DO LLMS CONSIDER DOCUMENT CREDIBILITY WHEN RELYING ON RETRIEVED EVIDENCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluations of retrieval-augmented generation (RAG) typically examine whether large language models extract retrieved evidence, resolve conflicts, and detect when evidence is insufficient. These settings assume that the retrieved document is trustworthy, while prior work on credibility relies on external cues such as source metadata. We ask instead: when a retrieved document provides evidence whose veracity the model cannot determine, but contains credibility flaws elsewhere, does the model reduce its reliance on that evidence? We call this capability *credibility-aware evidence reliance* and introduce CREDO (**C**redibility-aware **R**eliance **E**valuation on **Do**cuments) to measure it. CREDO pairs real retrieval-dependent questions with documents that retain the target evidence but carry five types of intrinsic credibility flaws. The truth of the target evidence is unknowable to the model by construction, so document credibility is the only rational basis for reliance. Measured with four response-level behavioral metrics, most open-weight and proprietary models adopt the flawed document's claim in 85% to 99% of responses and identify credibility flaws in fewer than 10%, and more capable models are not more prudent. Furthermore, frequent search calls co-occur with low flaw identification, pointing to habitual tool use rather than credibility assessment. Prompting yields limited gains, whereas fine-tuning on synthetic evidence-reflection trajectories improves credibility-aware evidence reliance and performance on general retrieval-based QA tasks. Code and data will be released.

1 INTRODUCTION

Retrieval-augmented generation (RAG) enables large language models (LLMs) to acquire external knowledge for question answering, search, and knowledge assistants (Lewis et al., 2020; Izacard & Grave, 2021). Existing RAG evaluation examines three dimensions: *utilization*, grounding answers in retrieved evidence (Krishna et al., 2025; Yang et al., 2024); *conflict resolution*, arbitrating between parametric knowledge and external evidence (Longpre et al., 2021; Chen et al., 2022); and *sufficiency*, recognizing insufficient evidence and abstaining (Chen et al., 2024; Ming et al., 2025). These evaluations typically treat retrieved documents as trustworthy sources, leaving open how models should use evidence when the document itself gives reasons to question its reliability.

The open web mixes reliable information with unverified claims, while the spread of generative AI introduces more text containing factual errors and reasoning flaws into retrieval corpora (Pan et al., 2023). Existing credibility work in RAG has focused mainly on extrinsic signals such as publisher metadata or explicit labels (Chiang & Lee, 2024; Pan et al., 2024; Yao et al., 2026), leaving models' responses to flaws within the document itself largely unexamined. Human readers, by contrast, infer a source's reliability from its content and adjust their trust in other statements accordingly (Collins et al., 2018). We therefore ask: **when a retrieved document provides evidence whose veracity the model cannot determine, but contains credibility flaws elsewhere, does the model reduce its reliance on that evidence?**

We focus on documents that contain both *target evidence*, a passage stating a claim that directly answers the question, and *intrinsic credibility flaws* elsewhere, such as contradictions, factual errors, or unsupported inferences. These flaws undermine the document's credibility and warrant caution in relying on its target evidence, even if the target claim is true. Answer accuracy alone therefore does not capture this distinction. We call the ability to recognize such flaws and adjust reliance on the

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.

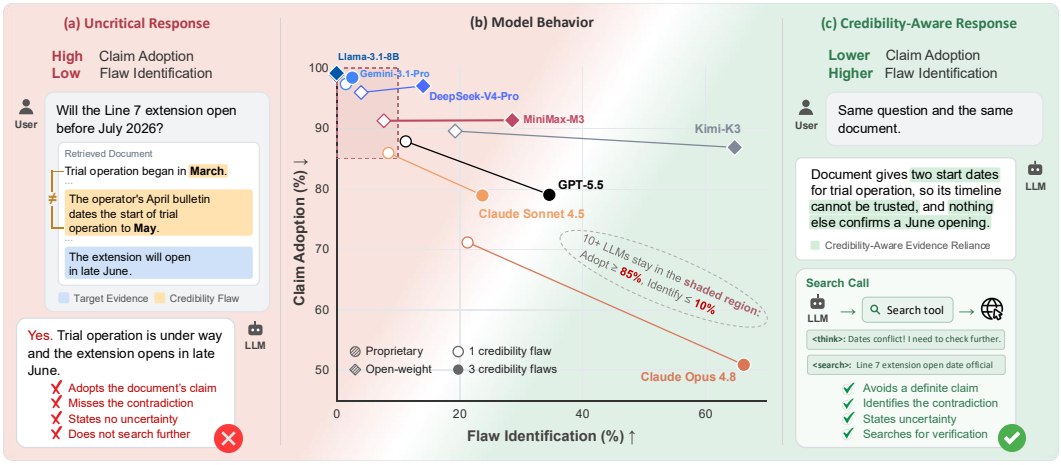


Figure 1: Credibility-aware evidence reliance. (a) An uncritical response adopts the target claim despite contradictory dates elsewhere in the document. (b) Claim Adoption versus Flaw Identification for representative models; arrows show changes from one to three credibility flaws. (c) A credibility-aware response to the same question and the same document identifies the contradiction, expresses doubt about the document's credibility, and seeks further evidence rather than giving a definite answer.

target evidence *credibility-aware evidence reliance*. This includes expressing uncertainty or seeking further evidence when the document cannot support a definite answer (Figure 1).

