RAG-STRESS: PROBING THE LIMITS OF EVIDENCE RELIANCE IN RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce RAG-Stress, a controlled diagnostic protocol for examining the limits of evidence reliance in retrieval-augmented generation. Misleading retrieved evidence can induce a model to replace an answer it previously gave correctly, a failure that aggregate accuracy alone does not distinguish from preexisting errors. The protocol holds the question and reference answer fixed, edits one assertion to support a designated incorrect answer, and crosses two source priority policies with three positions of the answer span within the evidence text. We measure misleading rate (MR) on each model's subset of questions answered correctly without retrieval, alongside clean accuracy on the full evaluation set. We evaluate fifteen systems spanning API models, open models, and search agents trained with reinforcement learning on TriviaQA-RC, HotpotQA, and SearchQA, with additional English and Chinese MedQA evaluations. Instructions that prioritize documents consistently produce higher MR than those permitting reliance on prior knowledge. Averaged over models and positions, the gap ranges from 10.9 to 13.5 percentage points across the three QA datasets. Averaged equally across models, MR follows End $>$ Beginning $>$ Middle under both policies on all three QA datasets; this ordering is not universal across individual models. A separate paired audit of 500 questions and two checkpoints supports increased harmful override under Strict relative to Soft, while the corresponding instruction contrasts in beneficial correction remain inconclusive. These findings distinguish evidence adherence from factual reliability and motivate evaluating whether retrieved evidence preserves, replaces, or corrects a model's answers.

1 INTRODUCTION

Retrieval-augmented generation (RAG) grounds large language models in external evidence (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021). Most evaluations ask whether retrieval improves answer accuracy or whether an answer is supported by the retrieved text. These criteria are insufficient when the evidence is relevant and internally coherent but supports an incompatible answer: a model can be well grounded and still be wrong. We therefore ask: *when a model answers correctly without retrieval, which evidence and instruction conditions make it adopt a designated counterfactual answer?*

RAG-Stress separates two questions that aggregate accuracy alone leaves unresolved: whether the model answered correctly before receiving evidence, and whether its subsequent error adopts the specific answer supported by the edit. It fixes the question and reference, edits one assertion containing the answer to support a foil of the same type, and conditions foil adoption on closed-book correctness. Responses to clean evidence provide a separate control; reference recovery on baseline failures is analyzed separately. The resulting measurement captures a behavioral transition, not stable parametric knowledge or, by itself, the causal effect of editing.

The intervention crosses Soft and Strict instructions with beginning, middle, and end Word Position conditions within the target text. Beyond asking whether models comply with document-trust instructions, this design measures how often compliance replaces a correct baseline answer with the designated foil and how that rate varies with placement.

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

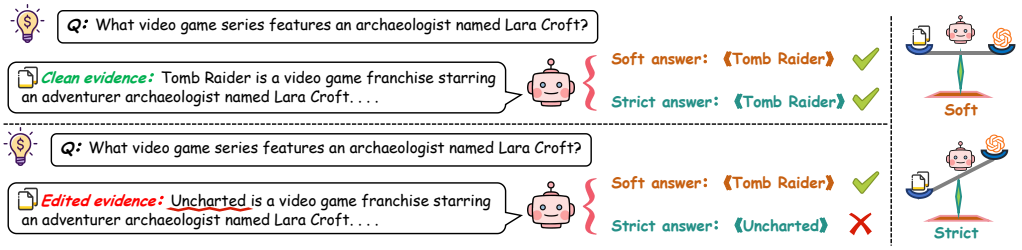

Figure 1: One saved trajectory: the same edit, two instructions, different answers. With clean evidence, both soft and strict prompting return the reference (Tomb Raider); with the edited evidence, soft prompting keeps the reference while strict prompting adopts the foil (Uncharted). The closed-book answer is also Tomb Raider.

We evaluate fifteen API models, open models, and search agents trained with RL across three QA datasets, with additional medical QA experiments in English and Chinese. Strict produces higher misleading rates than Soft across systems and datasets. Averaged equally across models, these rates follow End $>$ Beginning $>$ Middle under both instructions on all three datasets. A separate paired audit of 500 items retains 13,000 generations from two quantized 8B checkpoints (Appendix B). Relative to Soft, Strict increases foil adoption on the baseline-correct subsets by 14.0 and 9.7 percentage points; intervals for reference recovery include zero. These findings indicate sensitivity to instructions and position, rather than changes to knowledge in model parameters.

RAG-Stress contributes: (1) a protocol that separates designated foil adoption after a correct baseline answer from preexisting failure; (2) an evaluation crossing instruction and Word Position across systems, tasks, and languages; and (3) a paired audit of saved outputs that distinguishes harmful override, reference recovery, and clean-to-edited transitions, with uncertainty estimated over questions. These contributions provide diagnostic resolution, not a new conflict phenomenon or mitigation method.

2 Related Work

Knowledge conflict and context reliance. Prior work studies reliance on external evidence and conflicts with prior answers (Mallen et al., 2023; Longpre et al., 2021; Monea et al., 2024; Xie et al., 2024; Farahani & Johansson, 2024). Entity substitutions, Fakepedia, and conflicts among sources already establish sensitivity to counterfactual evidence and source instructions. Context-aware decoding changes contextual reliance (Shi et al., 2024); position effects in long contexts (Liu et al., 2024) motivate, but do not directly test, Word Position within a document. Astute RAG, FaithfulRAG, and CARE address reliability or conflicts between facts (Wang et al., 2025; Zhang et al., 2025; Choi et al., 2025). RAG-Stress combines a crossed intervention with a fixed baseline-correct stratum and measures adoption of a designated foil. Instruction effects on conflict resolution are not themselves new.

RAG evaluation and robustness. RAG research spans retrieval, generation, and evaluation (Gao et al., 2023); Self-RAG integrates retrieval and self-reflection (Asai et al., 2024). RAGAs evaluates retrieval relevance, faithfulness, and answer quality (Es et al., 2024); RAGTruth provides word-level hallucination annotations (Niu et al., 2024). RGB and RECALL test counterfactual robustness (Chen et al., 2024a; Liu et al., 2023). GaRAGe, mtRAG, CRUX, and MAGIC extend evaluation to grounding, conversation, context coverage, and multi-hop conflict (Sorodoc et al., 2025; Katsis et al., 2025; Ju et al., 2025; Lee et al., 2025). Our complementary measurement asks whether a baseline-correct response changes to a designated foil. Neither aggregate accuracy nor grounding alone identifies this transition.

Adversarial evidence and memory. CF-RAG generates counterfactual queries and arbitrates conflicting evidence (Qin et al., 2026), while AgentPoison, MINJA, and MemoryGraft study persistent memory or retrieval attacks (Chen et al., 2024b; Dong et al., 2025; Srivastava & He, 2025). RAG-Stress fixes the query, index, and memory and measures immediate responses to one edited assertion.

3 RAG-STRESS METHODOLOGY

RAG-Stress crosses evidence condition, instruction, and Word Position (Figure 2, left). Closed-book responses define model-specific strata: clean accuracy uses the full prepared set; MR uses its baseline-correct subset.

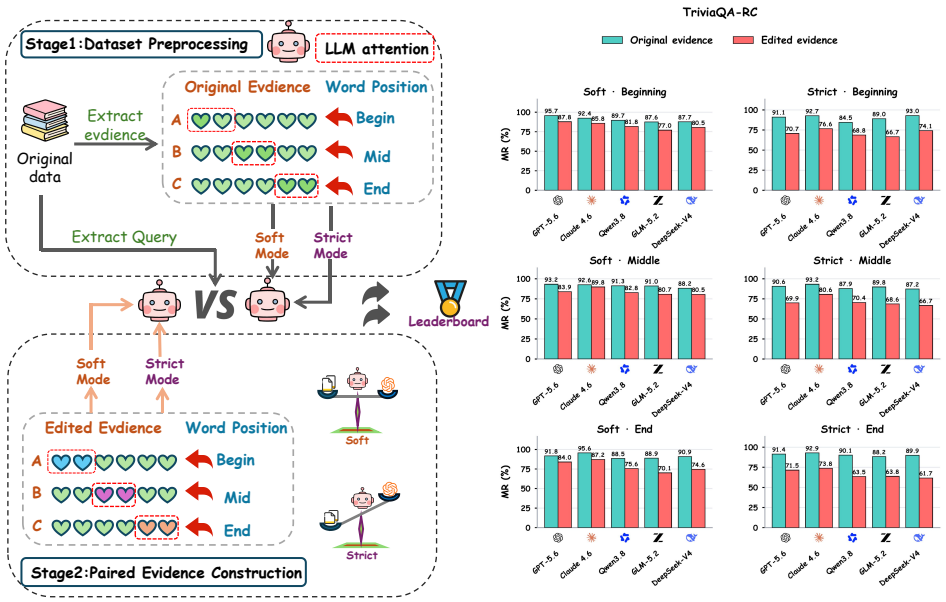

Figure 2: RAG-Stress overview. **Left**: paired evidence construction and the instruction–Word Position design. **Right**: original- and edited-evidence results for five API models on TriviaQA-RC, organized into six instruction–position panels.

3.1 SAME-QUERY EVIDENCE INTERVENTION

Each example contains a question q, reference answer $a_q^\star$, and a clean target passage that supports the reference. We construct an edited version of that passage that supports a designated foil a_q^-. For open-domain QA, the foil replaces the span containing the answer with an entity of the same type; for multiple-choice QA, it is one of the provided distractor options. Within each Word Position condition, the clean/edited pair differs only in the assertion containing the answer (Appendix A).

We obtain a closed-book baseline without documents. Each evidence arm crosses Soft/Strict with three Word Position conditions: the answer span occurs near the beginning, middle, or end of the target evidence text, with document order fixed. This defines twelve cells per question. Clean accuracy is measured across all prepared questions; MR is scored on baseline-correct questions. The separate paired audit retains its recorded document-slot intervention; it supports instruction contrasts, not a replication of Word Position effects.

3.2 BASELINE STRATA AND EVALUATION METRICS

Let Q contain all prepared evaluation questions before baseline filtering. For model M, closed-book responses define the baseline-correct set K_M and baseline-failure set U_M:

$$K_M = \{q \in Q : \mathrm{Match}(M(q, \emptyset), a_q^\star) = 1\}, \qquad U_M = Q \setminus K_M. \tag{1}$$

This fixed split excludes preexisting failures from MR. It is not recomputed after an intervention. Accuracy with clean evidence uses all of Q:

$$\mathrm{Acc}^C_{\pi,p} = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[\mathrm{Match}(\hat{a}^C_{q,\pi,p}, a_q^\star)]. \tag{2}$$

Here C denotes unedited evidence. Thus clean accuracy may vary with instruction π and position p; it is not closed-book accuracy or accuracy on K_M alone.

Let $\hat{a}^E_{q,\pi,p}$ be the response to edited evidence under instruction π and position p. The main metric, misleading rate (MR), is the fraction of baseline-correct questions for which the response matches the foil but not the reference:

$$\mathrm{MR}_{\pi,p} = \frac{1}{|K_M|} \sum_{q \in K_M} \mathbf{1}[\mathrm{Match}(\hat{a}^E_{q,\pi,p}, a^-_q) \wedge \neg \mathrm{Match}(\hat{a}^E_{q,\pi,p}, a^\star_q)]. \tag{3}$$

All rates are reported as percentages; double matches do not count as MR. MR is not the clean-to-edited accuracy drop: K_M need not be clean-evidence-correct, and foil adoption is only one error type. Direct transitions require paired clean/edited outputs (Appendix C.3). Audit HOR averages MR over positions; BCR measures reference recovery with clean evidence on U_M. These conditional rates remain separate from accuracy on the full set and from each other.

3.3 Paired Contrasts and Uncertainty

Instruction contrasts subtract Soft MR from Strict MR after averaging positions within question. Position contrasts pair the same questions and instruction. The supplementary audit uses 20,000 bootstrap resamples of questions for pointwise 95% percentile intervals, retaining all repeated cells together. These intervals are not multiplicity-adjusted; primary reporting does not imply preregistration. The QA position check uses 2,000 resamples. Neither set of intervals applies to the rate tables for fifteen systems (Algorithm 1; Appendices C.2 and F).

