.2]	11.0 [6.1, 16.1]
Both auditors valid (364 examples before output filtering)					
Gemma 31B	low	727/364	17.7 [11.5, 24.5]	5.6 [1.6, 9.9]	6.9 [2.9, 11.0]
Gemma 31B	high	660/344	22.6 [17.3, 27.9]	13.6 [7.6, 19.4]	17.4 [12.3, 22.2]
DeepSeek	low	728/364	−1.2 [−6.5, 4.7]	0.8 [−3.8, 5.8]	−0.8 [−5.5, 4.1]
DeepSeek	high	618/329	21.5 [15.4, 27.6]	15.4 [9.6, 21.6]	18.4 [13.6, 23.2]
Nemotron S	low	727/364	−2.2 [−8.0, 4.0]	−0.6 [−5.5, 4.5]	−1.2 [−5.6, 3.7]
Nemotron S	high	677/363	23.3 [17.3, 29.7]	16.8 [10.6, 22.9]	19.5 [14.7, 24.7]
Nemotron N	low	623/356	−0.3 [−4.7, 3.9]	−2.6 [−5.9, 0.6]	−1.1 [−4.7, 2.5]
Nemotron N	high	500/325	20.4 [13.3, 28.1]	14.2 [7.6, 21.4]	16.0 [9.5, 22.7]
Qwen 27B	low	623/356	8.7 [3.3, 14.3]	7.7 [2.9, 12.8]	5.9 [1.3, 10.5]
Qwen 27B	high	619/357	24.4 [18.5, 30.0]	19.4 [13.8, 24.8]	20.0 [15.2, 24.8]
Qwen 35B	low	646/363	15.0 [8.3, 21.7]	10.1 [4.3, 16.0]	11.2 [5.5, 16.8]
Qwen 35B	high	646/363	13.6 [7.1, 20.3]	8.2 [2.3, 14.2]	9.2 [3.6, 14.8]

Cells give estimates [pointwise 95% cluster-bootstrap intervals]. Containment and EM differences are accuracy percentage points; token-F1 differences are F1 points, with scores multiplied by 100. n/u denotes valid matched sets / distinct examples; the same sets are used for all three metrics. Low denotes the disabled-or-minimal setting. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 33: Scoring sensitivity of the hint increment over teacher labels.

Model	n/u	Containment	Normalized EM	Token F1
Complete (500 examples before output filtering)				
Gemma 31B	301/150	24.6 [16.4, 33.1]	15.3 [10.1, 21.0]	22.6 [17.7, 27.8]
DeepSeek	538/192	35.7 [28.9, 42.3]	29.0 [22.4, 35.2]	31.6 [25.4, 37.6]
Nemotron S	383/191	28.2 [21.8, 34.7]	21.9 [15.6, 28.3]	23.9 [17.8, 30.1]
Nemotron N	563/242	26.8 [21.8, 31.8]	21.8 [17.5, 26.5]	25.0 [20.3, 29.7]
Qwen 27B	371/158	28.0 [20.8, 35.6]	26.4 [19.3, 33.5]	28.4 [22.3, 34.8]
Qwen 35B	45/40	4.4 [−14.3, 23.9]	−6.7 [−23.4, 8.9]	1.7 [−14.6, 17.2]
Both auditors valid (364 examples before output filtering)				
Gemma 31B	227/113	25.1 [14.9, 35.8]	13.2 [7.6, 19.5]	21.0 [15.2, 26.8]
DeepSeek	430/151	39.3 [32.5, 45.8]	30.7 [24.1, 37.3]	33.6 [27.5, 39.5]
Nemotron S	285/140	30.5 [22.8, 39.0]	24.2 [16.5, 32.4]	26.3 [19.2, 33.9]
Nemotron N	471/194	26.5 [21.0, 32.3]	22.1 [17.1, 27.3]	24.9 [19.8, 30.5]
Qwen 27B	303/121	32.0 [23.7, 40.7]	29.4 [21.6, 37.3]	31.1 [24.1, 38.7]
Qwen 35B	35/31	8.6 [−10.8, 27.8]	0.0 [−17.7, 16.7]	6.7 [−10.2, 23.1]

Cells give estimates [pointwise 95% cluster-bootstrap intervals]. Containment and EM differences are accuracy percentage points; token-F1 differences are F1 points, with scores multiplied by 100. n/u denotes valid matched sets / distinct examples; the same sets are used for all three metrics. Original low-wrong/high-correct selection and accepted hints remain frozen; cases are not reselected under another scoring rule. Results pool lengths and source conditions on the matched accepted-hint set. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 34: Scoring sensitivity of the oracle-label increment over teacher labels.