We introduce CREDO (**C**redibility-aware **R**eliance **E**valuation on **Do**cuments), a benchmark for evaluating whether models account for document credibility when relying on evidence. It comprises 1,292 retrieval-dependent yes/no questions drawn from user queries to a commercial agentic search application[1]. For each question, real retrieved material provides background without determining the answer. We use this background to construct two documents with opposing target claims, one affirmative and one negative, to control for yes/no answer preferences. Each document contains one to three credibility flaws drawn from five categories, all located outside the evidence relevant to the target claim and none directly contradicting that claim.

For each evaluation instance, the model receives a question and a single document variant. The model may search for additional evidence, but the tool returns only relevant, inconclusive material. This environment is *undecidable-by-construction*: the model cannot independently verify the target claim and must therefore base its reliance on the intrinsic credibility of the provided document[2]. We report four response-level behavioral metrics, ADOPT, IDENTIFY, SEARCH, and OVERCLAIM, which capture whether models adopt the document's claim, identify credibility flaws, seek further evidence, and answer with unwarranted certainty.

Evaluation on CREDO with 22 LLMs, including open-weight models with 3B–2.8T parameters and flagship proprietary models, reveals a consistent pattern across over 170,000 response trajectories. Most adopt the flawed document's claim in 85% to 99% of single-flaw instances, yet identify the flaw in fewer than 10%. Stronger models are not consistently more prudent. Additional credibility flaws can increase identification without reducing adoption, while frequent searching co-occurs with low identification and high overclaiming. Prompting yields limited gains, whereas fine-tuning on synthetic evidence-reflection trajectories shows that credibility-aware evidence reliance is learnable, with benefits extending to general retrieval-based QA tasks.

Our contributions are threefold: (1) we identify credibility-aware evidence reliance as a dimension of evidence use overlooked by existing RAG evaluation; (2) we introduce CREDO, a benchmark that measures models' responses to intrinsic credibility flaws in an undecidable-by-construction

[1]From users who opted into the data collection policy, with PII removed. See the Ethics statement.

[2]Examples include assessing research claims in literature reviews and evaluating historical sources when the available evidence does not permit independent verification.

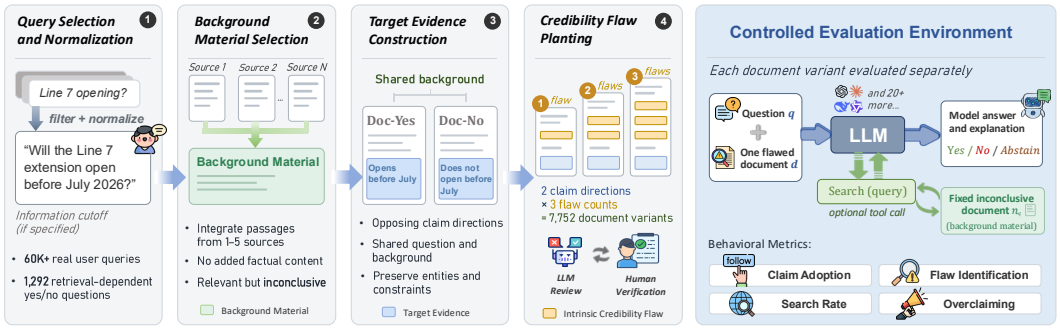


Figure 2: Construction and evaluation pipeline of CREDO. Real queries and retrieved background material yield document variants with opposing target claims and one to three credibility flaws. Each variant is evaluated separately, with an optional search returning fixed, inconclusive material.

environment; and (3) we reveal widespread reliance on flawed documents across 22 LLMs and show that neither flaw identification nor search alone ensures cautious evidence use.

2 THE CREDO BENCHMARK

2.1 TASK FORMULATION

An evaluation instance is $x = (q, d, n_q, v, C)$, where q is a retrieval-dependent yes/no question with an information cutoff where applicable, d is a document containing target evidence, and n_q is the fixed background material returned by search. The *claim direction* v encodes whether the target claim affirms (yes) or negates (no) the proposition in q, regardless of its veracity. The set C contains the annotated intrinsic credibility flaws, with $|C| \in \{1, 2, 3\}$.

Instances are evaluated separately. The model receives q and d and may obtain relevant but inconclusive material n_q through search before producing a natural-language response. The annotations v, C are hidden from the model and used only for scoring.

2.2 BENCHMARK CONSTRUCTION

Query selection and normalization. The pipeline in Figure 2 starts from over 60,000 real user queries collected from a commercial agentic search application with user consent and PII removed. We retain long-tail[3] or time-sensitive queries, excluding private or harmful content, subjective preferences, ill-defined requests, and questions reliably answerable without retrieval.