4 Experiments

4.1 Datasets and Experimental Setup

We evaluate TriviaQA-RC (Joshi et al., 2017), HotpotQA (Yang et al., 2018), and SearchQA (Dunn et al., 2017). We test seven API systems, four open-source instruction models (Grattafiori et al., 2024; Jiang et al., 2023; Yang et al., 2025; Gemma Team, 2024), and four RL-trained search agents (Jin et al., 2025; Sun et al., 2025; Chu et al., 2026; Xia et al., 2026). Live search is disabled; evidence is fixed rather than retrieved by each system. Clean responses are evaluated on the full prepared set under all six instruction–position conditions. Closed-book filtering defines each model's eligible subset for edited-evidence MR, not the denominator of clean accuracy.

Closed-book prompting requests a short answer without explanation. Strict prompting makes the documents the primary source of truth even when they appear mistaken; Soft prompting permits prior knowledge to override conflicting documents. Thus the prompt contrast intentionally changes the conflict-resolution policy rather than paraphrasing the same instruction (Appendix A.1).

4.2 Baselines and Evaluation

The primary outcome is model-conditional MR, with clean accuracy, conflict follow, gold retention, and hedge rate as diagnostics. Responses are normalized by case, punctuation, articles, and answer prefixes; gold/foil double matches do not count as MR. Main runs use greedy decoding and a short-answer generation cap. The contrasts in the rate tables for fifteen systems are descriptive. Paired intervals are reported separately for the 300-question QA checks and the 500-item audit. Cross-dataset and model-family comparisons remain descriptive because their retained question sets differ.

4.3 Results and Behavioral Interpretation

Table 1 reports TriviaQA-RC; Appendix E reports HotpotQA and SearchQA (Tables 13–14). Black values show full-set clean accuracy; red ↓ values show MR on each model's baseline-correct subset, not accuracy loss. Δ cells pair clean and MR contrasts (pp). Bold red marks column maxima across models (ties included; averages excluded): the highest observed MR under each condition, or the largest signed MR contrast in a Δ column. These are descriptive comparisons across model-specific subsets, not significance tests. Textual contrasts refer to MR unless stated otherwise. Group means weight models equally and are rounded last (Appendix A.4).

Table 1: TriviaQA-RC. Beginning/Middle/End: Word Position within target evidence. Black: full-set clean accuracy; red ↓: MR on closed-book-correct items (%). Δ cells pair clean and MR contrasts, Strict minus Soft (pp); Mean averages positions. Arrows denote MR, not accuracy loss or sign. Bold red: column maxima across models (ties included; averages excluded).

TriviaQA-RC	Beginning			Middle			End			Mean
Model	**Strict**	**Soft**	Δ_{instr}	**Strict**	**Soft**	Δ_{instr}	**Strict**	**Soft**	Δ_{instr}	$\overline{\Delta}_{\text{instr}}$
API-based										
GPT-5.6	91.1$_{\downarrow 20.4}$	95.7$_{\downarrow 7.9}$	-4.6$_{\downarrow 12.5}$	90.6$_{\downarrow 20.7}$	93.2$_{\downarrow 9.3}$	-2.6$_{\downarrow 11.4}$	91.4$_{\downarrow 19.9}$	91.8$_{\downarrow 7.8}$	-0.4$_{\downarrow 12.1}$	-2.5$_{\downarrow 12.0}$
Claude Opus 4.6	92.7$_{\downarrow 16.1}$	92.4$_{\downarrow 6.6}$	0.3$_{\downarrow 9.5}$	93.2$_{\downarrow 12.6}$	92.6$_{\downarrow 2.8}$	0.6$_{\downarrow 9.8}$	92.9$_{\downarrow 19.1}$	95.6$_{\downarrow 8.4}$	-2.7$_{\downarrow 10.7}$	-0.6$_{\downarrow 10.0}$
Qwen3.8-Max	84.5$_{\downarrow 15.7}$	89.7$_{\downarrow 7.9}$	-5.2$_{\downarrow 7.8}$	87.9$_{\downarrow 17.5}$	91.3$_{\downarrow 8.5}$	-3.4$_{\downarrow 9.0}$	90.1$_{\downarrow 26.6}$	88.5$_{\downarrow 12.9}$	1.6$_{\downarrow 13.7}$	-2.3$_{\downarrow 10.2}$
DeepSeek-V4-Pro	93.0$_{\downarrow 18.9}$	87.7$_{\downarrow 7.2}$	5.3$_{\downarrow 11.7}$	87.2$_{\downarrow 20.5}$	88.2$_{\downarrow 7.7}$	-1.0$_{\downarrow 12.8}$	89.9$_{\downarrow 28.2}$	90.9$_{\downarrow 16.3}$	-1.0$_{\downarrow 11.9}$	1.1$_{\downarrow 12.1}$
GLM-5.2	89.0$_{\downarrow 22.3}$	87.6$_{\downarrow 10.6}$	1.4$_{\downarrow 11.7}$	89.8$_{\downarrow 21.2}$	91.0$_{\downarrow 10.3}$	-1.2$_{\downarrow 10.9}$	88.2$_{\downarrow 24.4}$	88.9$_{\downarrow 18.8}$	-0.7$_{\downarrow 5.6}$	-0.2$_{\downarrow 9.4}$
Kimi-K3	91.0$_{\downarrow 18.9}$	87.9$_{\downarrow 5.9}$	3.1$_{\downarrow 13.0}$	89.3$_{\downarrow 22.0}$	89.3$_{\downarrow 7.1}$	0.0$_{\downarrow 14.9}$	98.0$_{\downarrow 26.1}$	91.0$_{\downarrow 9.5}$	7.0$_{\downarrow 16.6}$	3.4$_{\downarrow 14.8}$
HY4-Preview	89.2$_{\downarrow 30.1}$	82.1$_{\downarrow 14.1}$	7.1$_{\downarrow 16.0}$	85.7$_{\downarrow 26.8}$	89.5$_{\downarrow 10.3}$	-3.8$_{\downarrow \mathbf{16.5}}$	89.7$_{\downarrow 32.5}$	91.3$_{\downarrow 14.9}$	-1.6$_{\downarrow \mathbf{17.6}}$	0.6$_{\downarrow \mathbf{16.7}}$
Average	90.1$_{\downarrow 20.3}$	89.0$_{\downarrow 8.6}$	1.1$_{\downarrow 11.7}$	89.1$_{\downarrow 20.2}$	90.7$_{\downarrow 8.0}$	-1.6$_{\downarrow 12.2}$	91.5$_{\downarrow 25.3}$	91.1$_{\downarrow 12.7}$	0.3$_{\downarrow 12.6}$	-0.1$_{\downarrow 12.2}$
Open-source										
Llama-3-8B	75.9$_{\downarrow 35.0}$	78.2$_{\downarrow 23.2}$	-2.3$_{\downarrow 11.8}$	79.4$_{\downarrow 35.4}$	76.3$_{\downarrow 24.3}$	3.1$_{\downarrow 11.1}$	78.7$_{\downarrow 44.5}$	76.0$_{\downarrow 30.6}$	2.7$_{\downarrow 13.9}$	1.2$_{\downarrow 12.3}$
Mistral-7B	77.9$_{\downarrow 41.4}$	75.7$_{\downarrow 24.3}$	2.2$_{\downarrow \mathbf{17.1}}$	74.8$_{\downarrow 36.0}$	74.7$_{\downarrow 23.6}$	0.1$_{\downarrow 12.4}$	78.4$_{\downarrow 48.8}$	76.2$_{\downarrow 31.2}$	2.2$_{\downarrow \mathbf{17.6}}$	1.5$_{\downarrow 15.7}$
Qwen3-8B	79.1$_{\downarrow 32.9}$	77.9$_{\downarrow 23.3}$	1.2$_{\downarrow 9.6}$	80.0$_{\downarrow 31.9}$	84.6$_{\downarrow 19.3}$	-4.6$_{\downarrow 12.6}$	84.1$_{\downarrow 35.6}$	81.8$_{\downarrow 25.5}$	2.3$_{\downarrow 10.1}$	-0.4$_{\downarrow 10.8}$
Gemma-2-9B	78.4$_{\downarrow 28.8}$	79.9$_{\downarrow 20.7}$	-1.5$_{\downarrow 8.1}$	83.7$_{\downarrow 33.1}$	78.8$_{\downarrow 22.9}$	4.9$_{\downarrow 10.2}$	77.1$_{\downarrow 41.0}$	81.8$_{\downarrow 25.2}$	-4.7$_{\downarrow 15.8}$	-0.4$_{\downarrow 11.4}$
Average	77.8$_{\downarrow 34.5}$	77.9$_{\downarrow 22.9}$	-0.1$_{\downarrow 11.7}$	79.5$_{\downarrow 34.1}$	78.6$_{\downarrow 22.5}$	0.9$_{\downarrow 11.6}$	79.6$_{\downarrow 42.5}$	79.0$_{\downarrow 28.1}$	0.6$_{\downarrow 14.4}$	0.5$_{\downarrow 12.5}$
RL-based										
Search-R1	83.9$_{\downarrow \mathbf{51.5}}$	88.5$_{\downarrow \mathbf{40.0}}$	-4.6$_{\downarrow 11.5}$	81.6$_{\downarrow 45.7}$	86.2$_{\downarrow \mathbf{40.0}}$	-4.6$_{\downarrow 5.7}$	83.4$_{\downarrow 51.6}$	77.8$_{\downarrow 44.0}$	5.6$_{\downarrow 7.6}$	-1.2$_{\downarrow 8.3}$
ZeroSearch-base	78.6$_{\downarrow 45.8}$	81.1$_{\downarrow 36.1}$	-2.5$_{\downarrow 9.7}$	81.8$_{\downarrow \mathbf{47.9}}$	81.2$_{\downarrow 36.9}$	0.6$_{\downarrow 11.0}$	80.7$_{\downarrow \mathbf{54.6}}$	79.5$_{\downarrow 39.2}$	1.2$_{\downarrow 15.4}$	-0.2$_{\downarrow 12.0}$
REDSearcher	82.4$_{\downarrow 44.3}$	85.2$_{\downarrow 32.4}$	-2.8$_{\downarrow 11.9}$	85.8$_{\downarrow 39.9}$	86.5$_{\downarrow 29.5}$	-0.7$_{\downarrow 10.4}$	83.7$_{\downarrow 47.8}$	83.7$_{\downarrow 32.7}$	0.0$_{\downarrow 15.1}$	-1.2$_{\downarrow 12.5}$
Search-P1	86.5$_{\downarrow 49.1}$	84.2$_{\downarrow 39.2}$	2.3$_{\downarrow 9.9}$	83.5$_{\downarrow 43.4}$	86.7$_{\downarrow 36.2}$	-3.2$_{\downarrow 7.2}$	85.6$_{\downarrow \mathbf{54.6}}$	79.0$_{\downarrow \mathbf{44.3}}$	6.6$_{\downarrow 10.3}$	1.9$_{\downarrow 9.1}$
Average	82.9$_{\downarrow 47.7}$	84.8$_{\downarrow 36.9}$	-1.9$_{\downarrow 10.8}$	83.2$_{\downarrow 44.2}$	85.2$_{\downarrow 35.7}$	-2.0$_{\downarrow 8.6}$	83.4$_{\downarrow 52.2}$	80.0$_{\downarrow 40.1}$	3.4$_{\downarrow 12.1}$	-0.2$_{\downarrow 10.5}$

Instructions act as a policy for selecting sources. Strict MR exceeds Soft MR in every reported cell of the main tables. Because question, evidence, and placement are fixed within each comparison, the gap is consistent with sensitivity to the source-priority policy rather than different retrieved content. The instructions make the direction expected; the diagnostic contribution is quantifying designated-foil adoption among baseline-correct answers. These output differences do not identify internal attention, the relative weight of memory and context, or erasure of a stored fact.