Model	n/u	Containment	Normalized EM	Token F1
Complete (500 examples before output filtering)				
Gemma 31B	301/150	−0.3 [−5.6, 4.8]	−0.3 [−1.0, 0.0]	0.2 [−1.1, 1.3]
DeepSeek	538/192	−2.6 [−6.4, 1.4]	−1.5 [−5.4, 2.2]	−1.0 [−4.5, 2.5]
Nemotron S	383/191	−1.0 [−4.9, 2.7]	−0.3 [−3.6, 3.0]	−0.1 [−3.0, 2.7]
Nemotron N	563/242	−2.5 [−5.1, −0.2]	−1.1 [−2.9, 0.7]	−1.8 [−3.9, 0.1]
Qwen 27B	371/158	0.3 [−3.6, 4.1]	0.8 [−2.7, 3.8]	0.9 [−2.4, 3.7]
Qwen 35B	45/40	0.0 [−13.0, 14.3]	−6.7 [−16.3, 2.3]	−2.7 [−13.2, 7.9]
Both auditors valid (364 examples before output filtering)				
Gemma 31B	227/113	0.9 [−4.0, 5.5]	0.0 [0.0, 0.0]	0.6 [−0.2, 1.5]
DeepSeek	430/151	−1.6 [−5.5, 2.1]	−0.7 [−4.6, 3.5]	−0.6 [−4.2, 3.0]
Nemotron S	285/140	−0.4 [−4.6, 4.0]	0.4 [−3.1, 3.9]	0.6 [−2.7, 4.0]
Nemotron N	471/194	−1.5 [−4.1, 1.0]	−0.8 [−2.7, 1.0]	−1.2 [−3.3, 0.6]
Qwen 27B	303/121	2.0 [−2.4, 6.0]	1.7 [−2.1, 5.4]	1.9 [−1.6, 5.0]
Qwen 35B	35/31	0.0 [−17.1, 18.2]	−8.6 [−21.2, 3.0]	−3.5 [−17.0, 10.1]

Cells give estimates [pointwise 95% cluster-bootstrap intervals]. Containment and EM differences are accuracy percentage points; token-F1 differences are F1 points, with scores multiplied by 100. n/u denotes valid matched sets / distinct examples; the same sets are used for all three metrics. Original low-wrong/high-correct selection and accepted hints remain frozen; cases are not reselected under another scoring rule. Results pool lengths and source conditions on the matched accepted-hint set. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 35: Scoring sensitivity of the gold-only high-minus-low difference.

Model	n/u	Containment	Normalized EM	Token F1
Complete (500 examples before output filtering)				
Gemma 31B	490/490	$4.7\,[1.4, 8.0]$	$-7.1\,[-10.8, -3.5]$	$-7.6\,[-10.8, -4.5]$
DeepSeek	476/476	$11.6\,[7.6, 15.5]$	$7.1\,[2.9, 11.6]$	$10.8\,[7.2, 14.3]$
Nemotron S	498/498	$8.8\,[5.0, 12.7]$	$5.2\,[1.6, 9.0]$	$7.2\,[3.9, 10.5]$
Nemotron N	497/497	$27.4\,[22.9, 32.0]$	$25.8\,[21.3, 30.0]$	$26.5\,[22.8, 30.6]$
Qwen 27B	500/500	$11.0\,[7.4, 14.6]$	$8.0\,[4.6, 11.2]$	$10.7\,[7.8, 13.7]$
Qwen 35B	500/500	$2.6\,[0.2, 5.0]$	$1.4\,[-0.8, 3.6]$	$2.1\,[0.1, 4.1]$
Both auditors valid (364 examples before output filtering)				
Gemma 31B	360/360	$5.3\,[1.7, 9.2]$	$-7.2\,[-11.7, -2.5]$	$-7.4\,[-11.2, -3.5]$
DeepSeek	354/354	$11.9\,[7.3, 16.4]$	$5.4\,[0.3, 10.2]$	$10.4\,[6.2, 14.6]$
Nemotron S	362/362	$11.6\,[7.2, 16.0]$	$6.9\,[2.8, 11.3]$	$9.5\,[5.7, 13.3]$
Nemotron N	362/362	$27.9\,[22.7, 32.9]$	$25.7\,[20.4, 30.9]$	$26.8\,[22.0, 31.4]$
Qwen 27B	364/364	$12.1\,[8.2, 16.2]$	$7.7\,[3.8, 11.5]$	$11.6\,[8.2, 15.2]$
Qwen 35B	364/364	$2.2\,[-0.5, 4.9]$	$1.1\,[-1.4, 3.6]$	$1.9\,[-0.5, 4.2]$