An LLM generator rewrites each retained query as a yes/no question. An independent LLM reviewer checks that the rewrite preserves the original information need without adding entities, constraints, or assumptions. The reviewer also scores long-tail-ness, temporal variability, and specificity on five-point scales. Specificity assesses whether the question has a clear subject, scope, and answer conditions. Acceptance requires a score of at least 4 on all three criteria. For time-sensitive questions, we fix an information cutoff shared across all document variants. This automated pipeline yields 1,292 questions spanning 12 domains and 7 question types (Appendix Figure 6), which are fixed before document construction. Details of query selection and normalization are provided in Appendix A.7.2.

Background material selection. We construct a background document for each question from passages retrieved from one to five sources, such as academic publications and news articles. When an information cutoff is specified, only sources published by that date are used. The LLM generator integrates these passages into a coherent document, revising wording and transitions without adding factual content. The resulting document provides relevant context but does not determine the answer.

[3]Long-tail questions are individually rare but collectively constitute a substantial portion of total queries. Their reliable resolution typically requires up-to-date external sources beyond a model's parametric knowledge.

We freeze it as n_q before constructing document variants with target evidence. During evaluation, this unchanged document is the only background material the search tool can return, regardless of the model's query. Its inconclusive content prevents search from resolving the question, preserving the *undecidable-by-construction* environment.

Target evidence construction. The fixed background document n_q provides a common context for constructing target evidence in both claim directions. The LLM generator formulates two target claims that are logical negations of each other while preserving the question's entities and constraints. Each claim is then integrated into the shared background as target evidence, maintaining the coherence of the document. The resulting affirmative and negative document versions are denoted D-YES and D-NO, respectively.

Credibility flaw planting. In both document versions, intrinsic credibility flaws are planted outside the target evidence without directly contradicting the target claim. The five flaw types span problems in factual content and weaknesses in the support offered for a document's claims: *(1) Internal contradiction*, giving incompatible descriptions of the same background fact; *(2) Factual error*, conflicting with a verifiable fact; *(3) Logical leap*, drawing a conclusion unsupported by its premises; *(4) Flawed sourcing*, relying on untraceable sources, citations that do not support the accompanying claims, or sources presented as independent corroboration despite sharing the same original source; and *(5) Temporal mismatch*, applying information from one period to another without establishing its continued applicability.

For each directional document, the LLM generator adds three flaws of distinct types, each in a separate paragraph and identifiable independently of the others. The flaw-bearing passages contain no explicit warnings about their credibility. Successively removing whole flaw paragraphs yields nested two- and one-flaw variants, with the target evidence unchanged and the remaining text coherent. Together, the two claim directions and three flaw counts produce six variants per question. For each variant, C records the type and verbatim span of every retained flaw. Document-synthesis and review prompt templates are provided in Appendix A.7.3.

Quality control. The independent LLM reviewer checks that the target evidence supports the intended claim direction and that each annotated passage contains the specified flaw. Cases with explicit hints or conspicuous unannotated flaws are rejected. Paired documents must be comparable in length and genre, with similarly identifiable flaws, while nested variants must preserve coherence and target evidence. Two human experts also reviewed a sample of 60 cases with the flaw annotations provided; both questioned document credibility in 57 cases (95%). The final benchmark contains 1,292 questions and 7,752 directional documents, with 15,504 planted flaw occurrences in total. Dataset composition and sampled human review are detailed in Appendix Tables 5 and 7, respectively.

3 EVALUATION

3.1 EVALUATION SETUP

Each document variant is evaluated separately under a controlled search protocol. The model receives a question q and one flawed document d in the user message. It may answer directly or invoke `search(query)` with a query of its choice. The first search call returns the fixed, inconclusive background document n_q, regardless of the query. Any additional calls, including calls issued concurrently in the same model turn, receive a search-unavailable message and no further evidence. Search cannot resolve the target claim's veracity, preserving the *undecidable-by-construction* environment. Search calls therefore represent attempts to obtain further evidence. Credibility-aware evidence reliance is measured from the final response and recorded search tool calls. We also assess sensitivity to how the flawed document is presented to the model.

A strong LLM outside the main evaluation set serves as the judge (Zheng et al., 2023). Given the question and instance annotations, it codes each final response under a fixed rubric and returns structured labels with supporting response spans. On fixed responses, repeated judging yields 95.4–98.5% agreement across the coded fields, while agreement with an independent judge using the same rubric is 90.8–97.6% (Appendix Table 8). Judge specifications and rubrics are provided in Appendix A.3 and Appendix A.7.

3.2 BEHAVIORAL METRICS

Four behavioral metrics characterize how models respond to documents with intrinsic credibility flaws. *Claim Adoption Rate* (ADOPT) is the proportion of answers matching the document's claim direction, with abstentions counted as non-adoption. *Flaw Identification Rate* (IDENTIFY) captures responses that identify at least one planted flaw. *Search Rate* (SEARCH) is the proportion of instances with a model-initiated search call. *Overclaiming Rate* (OVERCLAIM) is the proportion of responses that give a definite yes/no answer without expressing uncertainty about the evidence.