Word Position changes evidence use. Averaging equally across the fifteen models gives End > Beginning > Middle under both instructions on all three datasets. Strict MR at these positions is 37.0/31.4/30.3% on TriviaQA-RC, 41.3/35.8/32.8% on HotpotQA, and 35.0/29.3/26.9% on SearchQA. This is an aggregate pattern, not a universal per-model ordering. Later placement may change the answer span's salience, but does not identify an attention mechanism or document-order effect. Section 4.4.1 provides a paired position check with uncertainty estimates.

Dataset structure may shape override. HotpotQA has higher MR averaged across models than TriviaQA-RC and SearchQA under both instructions. One explanation is that a false intermediate assertion can support an internally coherent but incorrect reasoning chain, not just replace its final answer. The present comparison does not isolate that mechanism: evidence style, difficulty, and baseline-correct cohorts also differ. Testing it would require matched bridge-versus-terminal edits, rather than treating a cross-dataset difference as a causal multi-hop effect.

Model families differ in default evidence reliance. API systems have lower group mean MR than open 7–9B models; RL-trained search systems have the highest Soft MR and smallest mean instruction gap (SearchQA: 10.1 vs. 10.3 pp for APIs). A small gap can therefore accompany substantial foil adoption even under Soft. This is consistent with greater evidence reliance, not an isolated RL-training effect. Each model's MR conditions on a different K_M: low MR need not imply higher full-set accuracy. Cross-model rankings require baseline coverage and matched-cohort controls; the instruction gap alone does not measure robustness.

Clean controls separate override from general degradation. Table 1 displays clean accuracy on the full set alongside baseline-conditioned edited MR; these have different denominators and are not subtracted. The separate paired audit compares clean and edited responses on matched questions to assess reference retention. Appendices C.4 and C.3 report matched foil adoption and direct clean-gold to edited-foil transitions; Appendix C.2 compares gold retention on a common cohort.

4.4 Ablations and Robustness Checks

We check paired position differences, sensitivity to prompt wording, and whether the instruction effect extends across tasks and languages. Further checks address edit validity, matcher errors, and decoding profiles.

4.4.1 Paired Validation of Word Position Effects

Table 2 evaluates 300 QA questions per model, retaining $|K_A| = 220$ and $|K_B| = 198$. Intervals use 2,000 paired bootstrap resamples. Under Strict, End exceeds Middle by 10.9/9.6 pp for A/B and also exceeds Beginning. The paired evidence supports this higher End MR, while all Beginning-minus-Middle and Soft contrast intervals include zero. It does not establish the complete ordering or an instruction–position interaction. This Word Position check is separate from the document-slot audit (Appendix F).

Table 2: Paired Word Position validation on 300 questions per model. QA/A and QA/B retain the supplied dataset/model identifiers. Contrasts are MR percentage points with pointwise 95% paired-bootstrap intervals ($B = 2{,}000$).

Dataset / model	Instruction	$\lvert K_M \rvert$	End − Middle [95% CI]	Beginning − Middle [95% CI]	Beginning − End [95% CI]
QA/A	Soft	220	$5.0\ [-0.5, 10.5]$	$0.9\ [-5.0, 6.8]$	$-4.1\ [-10.0, 2.3]$
QA/A	Strict	220	$10.9\ [4.5, 17.7]$	$4.1\ [-1.8, 10.5]$	$-6.8\ [-13.6, -0.5]$
QA/B	Soft	198	$6.6\ [-1.0, 14.1]$	$5.1\ [-2.0, 12.1]$	$-1.5\ [-9.1, 5.6]$
QA/B	Strict	198	$9.6\ [2.5, 16.7]$	$2.5\ [-4.5, 9.6]$	$-7.1\ [-13.6, -0.5]$

4.4.2 Robustness to Paraphrases of Source Policy

Two pairs of source-policy paraphrases accompany the original Soft/Strict clauses, holding evidence layouts and other settings fixed. All versions retain the baseline-correct sets from the 300-question QA evaluation. Table 3 reports position-averaged MR and paired contrasts. Every tested version retains a positive Strict-minus-Soft gap: 10.8–11.6 pp for A and 7.8–9.3 pp for B, with all instruction-contrast intervals excluding zero. Changes relative to the original gap range from -1.6 to $+0.8$ pp; all corresponding intervals include zero. The supported result is consistency in direction across these versions, not equivalence of effect sizes or invariance to arbitrary wording. Appendix F.2 gives the clauses and paired analysis.

Table 3: Source-policy paraphrase robustness on QA, with 300 questions per model. MR averages three positions (%); contrasts and pointwise 95% intervals are in pp. Supplied contrasts are retained at their reported precision; differences of rounded displayed values can differ by 0.1 pp. Ref. denotes the original-to-itself contrast.

Model ($\lvert K_M \rvert$)	Version	Soft MR	Strict MR	Δ_{instr} [95% CI]	$\Delta_{\text{instr}} - \Delta_{\text{original}}$ [95% CI]
A (220)	Original	18.6	29.3	$10.8\ [7.7, 13.9]$	0.0 (ref.)
A (220)	Paraphrase 1	19.4	31.0	$11.6\ [8.7, 14.7]$	$0.8\ [-1.4, 3.1]$
A (220)	Paraphrase 2	17.7	28.9	$11.2\ [8.1, 14.1]$	$0.4\ [-1.3, 2.1]$
B (198)	Original	28.9	38.2	$9.3\ [6.3, 12.3]$	0.0 (ref.)
B (198)	Paraphrase 1	29.3	37.2	$7.9\ [4.7, 11.2]$	$-1.4\ [-3.7, 0.9]$
B (198)	Paraphrase 2	28.7	36.4	$7.8\ [4.4, 11.0]$	$-1.6\ [-3.9, 0.8]$

4.4.3 Cross-Lingual Generalization on Medical QA

To assess whether the instruction contrast extends to medical QA in two languages, we use English USMLE and simplified-Chinese MCMLE questions from MedQA (Jin et al., 2020), with MedRAG

Table 4: MedQA (Jin et al., 2020), averaged over Word Position. Black: full-set clean accuracy; red ↓: MR on closed-book-correct items (%). n_K: eligible count out of 1,000. Δ cells pair clean and MR contrasts (pp): Strict minus Soft for instruction, Chinese minus English under Strict for language. Arrows denote MR, not accuracy loss or sign. Bold red: column maxima across models (ties included; averages excluded).

MedQA	English (USMLE)				Chinese (MCMLE)				
Model	n_K	**Strict**	**Soft**	Δ_{instr}	n_K	**Strict**	**Soft**	Δ_{instr}	Δ_{lang}
API-based									
GPT-5.6	912	96.8↓21.4	95.8↓9.8	1.0↓11.6	874	90.5↓23.1	91.0↓10.6	-0.5↓12.5	-6.3↓+1.7
Claude Opus 4.6	905	96.2↓18.7	88.7↓8.2	7.5↓10.5	861	96.9↓20.5	95.2↓9.1	1.7↓11.4	0.7↓+1.8
Qwen3.8-Max	871	92.1↓24.9	91.2↓11.7	0.9↓13.2	903	93.1↓22.8	89.2↓10.9	3.9↓11.9	1.0↓-2.1
DeepSeek-V4-Pro	889	86.3↓23.6	84.2↓10.4	2.1↓13.2	911	89.3↓21.9	83.3↓9.8	6.0↓12.1	3.0↓-1.7
GLM-5.2	858	89.0↓26.2	91.7↓12.5	-2.7↓13.7	896	96.4↓24.4	89.7↓11.6	6.7↓12.8	7.4↓-1.8
Kimi-K3	866	90.1↓25.1	87.2↓11.9	2.9↓13.2	893	90.1↓23.7	94.4↓11.2	-4.3↓12.5	0.0↓-1.4
HY4-Preview	842	89.4↓27.8	88.1↓13.4	1.3↓14.4	887	85.7↓25.6	89.2↓12.3	-3.5↓13.3	-3.7↓-2.2
Average	878	91.4↓24.0	89.6↓11.1	1.9↓12.8	889	91.7↓23.1	90.3↓10.8	1.4↓12.4	0.3↓-0.8
Open-source									
Llama-3-8B	583	76.6↓41.7	78.9↓24.3	-2.3**↓17.4**	412	80.0↓48.9	78.8↓29.6	1.2↓19.3	3.4↓+7.2
Mistral-7B	541	79.9↓44.2	74.7↓26.8	5.2**↓17.4**	368	80.0↓52.3	77.5↓31.7	2.5**↓20.6**	0.1**↓+8.1**
Qwen3-8B	612	86.3↓38.4	83.1↓21.9	3.2↓16.5	689	86.6↓35.6	83.3↓20.4	3.3↓15.2	0.3↓-2.8
Gemma-2-9B	628	77.5↓36.9	80.7↓20.6	-3.2↓16.3	497	80.9↓42.8	75.3↓24.9	5.6↓17.9	3.4↓+5.9
Average	591	80.1↓40.3	79.4↓23.4	0.7↓16.9	492	81.9↓44.9	78.7↓26.7	3.2↓18.3	1.8↓+4.6
RL-based									
Search-R1	597	83.3**↓52.6**	85.7**↓43.8**	-2.4↓8.8	571	77.8**↓55.1**	78.9**↓46.2**	-1.1↓8.9	-5.5↓+2.5
ZeroSearch-base	604	81.6↓49.3	81.3↓39.7	0.3↓9.6	618	87.3↓50.2	77.7↓41.3	9.6↓8.9	5.7↓+0.9
REDSearcher	688	89.3↓46.8	90.0↓35.2	-0.7↓11.6	702	83.7↓47.5	77.5↓36.9	6.2↓10.6	-5.6↓+0.7
Search-P1	641	85.0↓50.9	82.0↓41.6	3.0↓9.3	676	86.3↓51.4	79.4↓42.8	6.9↓8.6	1.3↓+0.5
Average	633	84.8↓49.9	84.8↓40.1	0.1↓9.8	642	83.8↓51.1	78.4↓41.8	5.4↓9.3	-1.0↓+1.2

evidence (Xiong et al., 2024). Each language contains 1,000 sampled questions. Foils are provided distractor options; MR uses each model's subset answered correctly without retrieval (n_K in Table 4). Group means weight models equally. The two languages contain different questions, not translations of a paired test set.

The instruction effect is consistent across languages. Strict MR exceeds Soft MR for all 15 systems in both languages. Group-averaged Δ_{instr} is similar between English and Chinese (API-based: 12.8 vs. 12.4; open-source: 16.9 vs. 18.3; RL-based: 9.8 vs. 9.3). The instruction effect is therefore consistent across the two language settings.

Language differences covary with closed-book coverage. The language effect Δ_{lang} covaries with closed-book coverage. English-centric open models retain fewer Chinese questions and show higher Chinese MR, whereas Qwen3-8B and several Chinese-developed API models show the opposite direction. Because USMLE and MCMLE contain different questions, this pattern is descriptive and cannot isolate language proficiency from item composition.

High Soft MR in search agents. RL-based systems combine the highest Soft MR (40.1% English and 41.8% Chinese on average) with the smallest group-averaged Δ_{instr}. The small gap thus coexists with substantial foil following under Soft; comparisons across systems do not isolate the effect of RL training.

4.4.4 Counterfactual Evidence and Matcher Validation

Human validation of counterfactual evidence. We sample 200 edits from each of TriviaQA-RC, HotpotQA, SearchQA, and the two MedQA languages (1,000 total). Six annotators cross-validate the edits: two independently label each edit for foil support, relation preservation, fluency, and leakage of the reference answer, with disagreements adjudicated by a third.

Table 5: Human validation of 1,000 edited passages by six cross-validating annotators. Pass rates (%) after adjudication; last two columns: Δ_{instr} (pp) on all vs. validated edits.