Cells give estimates [pointwise 95% cluster-bootstrap intervals]. Containment and EM differences are accuracy percentage points; token-F1 differences are F1 points, with scores multiplied by 100. n/u denotes valid matched sets / distinct examples; the same sets are used for all three metrics. The gold-only reference uses a separate population-wide set from the selected Stage-2 population. The Gemma difference reverses sign under EM and F1. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 36: Scoring sensitivity of the 2WikiMultiHopQA targeted penalty.

Model	n/u	Containment	Normalized EM	Token F1
Complete (120 examples before output filtering)				
DeepSeek	120/120	$22.5\,[15.0, 30.0]$	$19.2\,[12.5, 25.8]$	$19.2\,[12.6, 26.1]$
Nemotron S	120/120	$26.7\,[19.2, 35.0]$	$17.5\,[10.8, 24.2]$	$18.9\,[12.0, 26.1]$
Qwen 27B	118/118	$16.1\,[9.3, 22.9]$	$13.6\,[6.8, 21.2]$	$14.5\,[8.4, 21.3]$
Both auditors valid (112 examples before output filtering)				
DeepSeek	112/112	$23.2\,[16.1, 31.2]$	$19.6\,[12.5, 27.7]$	$19.7\,[12.8, 27.0]$
Nemotron S	112/112	$25.9\,[17.9, 33.9]$	$17.9\,[10.7, 25.0]$	$18.8\,[11.9, 26.1]$
Qwen 27B	110/110	$16.4\,[9.1, 23.6]$	$13.6\,[6.4, 20.9]$	$14.6\,[8.3, 21.8]$

Cells give estimates [pointwise 95% cluster-bootstrap intervals]. Containment and EM differences are accuracy percentage points; token-F1 differences are F1 points, with scores multiplied by 100. n/u denotes valid matched sets / distinct examples; the same sets are used for all three metrics. The replication compares hard and targeted contexts at approximately 25K tokens in the low setting. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

H.2 AUDIT-SELECTION SENSITIVITY

We construct six populations solely from the saved overall audit labels, before loading evaluated-model answers. Besides complete and consensus, we report each auditor's valid set, all cases with different overall labels, and cases with identical non-valid labels. In MuSiQue these contain respectively 500, 364, 478 (Qwen-valid), 368 (GPT-5.6-valid), 128 (disagreement), and eight (agreement-nonvalid) examples. The corresponding 2Wiki sizes are 120, 112, 116, 112, six, and two. Consensus, disagreement, and agreement-nonvalid partition the complete set; single-auditor valid sets overlap. In particular, 2Wiki GPT-5.6-valid and consensus have identical membership.

Disagreement refers to different three-way labels (valid, uncertain, invalid), not simply exclusion from consensus. MuSiQue's 128 disagreements include 118 cases with exactly one valid label and ten uncertain/invalid pairs; 2Wiki's six contain four and two, respectively. All agreement-nonvalid examples in these inputs are invalid/invalid. Passage-audit labels select source questions across all three conditions and their Stage-2 interventions. Hint acceptance uses the separate checks in Appendix C.2.

For each population, we retain the original containment scorer and matching requirements, then use 2,000 whole-example bootstrap draws with seed 20260919. The complete and consensus point estimates and denominators reproduce the canonical analyses. Non-OK responses remain missing. Population comparisons are descriptive; no between-population significance test is performed.

Complete, consensus, Qwen-valid, and GPT-5.6-valid preserve the same nine positive Δ_{gap} intervals and the same five positive hint-increment intervals. Main targeted-penalty estimates are positive in all 12 cells under each rule, with 12 positive intervals in complete/Qwen-valid and 11 in consensus/GPT-5.6-valid. Low/minimal Nemotron Nano is the boundary case in the latter two. All three 2Wiki penalties retain positive intervals in these four populations.