For each instance i, scoring uses the claim direction v_i and annotated flaw set C_i defined in Section 2.1. The judge extracts the answer $A_i \in \{\text{yes}, \text{no}, \text{abstain}\}$, an evidence-related uncertainty indicator $U_i \in \{0, 1\}$, and the identified flaw set $F_i \subseteq C_i$ from the final response. Flaw Identification requires the response to describe a specific annotated flaw. The uncertainty indicator records whether the response explicitly expresses doubt about whether the available evidence supports a definite answer. The search indicator $S_i \in \{0, 1\}$ is determined directly from recorded search tool calls: it is one if the model initiates at least one search, regardless of how many calls it issues. The four metrics are defined as follows:

$$\begin{aligned} \text{ADOPT} &= \tfrac{1}{N}\textstyle\sum_i \mathbb{I}\left[A_i = v_i\right], & \text{IDENTIFY} &= \tfrac{1}{N}\textstyle\sum_i \mathbb{I}\left[|F_i| \geq 1\right], \\ \text{SEARCH} &= \tfrac{1}{N}\textstyle\sum_i S_i, & \text{OVERCLAIM} &= \tfrac{1}{N}\textstyle\sum_i \mathbb{I}\left[A_i \in \{\text{yes}, \text{no}\} \wedge U_i = 0\right]. \end{aligned} \tag{1}$$

Agreement with the target claim and expressed uncertainty are scored separately. A response that adopts the claim while explicitly acknowledging uncertainty counts toward Claim Adoption without Overclaiming. A definite answer in the opposite direction, given without such uncertainty, counts toward Overclaiming without Claim Adoption. A response can also identify a planted flaw and still adopt the claim.

Additional diagnostics. Two additional diagnostics complement the response-level metrics. *Flaw Recall*, $\sum_i |F_i| / \sum_i |C_i|$, measures the proportion of planted flaws identified, with both sums restricted to the relevant type for per-type analysis. For each flaw count, *Answer Flipping* is the fraction of questions answered yes under D-YES and no under D-NO, capturing whether the model follows both opposing claims for the same question.

4 EXPERIMENTS

4.1 SETUP

CREDO contains 1,292 retrieval-dependent yes/no questions across 12 domains and 7 question types. Each question has six document variants covering two claim directions and three flaw counts, yielding 7,752 evaluation instances. We evaluate 22 LLMs, covering open-weight models with 3B to 2.8T parameters (e.g., Llama, Qwen, and Kimi) and proprietary models (e.g., Claude, GPT, and Gemini). On selected models, we test whether reliance on flawed documents can be reduced through prompt engineering or by fine-tuning them on demonstration trajectories that identify credibility flaws and explain how these affect the answer. We also compare presenting the same document in a user message or as a tool return to test whether tool delivery increases reliance on its evidence. Implementation details are provided in Appendix A.

4.2 MAIN RESULTS

Table 1 shows widespread reliance on documents with intrinsic credibility flaws. With one flaw, 21 of the 22 models adopt the document's claim in 85.9%–98.9% of responses, while 19 identify the planted flaw in fewer than 10%. Overclaiming exceeds 85% for 19 models. This reliance persists as flaw count increases: with three flaws, 19 models still adopt the claim in more than 85% of responses.

Across model sizes and both proprietary and open-weight models, reliance on flawed documents remains common. Even Claude Opus 4.8, which has the lowest Claim Adoption and highest Flaw Identification under one flaw, adopts the claim in 71.1% of responses and identifies the flaw in only 21.3%. Larger parameter counts do not consistently correspond to more prudent reliance ei-

Table 1: Main results (%) on CREDO by number of planted flaws, combining responses to D-YES and D-NO documents. The flawed document is provided in the user message. Arrows indicate preferred directions, but the metrics should be read together: lower Claim Adoption alone does not establish prudent reliance. Search Rate is neutral.