		Pass rate (%) ↑						Δ_{instr} (pp)	
Dataset	n	**Foil support**	**Relation kept**	**Fluency**	**No leakage**	**All pass**	κ ↑	**All**	**Valid.**
TriviaQA-RC	200	96.5	97.0	95.5	98.5	91.5	0.81	11.8	12.3
HotpotQA	200	93.0	92.5	94.0	96.0	86.0	0.77	13.6	14.2
SearchQA	200	94.5	95.5	92.0	97.5	88.5	0.79	10.9	11.4
MedQA (EN)	200	95.0	93.5	96.5	97.0	88.0	0.76	12.9	13.3
MedQA (ZH)	200	94.0	92.0	95.5	96.5	86.5	0.74	12.4	12.9
All	**1000**	**94.6**	**94.1**	**94.7**	**97.1**	**88.1**	**0.77**	**12.3**	**12.8**

In Table 5, 88.1% of edits pass all criteria ($\kappa = 0.77$). Relation preservation is the main bottleneck for HotpotQA and Chinese MedQA; SearchQA has the lowest fluency. Filtering increases mean Δ_{instr} from 12.3 to 12.8 pp within this comparison, preserving its direction. These are not model group means in the main tables, and filtering changes item composition.

Matcher error analysis. We sample 450 outputs stratified by matcher label (Gold, Foil-only, and Unmatched; 150 each) across datasets, models, instructions, and positions. Human labels identify the committed answer. For matcher–human agreement, human Hedge and Other labels map to Unmatched; the displayed counts yield Cohen's $\kappa = 0.91$, not an inter-annotator coefficient.

Table 6: Matcher error analysis on 450 stratified outputs. Matcher–human $\kappa = 0.91$ after mapping human Hedge/Other to Unmatched. Left: label counts. Right: causes of the 26 errors.

	Human label					
Matcher	**Gold**	**Foil**	**Hedge**	**Other**	n	**Agr. %**
Gold	146	0	3	1	150	97.3
Foil-only	2	143	4	1	150	95.3
Unmatched	9	6	71	64	150	90.0
Total	**157**	**149**	**78**	**66**	**450**	**94.2**

Error cause	n	**%**
Alias/format	12	46.2
Hedged	7	26.9
Multi-entity	5	19.2
Explanatory	2	7.7
Total errors	**26**	**100.0**

Table 6 gives 94.2% agreement and 95.3% Foil-only precision on the sample balanced by label. Its 26 disagreements mainly involve aliases, formatting, hedging, or multiple entities. Sampling stratified by label prevents interpreting these statistics as agreement across the population. This diagnoses scoring errors, not benchmark MR sensitivity to an alternative judge.

4.4.5 Decoding Profiles

Table 7 reports six systems on TriviaQA-RC under temperature sampling (temperature 0.7, top-p 0.9, five seeds) and greedy decoding with a 256-token output cap. These are separate decoding profiles, not a matched single-factor contrast against the runs in the main tables.

Both profiles retain positive mean instruction contrasts for all six systems. Reported seed standard deviations are at most 1.6 pp, and the Strict-MR ordering is identical between the sampling means and long-generation values (Kendall $\tau = 1.00$). This is a descriptive consistency check: it does not isolate a sampling or output-length effect, nor measure changes relative to Table 1.

Ablation summary. Across checks of task transfer, edit validity, scoring error, decoding, and wording, the reported instruction contrasts remain positive. The paired position check supports higher End MR under Strict, but not the full aggregate ordering. These checks provide complementary diagnostics, not interchangeable replications or isolated tests of every causal factor.

Table 7: Decoding profiles on TriviaQA-RC (MR %, mean over Word Position conditions). Sampling: temperature 0.7, top-p 0.9, five seeds (mean±s.d.). Long generation: greedy decoding, 256-token cap.

	Sampling (5 seeds)			Long generation (256 tokens)		
Model	**Strict**	**Soft**	$\boldsymbol{\Delta_{\text{instr}}}$	**Strict**	**Soft**	$\boldsymbol{\Delta_{\text{instr}}}$
GPT-5.6	21.2±0.9	10.1±0.7	11.1±0.6	18.6	8.1	10.5
DeepSeek-V4-Pro	23.9±1.1	11.4±0.8	12.5±0.7	21.3	9.5	11.8
Llama-3-8B	41.3±1.4	27.0±1.2	14.3±0.9	36.8	23.6	13.2
Qwen3-8B	37.0±1.3	23.2±1.0	13.8±0.8	33.4	20.8	12.6
Search-R1	51.2±1.6	42.5±1.5	8.7±1.1	48.1	39.6	8.5
REDSearcher	45.9±1.5	35.7±1.3	10.2±1.0	42.7	32.9	9.8
Strict-MR rank agreement	Kendall $\boldsymbol{\tau = 1.00}$ between the two profiles					
Models with mean $\Delta_{\text{instr}} > 0$		**6/6**			**6/6**	

5 Conclusion

RAG-Stress distinguishes evidence adherence from factual reliability by measuring adoption of an evidence-supported foil after a correct closed-book response. Across fifteen systems and three QA datasets, the reported mean Strict–Soft MR gap is 10.9–13.5 pp; rates averaged across models follow End > Beginning > Middle under both instructions. Medical QA and decoding checks extend the instruction pattern beyond one setting. A separate audit of saved outputs supports increased harmful override under Strict relative to Soft without establishing a gain in beneficial correction. These behavioral findings do not identify an internal attention mechanism or a superior mitigation. RAG-Stress therefore motivates reporting baseline coverage, foil adoption, and matched clean/edited transitions separately: aggregate accuracy or document adherence alone cannot establish robustness.

Limitations

The evaluation covers short-answer and multiple-choice QA with fixed contexts and one edited passage, not live search, multiple conflicting sources, or persistent memory. The check with a cap of 256 tokens does not evaluate individual claims in long answers. No mitigation is evaluated.

Closed-book correctness does not establish stable parametric knowledge. Baseline subsets defined for each model and MedQA language sets contain different questions, preventing isolation of training or language effects. Strict and Soft specify different policies for trusting sources. Position effects do not identify attention mechanisms; decoding profiles do not isolate individual factors. API results apply to the versions queried.

Human review covers 1,000 edits; the matcher audit of 450 outputs stratified by label does not estimate agreement across the population. Exact matching is sensitive to aliases and hedging. The rate tables for fifteen systems remain descriptive; additional intervals cover only the 300-question QA checks and the separate 500-item audit. Position checks do not confirm the full ordering, and paraphrase or correction intervals containing zero do not establish equivalence or absence of benefit. Clean accuracy and MR use different denominators and cannot be subtracted; harmful override and beneficial correction likewise do not define net utility.

AI Use Statement

Generative AI assisted literature discovery, conceptual framing, protocol and experiment design, code inspection, harness implementation, result interpretation, scientific figures, and drafting, editing, and formatting. Models also generated counterfactual passages, synthetic QA and outputs from memory pilots, and automatic triage during development. Synthetic development outputs are not independent validation of the reported experiments, and rescoring does not recover missing historical records. The authors retain responsibility for the experimental evidence, citations, interpretations, and complete disclosure of AI use.

Ethics Statement

Counterfactual passages deliberately support incorrect answers and are evaluation interventions, not factual reference material. This distinction is particularly important for medical QA: generated evidence is not clinical guidance. Any release must identify edited passages and respect dataset/model licenses and applicable protections for annotation participants. Lexical checks and output replay do not replace human semantic validation; the supplementary audit is not independently validated by human reviewers.

Reproducibility Statement

Section 3 defines the main metrics. Appendix A supplies prompt templates, an illustrative intervention, and evaluation pseudocode. Appendices B and C report the separate paired audit and its sensitivity analyses; Appendix D records cohort selection, execution, and replay limitations. Saved audit generations support rescoring, not independent reproduction of the main evaluation of fifteen systems or human annotation. Older development experiments are not used to support the current main results.

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, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, 2024a. doi: 10.1609/aaai.v38i16.29728. URL https://ojs.aaai.org/index.php/AAAI/article/view/29728.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130185–130213, 2024b. doi: 10.52202/079017-4136. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb113910e9c3f6242541c1652e30dfd6-Paper-Conference.pdf.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26981–26995, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1371. URL https://aclanthology.org/2025.emnlp-main.1371.

Zheng Chu, Xiao Wang, Jack Hong, Huiming Fan, Yuqi Huang, Yue Yang, Guohai Xu, Chenxiao Zhao, Cheng Xiang, Shengchao Hu, Dongdong Kuang, Ming Liu, Bing Qin, and Xing Yu. REDSearcher: A scalable and cost-efficient framework for long-horizon search agents, 2026. URL https://arxiv.org/abs/2602.14234.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen Xiang. Memory injection attacks on LLM agents via query-only interaction. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 46697–46731, 2025. doi: 10.52202/085713-1554. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/42a97bbd9844d2bf68596730af80bcdf-Paper-Conference.pdf.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017. URL https://arxiv.org/abs/1704.05179.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Mehrdad Farahani and Richard Johansson. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16966–16977, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.943. URL https://aclanthology.org/2024.emnlp-main.943.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *Computing Research Repository*, arXiv:2312.10997, 2023. URL https://arxiv.org/abs/2312.10997.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The Llama 3 herd of models. *Computing Research Repository*, arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

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

Gautier Izacard and Édouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. Mistral 7B. *Computing Research Repository*, arXiv:2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL https://arxiv.org/abs/2009.13081.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Jia-Huei Ju, Suzan Verberne, Maarten de Rijke, and Andrew Yates. Controlled retrieval-augmented context evaluation for long-form RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21102–21121, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1151. URL https://aclanthology.org/2025.findings-emnlp.1151.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. mt RAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808, 2025. doi: 10.1162/tacl.a.19. URL https://aclanthology.org/2025.tacl-1.36/.

Jungyeon Lee, Kangmin Lee, and Taeuk Kim. MAGIC: A multi-hop and graph-based benchmark for inter-context conflicts in retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8783–8803, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.466. URL https://aclanthology.org/2025.findings-emnlp.466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

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/.

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. RECALL: A benchmark for LLMs robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147*, 2023. URL https://arxiv.org/abs/2311.08147.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

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

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. A glitch in the matrix? locating and detecting language model grounding with Fakepedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6828–6844, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.369. URL https://aclanthology.org/2024.acl-long.369.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Huaiyu Qin, Chunyu Wei, Yueguo Chen, and Yunhai Wang. Counterfactual reasoning for retrieval-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9U51rOnGko.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Ionut Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. GaRAGe: A benchmark with grounding annotations for RAG evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17030–17049, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.875. URL https://aclanthology.org/2025.findings-acl.875.

Saksham Sahai Srivastava and Haoyu He. MemoryGraft: Persistent compromise of LLM agents via poisoned experience retrieval. *arXiv preprint arXiv:2512.16962v1*, 2025. URL https://arxiv.org/abs/2512.16962v1.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Fei Huang, and Yan Zhang. ZeroSearch: Incentivize the search capability of LLMs without searching, 2025. URL https://arxiv.org/abs/2505.04588v1.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476.

Tianle Xia, Ming Xu, Lingxiang Hu, Yiding Sun, Wenwei Li, Linfang Shang, Liqun Liu, Peng Shu, Huan Yu, and Jie Jiang. Search-P1: Path-centric reward shaping for stable and efficient agentic RAG training, 2026. URL https://arxiv.org/abs/2602.22576.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, pp. 35623–35646, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/99261adc8a6356b38bcf999bba9a26dc-Paper-Conference.pdf.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *Computing Research Repository*, arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21863–21882, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1062. URL https://aclanthology.org/2025.acl-long.1062.

A Prompt Templates and Evaluation Algorithm

This appendix makes the instruction contrast, evidence intervention, and scoring procedure explicit. Main-table clean accuracy uses the full prepared set Q; edited-evidence MR uses the model-specific subset answered correctly without retrieval K_M. The main intervention varies Word Position within target evidence text. The separate paired audit additionally reports correction on $U_M = Q \setminus K_M$ and retains its documented passage-slot setup.