Disagreement cases support the broad main pattern but have different exceptions. Their targeted penalties are positive in all 12 cells, with 11 positive intervals. Nine Δ_{gap} intervals are positive, but low/minimal DeepSeek replaces high Nemotron Nano relative to consensus. Effect sizes also vary: DeepSeek-low's targeted penalty is 22.66 accuracy percentage points [16.02, 30.08] on disagreement versus 8.24 [4.67, 11.95] on consensus.

Only four disagreement-set hint increments have positive intervals. Qwen 27B is +10.29 points [-2.47, 21.74] on 68 matched cases from 37 examples; Qwen 35B-A3B is -11.11 [-60.00, 50.00] on nine cases from eight examples. The six-example 2Wiki disagreement subset is likewise inconclusive: all three penalties have positive point estimates but intervals including zero. The eight/two agreement-nonvalid groups are descriptive only; some diagnostic cells are empty or contain one or two independent examples. Blank cells denote unavailable estimates, and degenerate small-sample intervals should not be read as precise effects.

Tables 15 and 16 report the primary consensus and complete-population estimates. Tables 37 and 38 give the other MuSiQue audit groups; Tables 39 and 40 report Stage-2 hint increments and 2Wiki results. Full membership lists, audit-run bootstrap intervals for the baseline populations, and the dataset-specific three-by-three audit cross-tabs are retained in the saved audit-selection summaries.

Table 37: Audit-selection sensitivity of MuSiQue penalties, pooling 25K and 112K contexts.

Model	Mode	n/u	Δ_{diffuse}	Δ_{targeted}	Δ_{gap}
Qwen auditor valid (478 examples before output filtering)					
Gemma 31B	low	954/478	5.6 [2.8, 8.5]	25.7 [21.7, 29.5]	20.1 [14.6, 25.4]
Gemma 31B	high	860/449	2.4 [0.5, 4.5]	26.9 [22.8, 30.9]	24.4 [19.4, 29.1]
DeepSeek	low	956/478	8.1 [5.3, 10.7]	11.5 [8.2, 14.9]	3.5 [−1.4, 8.4]
DeepSeek	high	806/429	2.2 [0.2, 4.4]	27.7 [23.3, 32.1]	25.4 [19.9, 30.9]
Nemotron S	low	955/478	7.7 [4.8, 10.7]	8.7 [5.2, 12.1]	0.9 [−4.1, 6.0]
Nemotron S	high	877/474	1.1 [−1.5, 3.8]	26.0 [22.0, 30.1]	24.9 [19.0, 30.2]
Nemotron N	low	810/467	2.3 [−0.2, 5.0]	2.2 [0.5, 4.1]	−0.1 [−3.7, 3.5]
Nemotron N	high	654/425	0.0 [−3.8, 3.7]	16.8 [12.8, 20.6]	16.8 [10.4, 23.1]
Qwen 27B	low	813/468	4.6 [1.5, 7.5]	14.8 [11.5, 18.2]	10.2 [5.2, 15.7]
Qwen 27B	high	807/468	4.0 [1.6, 6.4]	28.9 [24.5, 32.9]	24.9 [19.5, 30.2]
Qwen 35B	low	840/475	1.9 [−1.4, 5.3]	17.6 [13.9, 21.3]	15.7 [9.9, 21.3]
Qwen 35B	high	842/476	1.5 [−2.0, 4.9]	17.8 [14.0, 21.5]	16.3 [10.1, 22.3]
GPT-5.6 auditor valid (368 examples before output filtering)					
Gemma 31B	low	735/368	5.7 [2.4, 9.0]	23.7 [19.3, 28.2]	18.0 [12.0, 24.2]
Gemma 31B	high	666/347	2.1 [−0.2, 4.5]	24.8 [20.5, 29.4]	22.7 [17.2, 28.4]
DeepSeek	low	736/368	9.2 [6.2, 12.5]	8.4 [4.8, 12.2]	−0.8 [−6.2, 4.8]
DeepSeek	high	621/331	2.3 [−0.2, 4.7]	23.7 [18.8, 28.9]	21.4 [15.3, 27.7]
Nemotron S	low	735/368	8.6 [5.3, 12.1]	6.0 [2.2, 10.2]	−2.6 [−8.4, 3.4]
Nemotron S	high	685/367	0.4 [−2.7, 3.6]	24.2 [19.8, 28.8]	23.8 [17.6, 30.2]
Nemotron N	low	631/360	2.4 [−0.5, 5.3]	2.1 [0.0, 4.3]	−0.3 [−4.5, 3.9]
Nemotron N	high	505/328	−1.4 [−5.4, 2.6]	18.8 [14.5, 23.4]	20.2 [13.1, 27.5]
Qwen 27B	low	628/360	4.3 [1.3, 7.4]	12.9 [9.3, 16.6]	8.6 [3.2, 14.0]
Qwen 27B	high	624/361	3.0 [0.7, 5.6]	27.4 [22.5, 32.2]	24.4 [18.6, 30.3]
Qwen 35B	low	653/367	2.0 [−1.7, 5.5]	17.6 [13.2, 22.1]	15.6 [9.1, 22.7]
Qwen 35B	high	653/367	2.9 [−1.1, 6.5]	17.2 [12.9, 21.5]	14.2 [7.5, 20.7]