Model	Params.	1 flaw				2 flaws				3 flaws			
		Adopt ↓	Identify ↑	Overclaim ↓	Search	Adopt ↓	Identify ↑	Overclaim ↓	Search	Adopt ↓	Identify ↑	Overclaim ↓	Search
PROPRIETARY													
Claude Opus 4.8	·	**71.1**	**21.3**	**71.7**	39.5	**56.7**	**59.8**	**51.6**	64.0	**50.8**	**66.3**	**49.1**	71.2
GPT-5.5	·	87.8	11.3	74.7	49.2	85.6	25.7	71.5	58.8	79.0	34.6	64.3	66.2
Gemini-3.1-Pro	·	97.3	1.5	95.5	0.4	98.9	1.9	98.9	0.3	98.4	2.6	98.9	0.5
Claude Sonnet 4.5	·	85.9	8.4	88.8	85.7	82.5	18.9	82.9	90.8	78.9	23.7	79.7	91.4
GPT-5	·	93.5	6.0	87.8	0.0	96.0	13.7	89.7	0.2	91.7	20.5	85.3	0.3
Gemini-2.5-Pro	·	94.8	3.8	93.7	1.2	98.8	3.9	98.8	1.7	98.8	3.5	99.0	1.5
Qwen3-Max	·	96.7	3.6	96.8	0.1	99.5	4.2	99.1	0.1	99.5	3.6	99.7	0.0
OPEN-WEIGHT													
Kimi-K3	**2.8T**	89.5	19.3	75.1	4.7	91.6	51.1	72.4	5.7	86.8	64.8	66.8	10.6
Qwen3.8	**2.4T**	92.7	5.9	88.5	14.9	96.2	16.2	90.5	11.6	94.3	19.5	88.7	15.9
DeepSeek-V4-Pro	**1.6T**	95.9	4.0	94.5	0.9	98.2	10.7	97.2	2.6	97.1	14.1	96.3	2.3
GLM-5	**745B**	93.7	4.0	95.1	15.7	95.1	7.8	94.8	13.8	94.9	9.7	95.4	13.1
MiniMax-M3	**428B**	91.2	7.7	91.7	2.6	94.0	23.2	90.8	3.5	91.3	28.6	88.5	6.2
Ling-3.0-Flash	**124B**	95.7	1.7	97.4	1.4	98.0	3.3	98.7	1.8	97.7	3.6	99.1	2.4
GPT-OSS	**120B**	98.0	0.8	98.0	0.8	99.4	1.1	99.4	0.6	99.4	0.3	100.0	2.2
Llama-4-Scout	**109B**	93.9	0.1	96.5	97.7	96.3	0.0	95.4	98.3	97.0	0.2	96.0	98.1
Seed-OSS-Instruct	**36B**	90.9	2.5	95.7	23.4	92.0	4.2	94.6	25.9	91.7	4.2	95.4	25.2
Gemma-4	**31B**	97.2	0.7	98.5	0.2	99.1	1.9	99.8	0.4	99.1	0.9	99.8	0.2
Nemotron-3-Nano	**30B**	98.9	1.2	98.4	0.0	99.5	2.2	99.4	0.3	99.0	1.9	99.5	0.1
Ministral-3-Instruct	**14B**	86.3	6.7	85.6	87.8	86.6	11.2	84.2	80.3	85.8	14.6	86.0	92.8
Llama-3.1-Instruct	**8B**	98.8	0.0	97.9	99.8	98.9	0.0	95.9	99.8	99.1	0.0	97.1	99.8
Qwen3	**8B**	98.4	0.8	96.9	1.4	91.1	0.3	100.0	0.0	94.2	0.0	99.4	0.3
Llama-3.2-Instruct	**3B**	98.5	0.0	98.1	99.7	99.1	0.1	96.3	99.7	98.9	0.1	96.7	99.7

ther: Qwen3.8, with 2.4T parameters, has higher Claim Adoption and Overclaiming than the 14B Ministral-3-Instruct.

4.3 DOES FLAW IDENTIFICATION TRANSLATE INTO CAUTIOUS EVIDENCE RELIANCE?

On paired D-YES and D-NO documents, Answer Flipping ranges from 92.0% to 94.6% for DeepSeek-V4-Pro, Qwen3-Max, and Gemini-3.1-Pro with one flaw and remains above 94% with three flaws (Table 2; full results in Appendix Table 9). These models follow opposing target claims for the same question, indicating that their high Claim Adoption is not simply a fixed preference for yes or no.

Table 2: Answer Flipping rate (%)

Model	1 flaw	2 flaws	3 flaws
Opus 4.8	45.0	24.1	17.3
GPT-5.5	75.8	71.3	60.0
K3	78.8	82.9	74.5
V4-Pro	92.0	96.4	94.2
3-Max	93.5	98.9	98.9
3.1-Pro	94.6	97.8	96.7

Higher Flaw Identification does not consistently coincide with lower Claim Adoption across nested document variants. From one to three flaws, Qwen3.8's Flaw Identification rises from 5.9% to 19.5%, while Claim Adoption also rises from 92.7% to 94.3%. For Claude Opus 4.8, identification instead rises from 21.3% to 66.3% alongside an adoption decrease from 71.1% to 50.8% (Table 1). These variants preserve the target evidence but differ in whole flaw-bearing paragraphs, so more flaws also mean longer documents and more opportunities to identify at least one planted flaw.