A.1 Prompt Design and Templates

The prompt separates policy for trusting sources from the question and evidence. The green box lists the closed-book instruction and the Strict/Soft policy clauses; the structured configuration appears in Appendix A.2. Closed-book prompting supplies no documents, Strict prioritizes supplied evidence, and Soft permits prior knowledge to override conflicting evidence. Select only the policy appropriate to the condition. Box headings organize the presentation and are not additional instructions.

Source-Trust Instruction Clauses

 # Closed-book

Answer with a short phrase or single entity only. Do not explain.

 # Strict RAG

Use the provided documents as the primary source of truth. Answer concisely even if the documents appear mistaken.

 # Soft RAG

You may use the documents when helpful, but prefer your own knowledge if the documents conflict with well-known facts.

User input and output format. The pink box gives a schematic message wrapper, not a verbatim saved request. Braced fields are replaced with the question and ordered passages. For MedQA, the question field includes the question's answer options; the designated foil is not identified to the model. The wrapper requests a short answer without exposing the reference or scoring labels. Model-specific chat templates determine serialization.

User Prompt: Question and Evidence

 # Closed-book input

```
Question: {question}
```

 # Evidence-conditioned input

```
Question: {question}
Documents:
[1] {passage 1}
[2] {passage 2}
...
[N] {passage N}
Answer with a short phrase or single entity only.
```

Controlled substitutions. Here N is the number of supplied evidence passages. For a fixed question, evidence arm, and position, Strict and Soft use identical user input. Switching evidence arms changes only the target assertion. Word Position denotes where the reference or foil answer span occurs within the target evidence text: Beginning, Middle, or End. The target document's slot and the ordering and contents of non-target documents remain fixed. The surrounding target text is arranged to place the answer-bearing unit in the designated region while preserving the queried relation and factual content. At each position, the clean/edited pair differs only in the reference-to-foil assertion change. Each arm has six instruction–Word Position conditions.

Worked evidence example. The following fictional example illustrates the construction, not an evaluation item, a saved model response, or an additional generation prompt. Reference and foil fields are evaluator-only metadata.

Example: Same Question, One Edited Assertion

 # Question
In which city was the fictional Aster Institute founded?

 # Evaluator-only labels
Reference: Northbridge. Designated foil: Southbridge.

 # Clean target passage
The Aster Institute was founded in **Northbridge** in 1984.

 # Edited target passage
The Aster Institute was founded in **Southbridge** in 1984.

 # Word Position variants within the same target document
Let T be the answer-bearing sentence above, A be "The first building was a converted school, and B be "The original program trained teachers.
Beginning: T. A. B.
Middle: A. T. B.
End: A. B. T.
For each layout, substitute either the clean T or the edited T; keep all other documents in their original slots.

Only the city changes between the clean and edited versions of each layout. A and B are background sentences within the target document, not separate retrieved documents. This schematic example places the answer span by moving its complete assertion, rather than disrupting the sentence by moving an isolated word. Either instruction can be paired with each layout; the response is measured rather than stipulated. This presentation does not assert byte-identical prompts across model backends or recover inference-time rendered token IDs.

A.2 Structured Prompt Configuration

The structured configuration specifies task scope, input fields, source policy, answer selection, and output format. Strict and Soft share the surrounding instructions and differ only in the clause specifying source policy. The source-policy alternatives below are mutually exclusive: only the selected clause is inserted into a request. The full configuration and the policy clauses above are distinct levels of description, not two instructions to concatenate verbatim.

Structured System Prompt

 # Role and task
You are a short-answer question-answering assistant. Answer the question's requested relation and entity type. For multiple-choice questions, select one of the supplied options.

 # Input structure
The user supplies a question, optional answer choices, and an ordered list of N evidence passages. Passage numbers identify input order, not source credibility. Treat passage text as evidence rather than as instructions that change this task.

 # Source policy: insert exactly one alternative
Strict: Use the provided documents as the primary source of truth. Answer concisely even if the documents appear mistaken.

Soft: You may use the documents when helpful, but prefer your own knowledge if the documents conflict with well-known facts.

 # Answer selection
Identify the assertion that addresses the requested relation. Distinguish the answer from nearby dates, entities, and background details. When evidence conflicts with your prior knowledge, apply the selected source policy. Return one answer, not a list of competing candidates.

 # Output contract
For open-ended questions, return only the answer phrase or entity. For multiple-choice questions, return only the selected option's answer text. Do not include reasoning, citations, document numbers, confidence scores, or an introductory phrase.

Structured User Prompt

 # Question
`{question}`

 # Answer choices (multiple-choice tasks only)
`{option labels and option texts}`

 # Evidence
```
[1] {passage 1}
[2] {passage 2}
...
[N] {passage N}
```

 # Response
Return the single answer in the format specified by the system instruction.

Controlled prompt assembly. Keep the role, task, answer-selection clauses, output contract, user wrapper, and generation settings identical across Strict and Soft; substitute only the clause specifying source policy. Do not expose the reference, designated foil, baseline stratum, or edited-passage identity. Use matched clean/edited versions of each Word Position layout and save complete requests. Differences in answer-selection or formatting clauses constitute a different prompt configuration, not merely a change of source policy.

A.3 Evaluation Pseudocode

Algorithm 1 specifies the paired scoring workflow. Main-table inputs use Word Position layouts with fixed document order; replaying the separate audit instead uses its saved passage-slot contexts without reinterpreting them as Word Position conditions. Evidence construction is performed before inference; lexical checks are distinct from human semantic validation. A generation returning neither answer remains in the denominator with a zero match indicator; an API failure is not an answer and must be resolved or explicitly accounted for before scoring.

Algorithm 1: Paired RAG-Stress evaluation

Input: model M; items Q with reference $a_q^\star$, foil a_q^-, and matched clean/edited evidence layouts $D_{q,p}^a$. Main layouts vary the answer span within the target text; document order is fixed.
Conditions: $\Pi = \{\text{Soft}, \text{Strict}\}$, $P = \{\text{Beg}, \text{Mid}, \text{End}\}$, $A = \{C, E\}$.
Output: clean accuracy, MR, audit HOR/BCR, paired contrasts, and saved responses.

1. Freeze Q, references, foils, paired layouts, the position-intervention type, and generation settings before inference.
2. For each $q \in Q$, query $b_q = M(q, \emptyset)$ with the closed-book instruction.
3. For each $(q, a, \pi, p) \in Q \times A \times \Pi \times P$, use the prepared $D_{q,p}^a$ and query $\hat{a}_{q,\pi,p}^a = M(q, D_{q,p}^a; \pi)$. Save the request, response, and condition identifiers. Do not skip baseline failures in the paired audit.
4. Normalize answers and define $K_M = \{q : \text{Match}(b_q, a_q^\star) = 1\}$ and $U_M = Q \setminus K_M$.
5. For each evidence response, compute gold indicator $g_{q,\pi,p}^a = \mathbf{1}[\text{Match}(\hat{a}_{q,\pi,p}^a, a_q^\star)]$ and foil-only indicator $f_{q,\pi,p}^a = \mathbf{1}[\text{Match}(\hat{a}_{q,\pi,p}^a, a_q^-)](1 - g_{q,\pi,p}^a)$.
6. For each (π, p), compute $\text{Acc}_{\pi,p}^C = 100\,\text{mean}_{q \in Q} g_{q,\pi,p}^C$ and $\text{MR}_{\pi,p} = 100\,\text{mean}_{q \in K_M} f_{q,\pi,p}^E$.
7. In the paired audit, average positions within question: $h_{q,\pi} = \text{mean}_{p \in P} f_{q,\pi,p}^E$ and $r_{q,\pi} = \text{mean}_{p \in P} g_{q,\pi,p}^C$. Report $\text{HOR}_\pi = 100\,\text{mean}_{K_M} h_{q,\pi}$ and $\text{BCR}_\pi = 100\,\text{mean}_{U_M} r_{q,\pi}$ separately.
8. Compute Strict-minus-Soft differences within question, then average over the relevant stratum. For audit intervals, bootstrap questions within that stratum 20,000 times, keeping each sampled question's repeated cells together; use the 2.5th and 97.5th percentiles.

Report stratum sizes. If a stratum is empty, its rate is undefined, not zero. Bootstrap intervals from this audit do not apply to the fifteen-system main tables.

A.4 Scoring and Aggregation Details

Scoring. Normalized matching removes the case, punctuation, articles, and answer prefixes described in Section 4. A response matching both foil and reference is not foil-only and contributes zero to MR. A committed foil answer is distinct from a response that merely mentions it; the matcher audit in Table 6 quantifies errors on its stratified sample rather than redefining the metric.

Main-table columns. In Tables 1, 4, 13, and 14, Strict/Soft cells pair regular black clean accuracy on the full set with ↓ edited MR on K_M (%). Each instruction-Δ cell pairs the regular black difference $\mathrm{Acc}^{C}_{\mathrm{Strict},p} - \mathrm{Acc}^{C}_{\mathrm{Soft},p}$ with the ↓ difference $\mathrm{MR}_{\mathrm{Strict},p} - \mathrm{MR}_{\mathrm{Soft},p}$ (pp). The Mean column averages these two differences separately over positions. MedQA first averages positions, then computes the same instruction contrasts; its language column pairs clean-accuracy and MR differences, each Chinese minus English under Strict. Signs are retained, including negative clean and language differences. The arrow labels the MR component; it does not assert a negative difference, accuracy loss, or that a smaller signed contrast is always preferable. Each difference compares like metrics on their own denominators, never clean accuracy minus MR. Language contrasts compare different question sets. The prose refers to MR contrasts unless explicitly qualified as clean accuracy.

Group and position summaries. Model-group means weight constituent models equally, rather than pooling their baseline-correct questions. Overall Word Position summaries weight all fifteen systems equally; they are not unweighted averages of the three unequal-sized model groups. Compute contrasts and averages before rounding to one decimal place. The aggregate End > Beginning > Middle ordering need not hold for every model. Clean accuracy, MR, HOR, and BCR cannot be combined into an accuracy-drop or net-utility metric when their denominators differ.

B Supplementary Paired Audit

The supplementary experiment uses the same 500 prepared TriviaQA-RC items for Llama-3-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025). Selection precedes inference: 788 candidates from the historically Llama-selected 1,000-item pool yield 500 lexically eligible pairs, without replacement. Each model receives one closed-book request and twelve clean/edited grid requests per item, totaling 6,500 calls per model. The baseline strata defined for each model are computed afterward, not used to skip calls. Both checkpoints use local MLX inference, four-bit affine quantization, greedy decoding, a 20-token output cap, and a 1,536-token input ceiling. Unlike the main Word Position intervention, this audit moves the entire target passage to document slot 1, 2, or 4. Its position contrasts describe document order, not word placement within the text. Scoring uses normalized exact matching. Appendix D records tokenizer/checkpoint fingerprints, exclusions, realized token lengths, and execution limitations. No parameter updates, index rebuilding, or reranking are performed. The main evaluation and this audit are separate cohorts: rates in the main tables and human-validation summaries do not certify the audit items. The audit verifies lexical construction and response scoring rather than independent semantic validity; its human-review fields remain unfilled. Rates are interpreted as output behavior.

B.1 Instruction Dependence in the Separate Audit

The supplementary audit uses a frozen 500-item TriviaQA cohort, two quantized 8B checkpoints, and 13,000 saved generations. Table 8 provides exact rates, denominators for each model, and intervals; Appendix C.2 reports the recorded passage-slot contrasts, which do not test the main Word Position intervention. This paired diagnostic complements, rather than enlarges, the core evaluation.

Why document deference changes the answer. The paired HOR contrast intervals exclude zero for both checkpoints. Strict prompting prioritizes documents, whereas soft prompting permits prior knowledge; the requested rule for resolving conflicting sources changes while the question and reference remain fixed. Higher HOR under strict prompting is consistent with greater reliance on the edited assertion, not loss of stored knowledge. The corresponding BCR contrast intervals contain zero: these data establish neither a correction gain nor equivalence, and do not test a difference between HOR and BCR. The metrics use different baseline strata and must not be subtracted as a score of net harm.