Cells give accuracy differences in percentage points [pointwise 95% cluster-bootstrap intervals], using the original containment scorer. n/u denotes valid matched sets / distinct examples. Audit groups overlap; membership follows frozen passage-audit labels. Low denotes the disabled-or-minimal setting. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 38: Audit-selection sensitivity of MuSiQue penalties, pooling 25K and 112K contexts.

Model	Mode	n/u	Δ_{diffuse}	Δ_{targeted}	Δ_{gap}
Auditor disagreement (128 examples before output filtering)					
Gemma 31B	low	255/128	3.5 [−1.2, 8.3]	30.2 [23.4, 37.6]	26.7 [17.6, 36.3]
Gemma 31B	high	221/116	2.7 [−1.4, 7.1]	32.1 [24.0, 40.5]	29.4 [18.9, 39.4]
DeepSeek	low	256/128	2.7 [−2.3, 7.8]	22.7 [16.0, 30.1]	19.9 [10.5, 30.1]
DeepSeek	high	204/109	2.0 [−2.4, 6.5]	39.2 [30.8, 48.3]	37.3 [26.4, 48.0]
Nemotron S	low	256/128	7.4 [1.6, 13.3]	16.4 [10.2, 22.3]	9.0 [−0.8, 18.4]
Nemotron S	high	226/125	1.3 [−3.4, 6.8]	31.4 [23.2, 39.3]	30.1 [18.6, 40.4]
Nemotron N	low	207/123	0.5 [−4.7, 5.6]	3.9 [0.0, 7.8]	3.4 [−3.8, 11.1]
Nemotron N	high	175/113	4.0 [−2.9, 11.2]	8.6 [1.7, 16.2]	4.6 [−7.0, 17.2]
Qwen 27B	low	211/126	6.2 [−1.0, 13.1]	20.4 [13.3, 27.6]	14.2 [2.9, 25.9]
Qwen 27B	high	208/125	6.7 [1.0, 12.3]	32.2 [24.2, 40.8]	25.5 [13.7, 36.7]
Qwen 35B	low	217/126	0.0 [−6.8, 6.8]	19.8 [13.4, 26.8]	19.8 [8.9, 31.2]
Qwen 35B	high	219/127	−3.7 [−10.8, 3.6]	21.5 [14.4, 28.5]	25.1 [12.8, 37.1]
Auditors agree, non-valid (8 examples before output filtering)					
Gemma 31B	low	16/8	0.0 [0.0, 0.0]	18.8 [6.2, 37.5]	18.8 [6.2, 37.5]
Gemma 31B	high	11/6	0.0 [0.0, 0.0]	9.1 [0.0, 27.3]	9.1 [0.0, 27.3]
DeepSeek	low	16/8	0.0 [0.0, 0.0]	6.2 [0.0, 18.8]	6.2 [0.0, 18.8]
DeepSeek	high	12/6	8.3 [0.0, 25.0]	41.7 [8.3, 75.0]	33.3 [0.0, 66.7]
Nemotron S	low	16/8	6.2 [0.0, 18.8]	6.2 [−18.8, 37.5]	0.0 [−37.5, 37.5]
Nemotron S	high	13/8	0.0 [0.0, 0.0]	23.1 [0.0, 53.3]	23.1 [0.0, 53.3]
Nemotron N	low	12/8	−8.3 [−23.1, 0.0]	0.0 [0.0, 0.0]	8.3 [0.0, 23.1]
Nemotron N	high	14/7	−7.1 [−21.4, 0.0]	21.4 [0.0, 50.0]	28.6 [0.0, 71.4]
Qwen 27B	low	13/8	0.0 [0.0, 0.0]	7.7 [0.0, 27.3]	7.7 [0.0, 27.3]
Qwen 27B	high	13/8	0.0 [0.0, 0.0]	7.7 [0.0, 27.3]	7.7 [0.0, 27.3]
Qwen 35B	low	14/8	0.0 [0.0, 0.0]	7.1 [0.0, 25.0]	7.1 [0.0, 25.0]
Qwen 35B	high	14/8	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]