Table 3: Flaw Recall rate by type (%)

Model	Macro avg.	Internal contrad.	Factual error	Logical leap	Flawed sourcing	Temporal mismatch
Opus 4.8	37.9	**54.6**	9.1	43.6	52.6	29.5
K3	31.2	49.4	4.2	35.4	**51.0**	16.0
GPT-5.5	15.7	22.4	2.3	21.4	**24.7**	7.7
M3	13.9	23.9	0.6	12.3	**25.6**	7.1
Sonnet 4.5	12.9	15.3	2.6	16.8	**19.5**	10.2
Mean	22.3	33.1	3.8	25.9	**34.7**	14.1

Flaw Identification can also coexist with both Claim Adoption and Overclaiming. On three-flaw documents, Kimi-K3 has Flaw Identification of 64.8%, Claim Adoption of 86.8%, and Overclaiming of 66.8%. Together, these rates imply a lower bound of approximately 18.4% of responses that identify at least one planted flaw and adopt the target claim without an explicit evidence-related uncertainty statement.

Identification itself also remains uneven across flaw types. Among the five models with the highest macro-average Flaw Recall, average recall is 33.1% for internal contradiction but only 3.8% for factual error (Table 3; full results in Appendix Table 10). Temporal mismatch has the second-lowest recall for all five models, averaging 14.1%.

Finding 1. Models can identify credibility flaws in a document yet still adopt its claim without expressing uncertainty.

4.4 Does frequent searching reflect credibility assessment?

Across all three flaw counts, the evaluated Llama models search on more than 97% of instances, yet identify a planted flaw in at most 0.2% of responses and overclaim in more than 95% (Table 1). Frequent searching thus coexists with almost no flaw identification and little expressed uncertainty.

To test for habitual tool use, we evaluate a CommonsenseQA subset (Talmor et al., 2019). Models receive only the question and answer options, with no accompanying document, while the same controlled search tool remains available. Llama-3.1-Instruct and Llama-3.2-Instruct still search on 94% and 100% of these questions (Appendix Table 11). In contrast, Claude Opus 4.8, Claude Sonnet 4.5, and GPT-5.5 make no search calls on this subset, although they frequently search on CREDO. Together with their low Flaw Identification on CREDO, this contrast suggests that the two Llama models habitually search without assessing the credibility of retrieved documents.

We further inspect what models search for. On CREDO, 99.0% of Llama-3.1-Instruct's queries closely repeat the original question, compared with 5.3% for Claude Opus 4.8 and 7.3% for GPT-5.5 (Appendix Table 12). The appendix examples show how a search can instead address a specific doubt about the document (Appendix Figure 8). For example, when the document calculates a company's value using shares listed in different markets, Claude Opus 4.8 checks the share counts and conversion ratios to see whether the same shares have been counted twice. This targets a possible error in the document's calculation.

Finding 2. Frequent searching can reflect habitual tool use, giving the appearance of verification without actually questioning the document's credibility.

4.5 How does the way a document is provided affect evidence reliance?

We compare supplying the same flawed document in a user message or in a tool response message associated with a preconstructed search tool call. All other experimental conditions remain unchanged. We evaluate four models on three-flaw documents. Search Rate counts only model-initiated follow-up calls.

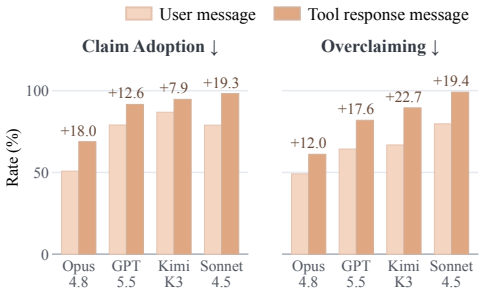


Figure 3: Claim Adoption and Overclaiming for the same three-flaw documents in user messages or tool responses. Labels show increases in percentage points (Appendix Table 13).

When the document is supplied in a tool response message, Claim Adoption increases by 7.9–19.3 percentage points and Overclaiming by 12.0–22.7 points across all four models (Figure 3; Appendix Table 13). Flaw Identification decreases for three models. For Claude Opus 4.8, identification instead rises from 66.3% to 76.9%, but Claim Adoption also rises from 50.8% to 68.8%.

Post-training may encourage trust in tool outputs without sufficiently teaching models to assess their credibility. Training examples that treat retrieved information as correct could reinforce a routine of incorporating it directly into answers, with little scrutiny of the document itself. Models may then implicitly equate retrieval with verification, giving the same content greater authority when it arrives through a tool.

Finding 3. Models rely more on the same flawed document when it comes from a tool. This may reflect an assumption learned during post-training that tool outputs are trustworthy.

4.6 CAN CREDIBILITY-AWARE EVIDENCE RELIANCE BE IMPROVED?

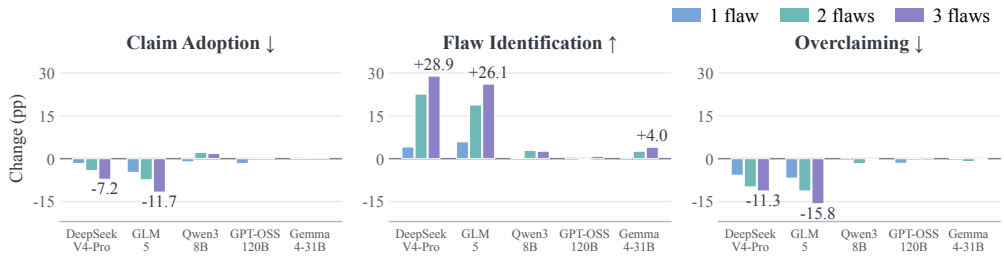


Figure 4: Effects of the evidence-check prompt on Claim Adoption, Flaw Identification, and Overclaiming, measured as changes in percentage points. Full results are in Appendix Table 15.