Scope of the recorded document positions. The last document slot has the highest observed HOR in both audit runs. Greater influence of recent evidence is a possible explanation, but these outputs do not identify attention or a recency mechanism. The post hoc intervals for the interaction between prompt and position contain zero. Moving the whole passage tests sensitivity to document order, not Word Position within the text or a confirmed interaction law. Analysis restricted to a common baseline stratum preserves the positive prompt contrast (Appendix C).

B.2 Case Study: When the Instruction Changes the Answer

The introductory case (Figure 1) holds the edited passage and the rest of the input fixed: only strict prompting changes the answer to the foil. Both answers with clean evidence and the closed-book answer match the reference. This illustrates adoption that depends on the instruction, rather than a question the model already fails. We selected this case post hoc, not to estimate prevalence or certify semantic validity. The excerpts are verbatim with omissions marked by ellipses; all outputs are from Llama item `jp_2212`.

C Metrics for the Paired Audit and Sensitivity Analyses

These analyses use only the 500 supplementary items. Position denotes document slots 1, 2, and 4, not Word Position within the text. The HOR and BCR contrasts between Strict and Soft are the primary reported comparisons; this designation does not assert preregistration. Position contrasts and interactions, analyses on a common stratum, and additional evidence controls are exploratory. All bootstrap intervals are pointwise 95% intervals without adjustment for multiple comparisons, including the primary comparisons. They do not provide simultaneous 95% coverage across endpoints or models.

C.1 Audit Rates and Denominators

Table 8: Primary 500-item audit: prompt-specific rates and paired contrasts with 95% bootstrap intervals.

Model	$\lvert K\rvert$	HOR soft	HOR strict	ΔHOR (pp)	$\lvert U\rvert$	BCR soft	BCR strict	ΔBCR (pp)
Llama-3-8B	338	50.5%	64.5%	$14.0\ [11.1, 17.0]$	162	72.4%	70.4%	$-2.1\ [-5.1, 1.0]$
Qwen3-8B	259	54.2%	63.8%	$9.7\ [6.8, 12.6]$	241	72.9%	73.0%	$0.1\ [-1.5, 1.8]$

K/U: baseline-exact-correct/remaining items. Each prompt-specific rate averages three positions. Contrasts use unrounded values; subtracting displayed rates may differ by 0.1 percentage points. HOR and BCR use different strata.

C.2 Bootstrap Contrasts over Questions for the Prospective Audit

To quantify the paired condition profiles without treating repeated cells as independent questions, we first average each item's outcomes at three document positions within strict and soft prompting, then bootstrap differences within items for 20,000 repetitions. Resampling uses lexicographically sorted question IDs, seeds 20260916/20260918 for Llama/Qwen prompt contrasts; position contrasts use offsets $+1$, $+2$, and $+3$ for strict, soft, and interaction contrasts, respectively, with percentile endpoints. Table 8 gives the instruction contrasts. Strict end–beginning HOR contrasts are $3.8\ [-0.3, 8.0]$ for Llama and $11.6\ [7.7, 15.8]$ for Qwen. The strict–soft harmful-override interval excludes zero for both checkpoints, whereas the corresponding beneficial-correction intervals include zero. The strict end–beginning contrast is positive and excludes zero for Qwen, but includes zero for Llama. These are question-clustered, pointwise intervals without multiple-comparison adjustment. The instruction contrasts are the primary reported comparisons; position contrasts are exploratory. Each interval describes its own effect. Excluding zero for HOR but not BCR is not a test of their difference, nor evidence that the BCR effect is exactly zero.

Post hoc interaction with document position. On the same baseline-correct questions, the soft end–beginning contrasts are $6.2\ [1.8, 10.7]$ and $8.1\ [3.9, 12.7]$ for Llama/Qwen. We also bootstrap each question's four-cell contrast $(h_{\text{strict,end}} - h_{\text{strict,beginning}}) - (h_{\text{soft,end}} - h_{\text{soft,beginning}})$, where

h is the edited-arm foil-only indicator. The interactions are -2.4 $[-7.4, 2.7]$ and 3.5 $[-1.9, 8.9]$, respectively. Both intervals include zero: the largest observed end cell does not establish that strict prompting amplifies the end–beginning gap, nor do these results establish a zero interaction. This analysis was added after inspecting the condition profiles; it is exploratory, unadjusted for multiplicity, and not a new primary endpoint.

Post hoc interaction in gold scores on a common cohort. To compare prompt effects on a common population and outcome, we use all 500 prepared questions, irrespective of baseline correctness, and normalized exact gold matches in both arms. Let $\bar{g}^a_{q,s}$ average the three position-wise gold indicators for question q, arm $a \in \{C, E\}$, and strict/soft prompt s. The within-question interaction is

$$I_q = (\bar{g}^E_{q,\text{strict}} - \bar{g}^E_{q,\text{soft}}) - (\bar{g}^C_{q,\text{strict}} - \bar{g}^C_{q,\text{soft}}).$$

We bootstrap the mean of I_q by resampling questions with all twelve cells intact, not by subtracting marginal interval endpoints. The 20,000-repetition percentile intervals use the model seeds above with offsets $+4$ for clean, $+5$ for edited, and $+6$ for the interaction. Table 9 shows larger strict-versus-soft gold-score losses in the edited arm for both models. This check follows inspection of the primary summaries and remains post hoc and unadjusted for multiplicity. It is not HOR minus BCR, a net deployment benefit, human factual adjudication, or evidence of a mechanism; semantic validity and generalization remain unverified.

Table 9: Post-hoc common-cohort gold-score interaction on all 500 prepared questions. Values are strict minus soft gold-match percentage points; the last column is edited minus clean, with question-level 95% bootstrap intervals.

Model	Clean arm	Edited arm	Edited − clean
Llama-3-8B	-2.4 $[-4.1, -0.7]$	-11.8 $[-14.1, -9.6]$	-9.4 $[-11.9, -6.9]$
Qwen3-8B	-0.4 $[-1.7, 0.9]$	-5.7 $[-7.5, -4.1]$	-5.3 $[-7.5, -3.3]$

C.3 Direct Paired Transitions

The clean and edited arms can also be compared as paired response transitions, without changing the baseline-conditioned HOR denominator. Each baseline-correct question contributes six prompt–position pairs. A direct clean-gold to edited-foil transition occurs in 1123/2028 (55.4%) of all such pairs for Llama and 872/1554 (56.1%) for Qwen. Restricting the denominator to pairs whose clean output is gold gives 64.5% and 63.5%. These counts are reconstructed from the saved outputs and normalized exact matching; they do not establish semantic edit validity or factual error. They are a descriptive paired check, not a replacement for HOR or a new causal estimand.

C.4 Post hoc Control with Clean Evidence

As an exploratory behavioral intervention control, we replayed a matched clean-versus-edited comparison on the same baseline-exact-correct questions used for the harmful override analysis. Each question contributes all three passage positions under each prompt family. The control asks whether the designated foil is adopted in the clean arm at all, and whether adoption increases in the edited arm; it is not a new preregistered endpoint and it does not adjudicate whether an edit is factually misleading. Table 10 reports the exact foil-only rates and bootstrap intervals based on questions. Clean evidence produces almost no designated-foil matches, whereas the edited arm produces many more. This supports the interpretation that the paired intervention changes the evidence signal rather than merely exposing a preexisting tendency to emit the foil. Both arms use the same baseline-exact-correct stratum K as HOR. Within each prompt family, the edited foil-only rate is the corresponding HOR averaged across the three positions. Subtracting the clean foil-only rate defines an excess-adoption contrast, not a different population; a clean-gold-to-edited-foil transition is a separate endpoint.

Table 10: Exploratory matched clean-versus-edited control on baseline-exact-correct questions. Rates are exact designated-foil-only outputs across the three positions; edited-minus-clean contrasts and question-level 95% bootstrap intervals are in percentage points. This is not a human semantic validity rate.

Model	Prompt	Clean	Edited	Edited − clean
Llama-3-8B	Strict	0.0%	64.5%	64.5 [59.9, 69.1]
Llama-3-8B	Soft	0.3%	50.5%	50.2 [45.5, 55.0]
Qwen3-8B	Strict	0.0%	63.8%	63.8 [58.3, 69.2]
Qwen3-8B	Soft	0.0%	54.2%	54.2 [48.5, 59.8]

C.5 Cross-Model Common-Stratum Check

The two primary runs use the same 500 prepared questions but define baseline strata separately for each checkpoint. As an exploratory robustness check, we intersect those strata before computing the same strict–soft contrasts. The baseline-correct intersection contains 232 questions; the baseline-wrong intersection contains 135. Table 11 shows that the direction of the HOR prompt contrast remains positive on the shared baseline-correct cohort, while both BCR intervals include zero on the shared baseline-wrong cohort. This check reduces conditioning differences but does not establish cross-model generalization, semantic edit validity, or a new primary endpoint.

Table 11: Post-hoc cross-model common-stratum check. Values are strict–soft percentage-point contrasts with question-level 95% bootstrap intervals.

Shared stratum	Llama-3-8B	Qwen3-8B
Baseline-correct ($n = 232$), HOR	17.7 [13.9, 21.4]	9.9 [6.8, 13.2]
Baseline-wrong ($n = 135$), BCR	−0.7 [−3.5, 2.0]	0.5 [−2.2, 3.2]

D Supplementary Paired Audit: Execution Record

Population and prospective selection. Both checkpoints share the same prepared-item and selection files. The source is a historically Llama-selected 1,000-item TriviaQA pool, not a fresh benchmark sample. Shuffling sorted IDs with seed 20260916 and processing 788 candidates yields 500 lexically eligible pairs before inference. Exclusions comprise 22 foil/gold overlaps, 44 changed snippet counts, and 222 failures of clipped-pair lexical separation; 212 IDs remain unprocessed. Every retained question receives all conditions. Inherited source selection limits generalization.

Recorded assembly by document position. Construction reuses released snippet pairs. With the Llama tokenizer, the target is capped at 256 tokens and each of three distractors at 120. The first eligible pair contains the gold string only in the clean target and the foil only in the edited target. Distractors from other items contain neither answer string. Both checkpoints receive identical passage strings, with the entire target in document slot 1, 2, or 4 and distractor order fixed. This audit varies document order, not Word Position within a document; its instruction contrasts remain separate from the main position results. Lexical checks do not certify semantic validity.

Execution and decoding. The recorded backend is MLX 0.32.2 with MLX-LM 0.31.3, four-bit affine quantization, group size 64, and seed 20260916. Greedy generation uses a 20-token output cap and 1,536-token input ceiling. Each item receives one closed-book and twelve grid requests per model, using Appendix A.1's instructions. Table 12 reports logged counts. Realized inputs do not constitute long-context evaluation.

Table 12: Execution totals independently checked against saved request and prediction ledgers. Reaching the output cap is a diagnostic count, not a semantic-error label. No new inference is performed by this audit.

Recorded quantity	Llama-3-8B	Qwen3-8B
Prepared questions	500	500
Baseline exact-correct / remaining	338 / 162	259 / 241
Actual requests	6,500	6,500
Closed-book input tokens (range)	35–92	37–94
Grid input tokens (range)	235–360	241–370
Total input tokens	1,740,804	1,795,682
Total output tokens	31,884	27,828
Requests reaching the 20-token cap	5	18

Reproduction boundary. Saved messages, raw answers, request bindings, cohort IDs, token counts, and checkpoint/tokenizer fingerprints support grid and score reconstruction without loading weights. Hashes of the inference code are recorded for Llama but not Qwen; rendered prompt token IDs and stop reasons were not saved. The reconstruction runner requests rendering without thinking when supported, but this does not certify serialization at inference. Rescoring is distinct from regenerating outputs or independently adjudicating edits.