Cells give accuracy differences in percentage points [pointwise 95% cluster-bootstrap intervals], using the original containment scorer. n/u denotes valid matched sets / distinct examples. Disagreement means unequal overall decisions, including uncertain versus invalid; identical non-valid decisions form a separate stratum. Small strata are descriptive: intervals can be wide or degenerate. A dash denotes no valid matched cases, not a zero effect. Low denotes the disabled-or-minimal setting. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

Table 39: Audit-selection sensitivity of the Stage-2 hint increment.

	Consensus (364 questions)		Complete (500 questions)	
Model	n/u	Hint increment (pp)	n/u	Hint increment (pp)
Gemma 31B	227/113	25.1 [15.3, 35.3]	301/150	24.6 [15.8, 32.8]
DeepSeek V4 Flash	430/151	39.3 [32.4, 45.6]	538/192	35.7 [29.3, 41.9]
Nemotron Super	285/140	30.5 [22.7, 38.2]	383/191	28.2 [22.2, 34.5]
Nemotron Nano	471/194	26.5 [21.2, 32.0]	563/242	26.8 [21.8, 32.0]
Qwen 27B	303/121	32.0 [23.5, 40.6]	371/158	28.0 [20.8, 35.6]
Qwen 35B-A3B	35/31	8.6 [−9.7, 28.6]	45/40	4.4 [−14.9, 23.3]
	Qwen-valid (478 questions)		GPT-5.6-valid (368 questions)	
Model	n/u	Hint increment (pp)	n/u	Hint increment (pp)
Gemma 31B	298/147	24.5 [16.0, 33.5]	228/114	25.0 [14.9, 35.0]
DeepSeek V4 Flash	531/190	35.8 [29.0, 42.2]	434/152	39.4 [32.3, 46.1]
Nemotron Super	374/186	28.3 [21.8, 34.8]	293/144	30.4 [22.8, 37.9]
Nemotron Nano	552/236	26.1 [21.2, 30.8]	475/196	26.9 [21.6, 32.7]
Qwen 27B	371/158	28.0 [20.8, 35.8]	303/121	32.0 [23.4, 40.6]
Qwen 35B-A3B	44/39	4.5 [−14.3, 24.4]	35/31	8.6 [−11.4, 27.8]
	Disagreement (128 questions)		Agreement-nonvalid (8 questions)	
Model	n/u	Hint increment (pp)	n/u	Hint increment (pp)
Gemma 31B	73/36	23.3 [7.8, 40.0]	1/1	0.0 [0.0, 0.0]
DeepSeek V4 Flash	105/40	21.9 [5.4, 37.5]	3/1	0.0 [0.0, 0.0]
Nemotron Super	98/51	21.4 [11.1, 32.1]	0/0	—
Nemotron Nano	88/46	27.3 [16.7, 38.4]	4/2	50.0 [0.0, 100.0]
Qwen 27B	68/37	10.3 [−2.5, 21.7]	0/0	—
Qwen 35B-A3B	9/8	−11.1 [−60.0, 50.0]	1/1	0.0 [0.0, 0.0]

Increments are teacher+hint minus teacher-label accuracy in percentage points, with pointwise 95% example-clustered bootstrap intervals; n/u gives matched context instances / distinct questions. Populations follow passage-audit labels, with question counts before output filtering. Disagreement means unequal valid/uncertain/invalid decisions; agreement-nonvalid contains identical non-valid decisions. Original Stage-2 selection and hint acceptance remain fixed. Dashes indicate no matched cases. The last group includes only zero to two independent questions per model. Intervals use 2,000 question-cluster bootstrap draws with seed 20260919.