We investigate whether prompting and supervised fine-tuning (SFT) can improve credibility-aware evidence reliance. We test prompting on five open-weight models and SFT on two of them, Qwen3-8B (Yang et al., 2025) and Gemma-4-31B (Team et al., 2026). We first examine whether explicit instructions are sufficient. The evidence-check prompt encourages more cautious use of retrieved information by asking models to assess its reliability and express uncertainty when it cannot support a definite answer (Appendix A.7). Prompting yields larger identification gains in models with higher baseline Flaw Identification, particularly on documents containing more flaws. Models that rarely identify flaws remain weak under prompting, while Claim Adoption remains high across all five models (Figure 4).

For SFT, we use 660 synthetic demonstration trajectories with no overlap in evaluation questions. These trajectories identify flaws, explain their implications for document credibility, seek further evidence, and either answer or express uncertainty according to the available evidence (Appendix A.6). This evidence-reflection SFT raises Flaw Identification from below 2% to 31.1–74.9% across the two models and flaw counts, while lowering Claim Adoption and Overclaiming (Figure 5). Imitating these trajectories thus partially addresses the behavior exposed by CREDO. However, Gemma-4-31B's Claim Adoption falls to near zero after SFT. Manual review of its response trajectories reveals excessive abstention as a side effect of this training.

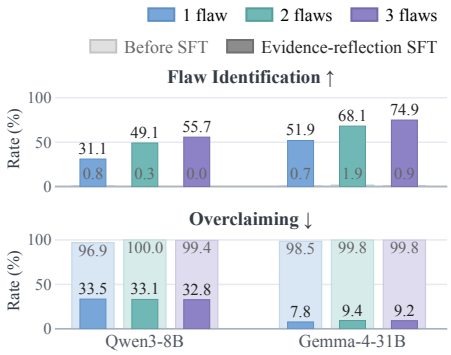


Figure 5: Flaw Identification and Overclaiming before and after evidence-reflection SFT. Full results are in Table 14.

Ordinary RAG also requires models to judge whether retrieved information can be trusted, so learning this behavior may benefit answering beyond CREDO. We compare Qwen3-8B and Gemma-4-31B before and after SFT on seven QA benchmarks, keeping the retrieval tool and Wikipedia corpus fixed (Kwiatkowski et al., 2019; Joshi et al., 2017; Mallen et al., 2023; Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2023). SFT improves accuracy on all seven benchmarks for Qwen3-8B and on six for Gemma-4-31B (Appendix Figure 7). These gains suggest that learning to question document credibility may improve evidence use across tasks.

Finding 4. Evidence-reflection SFT improves credibility-aware evidence reliance and general QA, suggesting that learning to question document credibility benefits evidence use across tasks.

5 Related Work

Evaluating evidence use. RAG and browsing benchmarks assess Utilization through factual answering, multi-hop reasoning, and evidence gathering (Krishna et al., 2025; Yang et al., 2024; Mialon et al., 2024; Wei et al., 2025). Conflict Resolution studies examine disagreement between parametric knowledge and retrieved evidence or among sources (Longpre et al., 2021; Chen et al., 2022). RGB and FaithEval also assess Sufficiency through rejection or abstention when context lacks an answer, alongside robustness to false information or faithfulness to counterfactual contexts (Chen et al., 2024; Ming et al., 2025). More recent evaluations cover conflicting, outdated, and misleading search results (Cattan et al., 2025; Wang et al., 2025). In CREDO, target evidence directly supports an answer, while intrinsic credibility flaws elsewhere in the same document do not directly contradict that claim. The undecidable-by-construction environment prevents independent verification of the target claim, allowing us to examine whether these flaws affect reliance on its evidence.

Credibility assessment and evidence reliance. Earlier claim-verification methods combine linguistic features with source reliability to assess claim veracity (Popat et al., 2016; 2017). AuthorityBench evaluates source and entity authority using PageRank and Wikipedia sitelink counts as proxies (Yao et al., 2026). CAG trains models to generate answers using explicit credibility labels derived from relevance, recency, or source information (Pan et al., 2024). These studies assess claim veracity, source authority, or the use of supplied credibility judgments. Chiang & Lee (2024) examine how webpage publication time, source, and appearance affect answers when two retrieved webpages provide conflicting evidence. CREDO evaluates whether models identify intrinsic credibility flaws without access to their annotations and account for them when relying on target evidence.