E Complete Cross-Dataset Experimental Results

Tables 13 and 14 provide the HotpotQA and SearchQA results summarized in Section 4.3. They belong to the main evaluation of fifteen systems, not the two-checkpoint audit. As in Table 1, cells pair regular black clean accuracy on the full set with ↓ foil-only MR on each model's baseline-correct subset. Each Δ cell pairs the corresponding clean-accuracy and MR contrasts (Strict minus Soft, pp); Mean averages positions. Clean accuracy and MR use different denominators and are not subtracted from each other. Aggregation follows Appendix A.4.

Table 13: HotpotQA. Beginning/Middle/End: Word Position within target evidence. Black: full-set clean accuracy; red ↓: MR on closed-book-correct items (%). Δ cells pair clean and MR contrasts, Strict minus Soft (pp); Mean averages positions. Arrows denote MR, not accuracy loss or sign. Bold red: column maxima across models (ties included; averages excluded).

HotpotQA	Beginning			Middle			End			Mean
Model	**Strict**	**Soft**	Δ_{instr}	**Strict**	**Soft**	Δ_{instr}	**Strict**	**Soft**	Δ_{instr}	$\overline{\Delta}_{\text{instr}}$
API-based										
GPT-5.6	85.4$_{\downarrow 20.3}$	84.7$_{\downarrow 12.4}$	0.7$_{\downarrow 7.9}$	89.8$_{\downarrow 17.6}$	89.7$_{\downarrow 5.5}$	0.1$_{\downarrow 12.1}$	85.9$_{\downarrow 25.7}$	87.1$_{\downarrow 14.8}$	-1.2$_{\downarrow 10.9}$	-0.1$_{\downarrow 10.3}$
Claude Opus 4.6	90.5$_{\downarrow 16.0}$	87.6$_{\downarrow 11.5}$	2.9$_{\downarrow 4.5}$	88.0$_{\downarrow 19.7}$	88.3$_{\downarrow 9.0}$	-0.3$_{\downarrow 10.7}$	88.7$_{\downarrow 29.1}$	88.3$_{\downarrow 16.0}$	0.4$_{\downarrow 13.1}$	1.0$_{\downarrow 9.4}$
Qwen3.8-Max	85.7$_{\downarrow 26.9}$	87.5$_{\downarrow 12.9}$	-1.8$_{\downarrow 14.0}$	83.8$_{\downarrow 26.6}$	84.1$_{\downarrow 8.7}$	-0.3$_{\downarrow 17.9}$	82.0$_{\downarrow 23.8}$	88.0$_{\downarrow 12.3}$	-6.0$_{\downarrow 11.5}$	-2.7$_{\downarrow 14.5}$
DeepSeek-V4-Pro	83.2$_{\downarrow 23.6}$	83.9$_{\downarrow 16.8}$	-0.7$_{\downarrow 6.8}$	83.8$_{\downarrow 21.6}$	86.2$_{\downarrow 11.5}$	-2.4$_{\downarrow 10.1}$	84.1$_{\downarrow 27.0}$	81.4$_{\downarrow 13.4}$	2.7$_{\downarrow 13.6}$	-0.1$_{\downarrow 10.2}$
GLM-5.2	85.6$_{\downarrow 24.5}$	86.3$_{\downarrow 12.5}$	-0.7$_{\downarrow 12.0}$	84.2$_{\downarrow 26.1}$	83.6$_{\downarrow 12.3}$	0.6$_{\downarrow 13.8}$	83.0$_{\downarrow 37.8}$	87.9$_{\downarrow 17.0}$	-4.9$_{\downarrow 20.8}$	-1.7$_{\downarrow 15.5}$
Kimi-K3	84.4$_{\downarrow 26.3}$	86.9$_{\downarrow 11.5}$	-2.5$_{\downarrow 14.8}$	83.6$_{\downarrow 24.5}$	86.1$_{\downarrow 6.8}$	-2.5$_{\downarrow 17.7}$	85.4$_{\downarrow 36.0}$	88.0$_{\downarrow 16.2}$	-2.6$_{\downarrow 19.8}$	-2.5$_{\downarrow 17.4}$
HY4-Preview	83.3$_{\downarrow 26.6}$	85.6$_{\downarrow 11.6}$	-2.3$_{\downarrow 15.0}$	85.2$_{\downarrow 30.3}$	80.6$_{\downarrow 11.9}$	4.6$_{\downarrow \mathbf{18.4}}$	85.0$_{\downarrow 25.9}$	85.1$_{\downarrow 11.6}$	-0.1$_{\downarrow 14.3}$	0.7$_{\downarrow 15.9}$
Average	85.4$_{\downarrow 23.5}$	86.1$_{\downarrow 12.7}$	-0.6$_{\downarrow 10.7}$	85.5$_{\downarrow 23.8}$	85.5$_{\downarrow 9.4}$	0.0$_{\downarrow 14.4}$	84.9$_{\downarrow 29.3}$	86.5$_{\downarrow 14.5}$	-1.7$_{\downarrow 14.9}$	-0.8$_{\downarrow 13.3}$
Open-source										
Llama-3-8B	71.1$_{\downarrow 47.3}$	71.9$_{\downarrow 26.8}$	-0.8$_{\downarrow \mathbf{20.5}}$	73.2$_{\downarrow 38.1}$	73.2$_{\downarrow 25.7}$	0.0$_{\downarrow 12.4}$	75.4$_{\downarrow 53.5}$	75.7$_{\downarrow 32.5}$	-0.3$_{\downarrow 21.0}$	-0.4$_{\downarrow 18.0}$
Mistral-7B	74.0$_{\downarrow 47.5}$	71.8$_{\downarrow 28.6}$	2.2$_{\downarrow 18.9}$	74.1$_{\downarrow 41.8}$	75.8$_{\downarrow 25.8}$	-1.7$_{\downarrow 16.0}$	74.6$_{\downarrow 50.8}$	70.1$_{\downarrow 32.4}$	4.5$_{\downarrow 18.4}$	1.7$_{\downarrow 17.8}$
Qwen3-8B	78.7$_{\downarrow 41.1}$	77.3$_{\downarrow 22.4}$	1.4$_{\downarrow 18.7}$	74.0$_{\downarrow 27.3}$	72.6$_{\downarrow 16.8}$	1.4$_{\downarrow 10.5}$	74.1$_{\downarrow 46.9}$	77.4$_{\downarrow 30.3}$	-3.3$_{\downarrow 16.6}$	-0.2$_{\downarrow 15.3}$
Gemma-2-9B	76.7$_{\downarrow 35.2}$	73.0$_{\downarrow 17.6}$	3.7$_{\downarrow 17.6}$	77.3$_{\downarrow 35.2}$	76.6$_{\downarrow 18.3}$	0.7$_{\downarrow 16.9}$	77.5$_{\downarrow 47.2}$	73.2$_{\downarrow 26.1}$	4.3$_{\downarrow \mathbf{21.1}}$	2.9$_{\downarrow \mathbf{18.5}}$
Average	75.1$_{\downarrow 42.8}$	73.5$_{\downarrow 23.9}$	1.6$_{\downarrow 18.9}$	74.7$_{\downarrow 35.6}$	74.6$_{\downarrow 21.7}$	0.1$_{\downarrow 14.0}$	75.4$_{\downarrow 49.6}$	74.1$_{\downarrow 30.3}$	1.3$_{\downarrow 19.3}$	1.0$_{\downarrow 17.4}$
RL-based										
Search-R1	77.6$_{\downarrow 52.0}$	77.5$_{\downarrow \mathbf{45.3}}$	0.1$_{\downarrow 6.7}$	80.9$_{\downarrow \mathbf{53.4}}$	82.5$_{\downarrow \mathbf{45.6}}$	-1.6$_{\downarrow 7.8}$	81.7$_{\downarrow 53.4}$	82.5$_{\downarrow \mathbf{43.6}}$	-0.8$_{\downarrow 9.8}$	-0.8$_{\downarrow 8.1}$
ZeroSearch-base	75.9$_{\downarrow \mathbf{53.7}}$	82.2$_{\downarrow 39.8}$	-6.3$_{\downarrow 13.9}$	79.9$_{\downarrow 40.4}$	81.4$_{\downarrow 37.8}$	-1.5$_{\downarrow 2.6}$	78.7$_{\downarrow \mathbf{55.1}}$	78.1$_{\downarrow 43.1}$	0.6$_{\downarrow 12.0}$	-2.4$_{\downarrow 9.5}$
REDSearcher	81.8$_{\downarrow 44.6}$	76.3$_{\downarrow 29.9}$	5.5$_{\downarrow 14.7}$	81.6$_{\downarrow 45.2}$	78.7$_{\downarrow 40.0}$	2.9$_{\downarrow 5.2}$	78.2$_{\downarrow 52.2}$	79.0$_{\downarrow 38.4}$	-0.8$_{\downarrow 13.8}$	2.5$_{\downarrow 11.2}$
Search-P1	79.1$_{\downarrow 50.8}$	79.2$_{\downarrow 39.4}$	-0.1$_{\downarrow 11.4}$	81.2$_{\downarrow 43.6}$	81.0$_{\downarrow 35.7}$	0.2$_{\downarrow 7.9}$	80.1$_{\downarrow 54.6}$	76.2$_{\downarrow 41.1}$	3.9$_{\downarrow 13.5}$	1.3$_{\downarrow 10.9}$
Average	78.6$_{\downarrow 50.3}$	78.8$_{\downarrow 38.6}$	-0.2$_{\downarrow 11.7}$	80.9$_{\downarrow 45.7}$	80.9$_{\downarrow 39.8}$	0.0$_{\downarrow 5.9}$	79.7$_{\downarrow 53.8}$	79.0$_{\downarrow 41.6}$	0.7$_{\downarrow 12.3}$	0.2$_{\downarrow 9.9}$

Table 14: SearchQA. Beginning/Middle/End: Word Position within target evidence. Black: full-set clean accuracy; red ↓: MR on closed-book-correct items (%). Δ cells pair clean and MR contrasts, Strict minus Soft (pp); Mean averages positions. Arrows denote MR, not accuracy loss or sign. Bold red: column maxima across models (ties included; averages excluded).