Table 40: Audit-selection sensitivity of the 2WikiMultiHopQA targeted penalty.

Model	n/u	Δ_{targeted}
Complete (120 examples before output filtering)		
DeepSeek	120/120	22.5 [15.0, 30.0]
Nemotron S	120/120	26.7 [19.2, 35.0]
Qwen 27B	118/118	16.1 [9.3, 23.7]
Both auditors valid (112 examples before output filtering)		
DeepSeek	112/112	23.2 [15.2, 31.2]
Nemotron S	112/112	25.9 [18.8, 34.8]
Qwen 27B	110/110	16.4 [9.1, 23.6]
Qwen auditor valid (116 examples before output filtering)		
DeepSeek	116/116	23.3 [15.5, 31.0]
Nemotron S	116/116	25.9 [18.1, 34.5]
Qwen 27B	114/114	16.7 [9.6, 24.6]
GPT-5.6 auditor valid (112 examples before output filtering)		
DeepSeek	112/112	23.2 [16.1, 31.2]
Nemotron S	112/112	25.9 [18.7, 33.9]
Qwen 27B	110/110	16.4 [10.0, 23.6]
Auditor disagreement (6 examples before output filtering)		
DeepSeek	6/6	16.7 [0.0, 50.0]
Nemotron S	6/6	33.3 [0.0, 66.7]
Qwen 27B	6/6	16.7 [0.0, 50.0]
Auditors agree, non-valid (2 examples before output filtering)		
DeepSeek	2/2	0.0 [0.0, 0.0]
Nemotron S	2/2	50.0 [0.0, 100.0]
Qwen 27B	2/2	0.0 [0.0, 0.0]

Cells give accuracy differences in percentage points [pointwise 95% cluster-bootstrap intervals], using the original containment scorer. n/u denotes valid matched sets / distinct examples. Audit groups overlap; membership follows frozen passage-audit labels. Disagreement means unequal overall decisions, including uncertain versus invalid; identical non-valid decisions form a separate stratum. Small strata are descriptive: intervals can be wide or degenerate. A dash denotes no valid matched cases, not a zero effect. All cells use approximately 25K tokens and the low setting. Consensus and GPT-5.6-valid contain the same 112 examples; their different bootstrap endpoints reflect resampling, not separate replications. Intervals use 2,000 example-cluster resamples (seed 20260919); displayed values are rounded to one decimal. Positive-interval counts use unrounded endpoints. Non-OK outputs are excluded. DeepSeek denotes V4 Flash; Nemotron S/N denote Super/Nano; Qwen 35B denotes 35B-A3B.

I Reproducibility and Research Artifacts

Experimental and analysis code, model configurations, frozen scored responses, audit records, and statistical results are provided in a supplementary ZIP archive uploaded to OpenReview with the submission. The archive includes a README describing the file layout and commands for reproducing the analyses. Table 41 lists the main entry points, with paths relative to the archive's `long_context_qa/` directory. JSON analysis summaries record SHA-256 hashes for input and output files. Statistical analyses reuse saved outputs; the semantic-audit runner documents the earlier model-based screening procedure.

Table 41: Main entry points for reproducing the reported analyses.

Paper component	Input or analysis entry point	Saved outputs or documentation
Populations and audits	`analysis/manifests/semantic_analysis_populations_v1.json`	`data/processed/semantic_audit/`
Stage-1 context effects	`results/current_main_scores.jsonl`; `analysis/information_structure_analysis.py`	`results/paper/information_structure/`; `results/paper/semantic_populations/`
Stage 2 and gold-only	`analysis/evidence_diagnostic_audit.py`	`results/paper/evidence_diagnostics/`
Stage-2 post-hoc checks	`analysis/stage2_followup_analysis.py`	`results/paper/stage2_followup/`
Scoring sensitivity	`analysis/scoring_sensitivity.py`	`results/paper/scoring_sensitivity/`
Audit-selection sensitivity	`analysis/audit_selection_sensitivity.py`	`results/paper/audit_selection_sensitivity/`

The sensitivity summaries record bootstrap seeds, scope, and input/output hashes. Frozen scored records retain final answers, correctness and parsing status, run metadata, and provider usage fields. All 29 constructed input files, including the long contexts and Stage-2 interventions, are stored with lossless deduplication. The supplied restoration script reconstructs their original bytes and verifies each file's SHA-256 hash; the README gives the commands.