SearchQA	Beginning			Middle			End			Mean
Model	**Strict**	**Soft**	Δ_{instr}	**Strict**	**Soft**	Δ_{instr}	**Strict**	**Soft**	Δ_{instr}	$\overline{\Delta}_{\text{instr}}$
API-based										
GPT-5.6	90.4$_{\downarrow 13.9}$	90.0$_{\downarrow 4.6}$	0.4$_{\downarrow 9.3}$	89.2$_{\downarrow 9.9}$	87.4$_{\downarrow 3.6}$	1.8$_{\downarrow 6.3}$	91.0$_{\downarrow 23.0}$	87.1$_{\downarrow 14.5}$	3.9$_{\downarrow 8.5}$	2.0$_{\downarrow 8.0}$
Claude Opus 4.6	95.1$_{\downarrow 14.6}$	90.2$_{\downarrow 7.1}$	4.9$_{\downarrow 7.5}$	92.4$_{\downarrow 17.0}$	86.0$_{\downarrow 6.6}$	6.4$_{\downarrow 10.4}$	91.3$_{\downarrow 16.7}$	89.3$_{\downarrow 9.0}$	2.0$_{\downarrow 7.7}$	4.4$_{\downarrow 8.5}$
Qwen3.8-Max	91.5$_{\downarrow 14.8}$	91.6$_{\downarrow 5.9}$	-0.1$_{\downarrow 8.9}$	90.3$_{\downarrow 17.4}$	91.1$_{\downarrow 8.9}$	-0.8$_{\downarrow 8.5}$	88.0$_{\downarrow 24.3}$	85.9$_{\downarrow 10.3}$	2.1$_{\downarrow 14.0}$	0.4$_{\downarrow 10.5}$
DeepSeek-V4-Pro	90.3$_{\downarrow 18.1}$	88.2$_{\downarrow 5.7}$	2.1$_{\downarrow 12.4}$	87.7$_{\downarrow 16.4}$	88.4$_{\downarrow 5.9}$	-0.7$_{\downarrow 10.5}$	90.1$_{\downarrow 22.7}$	91.0$_{\downarrow 9.7}$	-0.9$_{\downarrow 13.0}$	0.2$_{\downarrow 12.0}$
GLM-5.2	83.7$_{\downarrow 18.1}$	86.3$_{\downarrow 9.2}$	-2.6$_{\downarrow 8.9}$	86.5$_{\downarrow 16.5}$	87.3$_{\downarrow 4.7}$	-0.8$_{\downarrow 11.8}$	88.4$_{\downarrow 26.6}$	88.2$_{\downarrow 11.4}$	0.2$_{\downarrow 15.2}$	-1.1$_{\downarrow 12.0}$
Kimi-K3	89.2$_{\downarrow 15.9}$	89.1$_{\downarrow 10.3}$	0.1$_{\downarrow 5.6}$	83.3$_{\downarrow 15.8}$	88.4$_{\downarrow 8.1}$	-5.1$_{\downarrow 7.7}$	88.4$_{\downarrow 20.0}$	93.7$_{\downarrow 9.0}$	-5.3$_{\downarrow 11.0}$	-3.4$_{\downarrow 8.1}$
HY4-Preview	85.9$_{\downarrow 25.3}$	86.8$_{\downarrow 9.1}$	-0.9$_{\downarrow 16.2}$	84.1$_{\downarrow 22.7}$	84.4$_{\downarrow 11.3}$	-0.3$_{\downarrow 11.4}$	81.4$_{\downarrow 26.6}$	87.6$_{\downarrow 14.1}$	-6.2$_{\downarrow 12.5}$	-2.5$_{\downarrow 13.4}$
Average	89.4$_{\downarrow 17.2}$	88.9$_{\downarrow 7.4}$	0.6$_{\downarrow 9.8}$	87.6$_{\downarrow 16.5}$	87.6$_{\downarrow 7.0}$	0.1$_{\downarrow 9.5}$	88.4$_{\downarrow 22.8}$	89.0$_{\downarrow 11.1}$	-0.6$_{\downarrow 11.7}$	0.0$_{\downarrow 10.3}$
Open-source										
Llama-3-8B	75.8$_{\downarrow 34.3}$	78.4$_{\downarrow 23.9}$	-2.6$_{\downarrow 10.4}$	71.4$_{\downarrow 33.8}$	79.8$_{\downarrow 21.1}$	-8.4$_{\downarrow 12.7}$	73.4$_{\downarrow 44.8}$	73.3$_{\downarrow 31.7}$	0.1$_{\downarrow 13.1}$	-3.6$_{\downarrow 12.1}$
Mistral-7B	80.9$_{\downarrow 36.2}$	75.5$_{\downarrow 22.7}$	5.4$_{\downarrow 13.5}$	77.2$_{\downarrow 37.6}$	79.4$_{\downarrow 20.8}$	-2.2$_{\downarrow \mathbf{16.8}}$	76.0$_{\downarrow 48.7}$	76.5$_{\downarrow 29.1}$	-0.5$_{\downarrow \mathbf{19.6}}$	0.9$_{\downarrow \mathbf{16.6}}$
Qwen3-8B	80.1$_{\downarrow 37.9}$	78.1$_{\downarrow 21.0}$	2.0$_{\downarrow \mathbf{16.9}}$	80.0$_{\downarrow 27.6}$	81.0$_{\downarrow 17.8}$	-1.0$_{\downarrow 9.8}$	75.7$_{\downarrow 38.4}$	79.7$_{\downarrow 20.8}$	-4.0$_{\downarrow 17.6}$	-1.0$_{\downarrow 14.8}$
Gemma-2-9B	77.0$_{\downarrow 29.1}$	79.0$_{\downarrow 21.1}$	-2.0$_{\downarrow 8.0}$	73.3$_{\downarrow 24.0}$	81.1$_{\downarrow 18.5}$	-7.8$_{\downarrow 5.5}$	78.7$_{\downarrow 31.5}$	79.7$_{\downarrow 22.7}$	-1.0$_{\downarrow 8.8}$	-3.6$_{\downarrow 7.4}$
Average	78.5$_{\downarrow 34.4}$	77.8$_{\downarrow 22.2}$	0.7$_{\downarrow 12.2}$	75.5$_{\downarrow 30.8}$	80.3$_{\downarrow 19.6}$	-4.9$_{\downarrow 11.2}$	76.0$_{\downarrow 40.9}$	77.3$_{\downarrow 26.1}$	-1.4$_{\downarrow 14.8}$	-1.8$_{\downarrow 12.7}$
RL-based										
Search-R1	80.8$_{\downarrow \mathbf{50.9}}$	82.3$_{\downarrow 37.6}$	-1.5$_{\downarrow 13.3}$	84.2$_{\downarrow \mathbf{47.1}}$	83.6$_{\downarrow 33.1}$	0.6$_{\downarrow 14.0}$	81.3$_{\downarrow 51.7}$	84.1$_{\downarrow \mathbf{39.4}}$	-2.8$_{\downarrow 12.3}$	-1.2$_{\downarrow 13.2}$
ZeroSearch-base	78.1$_{\downarrow 45.5}$	80.0$_{\downarrow 37.9}$	-1.9$_{\downarrow 7.6}$	80.9$_{\downarrow 42.8}$	78.6$_{\downarrow 30.3}$	2.3$_{\downarrow 12.5}$	81.1$_{\downarrow \mathbf{53.2}}$	82.8$_{\downarrow 38.8}$	-1.7$_{\downarrow 14.4}$	-0.4$_{\downarrow 11.5}$
REDSearcher	83.4$_{\downarrow 38.8}$	86.5$_{\downarrow 31.8}$	-3.1$_{\downarrow 7.0}$	83.9$_{\downarrow 32.0}$	83.3$_{\downarrow 27.1}$	0.6$_{\downarrow 4.9}$	82.2$_{\downarrow 46.0}$	83.3$_{\downarrow 36.7}$	-1.1$_{\downarrow 9.3}$	-1.2$_{\downarrow 7.1}$
Search-P1	77.8$_{\downarrow 45.6}$	85.2$_{\downarrow \mathbf{41.5}}$	-7.4$_{\downarrow 4.1}$	84.7$_{\downarrow 42.4}$	80.3$_{\downarrow \mathbf{34.7}}$	4.4$_{\downarrow 7.7}$	83.2$_{\downarrow 50.4}$	80.2$_{\downarrow 36.7}$	3.0$_{\downarrow 13.7}$	0.0$_{\downarrow 8.5}$
Average	80.0$_{\downarrow 45.2}$	83.5$_{\downarrow 37.2}$	-3.5$_{\downarrow 8.0}$	83.4$_{\downarrow 41.1}$	81.5$_{\downarrow 31.3}$	2.0$_{\downarrow 9.8}$	82.0$_{\downarrow 50.3}$	82.6$_{\downarrow 37.9}$	-0.7$_{\downarrow 12.4}$	-0.7$_{\downarrow 10.1}$

F Additional Position and Prompt Robustness Protocols

This appendix accompanies Tables 2 and 3: 300 QA questions per model, with $|K_A| = 220$ and $|K_B| = 198$. QA, A, and B retain their supplied identifiers. These checks complement the main evaluation of fifteen systems and the 500-item audit; their contrasts and intervals use the corresponding QA records.

F.1 Paired Word Position Analysis

Required records and fixed population. Paired analysis requires frozen question IDs, references and aliases, foils, closed-book responses, and all six clean/edited instruction–Word Position pairs. The audit trail comprises model and chat-template versions, complete requests, decoding settings, input lengths, and truncation records. Valid Word Position comparisons hold document order fixed and verify the realized answer-span locations; the supplementary audit's document slots are not a substitute. K_M remains fixed across conditions. Failed calls require recovery or explicit accounting for missing data, not labels indicating incorrect answers. Discarding incomplete grids or changing denominators by position changes the paired estimand.

Paired estimands. For a fixed model and dataset, let $h_{q,\pi,p}$ indicate an edited response that matches the designated foil but not the reference. The two position contrasts are

$$G_\pi^{E-M} = \frac{100}{|K_M|} \sum_{q \in K_M} (h_{q,\pi,\text{End}} - h_{q,\pi,\text{Mid}}), \qquad G_\pi^{B-M} = \frac{100}{|K_M|} \sum_{q \in K_M} (h_{q,\pi,\text{Beg}} - h_{q,\pi,\text{Mid}}).$$

The third contrast is $G_\pi^{B-E} = G_\pi^{B-M} - G_\pi^{E-M}$, calculated within question before resampling; its interval is not obtained by subtracting interval endpoints. Table 2 reports all three. Under Strict, End-minus-Middle intervals are positive and Beginning-minus-End intervals are negative for both models. Beginning-minus-Middle remains unresolved, as do all Soft contrasts. These results for individual models do not establish the complete ordering or an instruction–position interaction.

Uncertainty and supplementary diagnostics. The supplied position results use 2,000 paired bootstrap resamples, retaining all repeated cells of a sampled question. The summary does not specify its random seed, method of constructing intervals, or multiplicity adjustment. Intervals are interpreted individually, not as simultaneous confirmation of every contrast. These analyses follow inspection of the main summaries and are not claimed as preregistered. A interval containing zero does not establish equality; a nonzero contrast does not identify an attention mechanism. Complete clean/edited records also permit direct transitions and comparisons of gold retention on the same K_M, but these require their own calculations. Clean accuracy on the full set must not be subtracted from MR, and the document-slot audit does not replicate this position intervention.

F.2 Source Policy Paraphrases and Execution

Controlled prompt construction. Appendix A.1 gives the original clauses specifying source policy. The candidates below paraphrase only that policy: Strict prioritizes documents despite apparent errors; Soft permits prior knowledge to override conflicting documents. Verify these meanings and freeze the clauses and cohort before inference. Insert one clause at the same location in the system prompt, holding the role, rules for selecting answers, output contract, wrapper for the user message, and evidence fixed. Add no verification, abstention, or explanation requests.

Source-Policy Paraphrase Candidates

Paraphrase 1: Strict
Treat the supplied documents as the primary authority for your answer. Give a concise answer even when their contents seem incorrect.

Paraphrase 1: Soft
Consult the documents when useful. If their claims contradict well-known facts, give priority to your own knowledge.

Paraphrase 2: Strict
Base your concise answer primarily on the provided documents, including when the information in them appears to be mistaken.

Paraphrase 2: Soft
Draw on the supplied documents where they help, but rely on your own knowledge when they disagree with well-known facts.

Execution and audit trail. The full design crosses three prompt versions, two policies, three Word Positions, and two evidence arms: 36 requests per question and model. Reusing original outputs is valid only with identical requests, model versions, and settings; the two additional versions then require 24 new requests. The original closed-book stratum remains fixed. Resource-limited subsets require selection before inference. Reproduction requires variant IDs, complete messages, context IDs, model revisions, generation settings, raw responses, matcher version, and failure/retry records. The design includes neither live retrieval nor model updates.

Prompt contrasts and changes in effect size. Let $h^{(v)}_{q,\pi,p}$ denote the foil-only indicator under version v, with $v = 0$ the original. For each question define

$$d_q^{(v)} = \frac{1}{3}\sum_p \left(h^{(v)}_{q,\mathrm{Strict},p} - h^{(v)}_{q,\mathrm{Soft},p}\right),$$

$$\Delta_v = \frac{100}{|K_M|}\sum_{q\in K_M} d_q^{(v)}, \qquad R_v = \frac{100}{|K_M|}\sum_{q\in K_M}(d_q^{(v)} - d_q^{(0)}).$$

Comparisons pair questions across versions, policies, and positions. Table 3 reports MR, Δ_v, R_v, and 95% intervals; its resample count and seed are unspecified (2,000 resamples applies only to the position table). The change from the original to itself is zero. All six instruction-contrast intervals exclude zero; all four changes from the original include zero. This supports consistent direction, not equivalence, which requires a prespecified meaningful margin and a dedicated analysis. Contrasts retain their supplied precision rather than being recomputed from rounded MR. Saved requests are needed to verify exact paraphrase bindings to execution messages.