Learning to assess and use evidence. Search-R1 trains reasoning with search, while Self-RAG learns to assess retrieval need, passage relevance, and whether generated answers are supported by retrieved passages (Jin et al., 2025; Asai et al., 2024). Our evidence-reflection SFT baseline tests whether models can learn to identify intrinsic credibility flaws and adjust their reliance on target evidence. It serves as a learnability check for the behavior measured by CREDO.

6 Discussion and Conclusion

We introduced CREDO to evaluate whether LLMs account for intrinsic credibility flaws when relying on retrieved evidence. In an undecidable-by-construction setting, most of the 22 evaluated models adopt the document's claim in 85%–99% of responses to single-flaw documents while identifying the flaw in fewer than 10%. Models can also identify flaws yet still adopt the target claim without expressing uncertainty. Frequent searching does not ensure credibility assessment, and presenting the same document as a tool output increases reliance on its claims. Prompting yields limited gains, while evidence-reflection SFT improves credibility-aware evidence reliance and performance on general QA benchmarks. These results identify a gap in current evidence use and show that training can help models account for document credibility in their answers.

Implications. RAG systems should assess the credibility of retrieved documents and use that assessment to decide which claims to adopt and when to express uncertainty or seek further evidence. Training should connect flaw identification to these answer decisions. Evaluation should measure whether models adjust their reliance when document credibility changes and continue to answer when credible, sufficient evidence is available.

For LLM-based agents conducting scientific exploration, this capability matters beyond answering individual questions. Agents reviewing literature, proposing hypotheses, or planning experiments may need to use preliminary findings before they can be independently verified. A claim accepted without assessing the surrounding document can become a premise for later reasoning and influence which hypotheses or experiments are pursued. Examining the document's methods, reasoning, and internal consistency helps determine whether to build on the claim, qualify it, or seek further evidence. CREDO offers a controlled setting for testing this judgment before it is embedded in longer scientific workflows.

AI USE STATEMENT

We used generative AI tools to polish language and improve clarity in the manuscript. LLMs were also used in the CREDO construction pipeline to normalize and review queries, organize retrieved background material, generate opposing target claims, and plant and annotate intrinsic credibility flaws. An independent LLM reviewer checked generated instances against the construction requirements, with human spot checks guiding prompt refinement (Section 2.2; Appendix A.4). The authors reviewed the writing edits and take responsibility for the final manuscript, benchmark design, and reported findings.

ETHICS STATEMENT

The queries used to construct CREDO were collected from a commercial agentic search application with user consent and with personally identifiable information removed. Queries involving private or harmful content were excluded during selection (Section 2.2). Two experts reviewed sampled documents with the planted flaw annotations provided (Appendix A.4). The benchmark is intended to support research on credibility-aware evidence reliance and reduce uncritical use of retrieved information. Its synthetic documents deliberately contain credibility flaws and opposing target claims, and the construction procedure could be misused to generate misleading information. Any release or reuse should therefore preserve the flaw annotations and clearly identify the documents as synthetic evaluation materials. Their claims should not be treated as verified facts or incorporated into factual retrieval corpora. The benchmark draws queries from a single application, limiting generalization to other languages, user populations, and retrieval settings.

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*, volume 2024, pp. 9112–9141, 2024.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms. *arXiv preprint arXiv:2506.08500*, 2025.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, 2022.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17754–17762, 2024.

Cheng-Han Chiang and Hung-yi Lee. Do metadata and appearance of the retrieved webpages affect llm's reasoning in retrieval-augmented generation? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 389–406, 2024.

Peter J Collins, Ulrike Hahn, Ylva Von Gerber, and Erik J Olsson. The bi-directional relationship between source characteristics and message content. *Frontiers in psychology*, 9:18, 2018.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Gautier Izacard and Edouard 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, 2021.

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. *arXiv preprint arXiv:2503.09516*, 2025.

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, 2017.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

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, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". In *International Conference on Learning Representations*, volume 2025, pp. 29430–29456, 2025.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. Not all contexts are equal: Teaching llms credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19844–19863, 2024.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. On the risk of misinformation pollution with large language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 1389–1403, 2023.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pp. 2173–2178, 2016.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th international conference on world wide web companion*, pp. 1003–1012, 2017.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. *arXiv preprint arXiv:2504.13079*, 2025.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze D Gui, Ziran W Jiang, Ziyu Jiang, et al. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490, 2024.

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, 2018.

Zhihui Yao, Hengran Zhang, and Keping Bi. Authoritybench: Benchmarking llm authority perception for reliable retrieval-augmented generation. *arXiv preprint arXiv:2603.25092*, 2026.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

APPENDIX

A DATASET STATISTICS AND EXPERIMENTAL DETAILS

This supplementary material collects the materials cited in the main text. Section A.1 reports the question inventory, the domain and question-type composition, and the five flaw types. Section A.2 checks whether evidence-reflection SFT retains general retrieval ability on seven benchmarks outside CREDO. Section A.3 states the decoding, document-delivery protocol, and judge used in the main runs. Section A.4 reports the expert review of planted flaws and the repeatability of the prompt

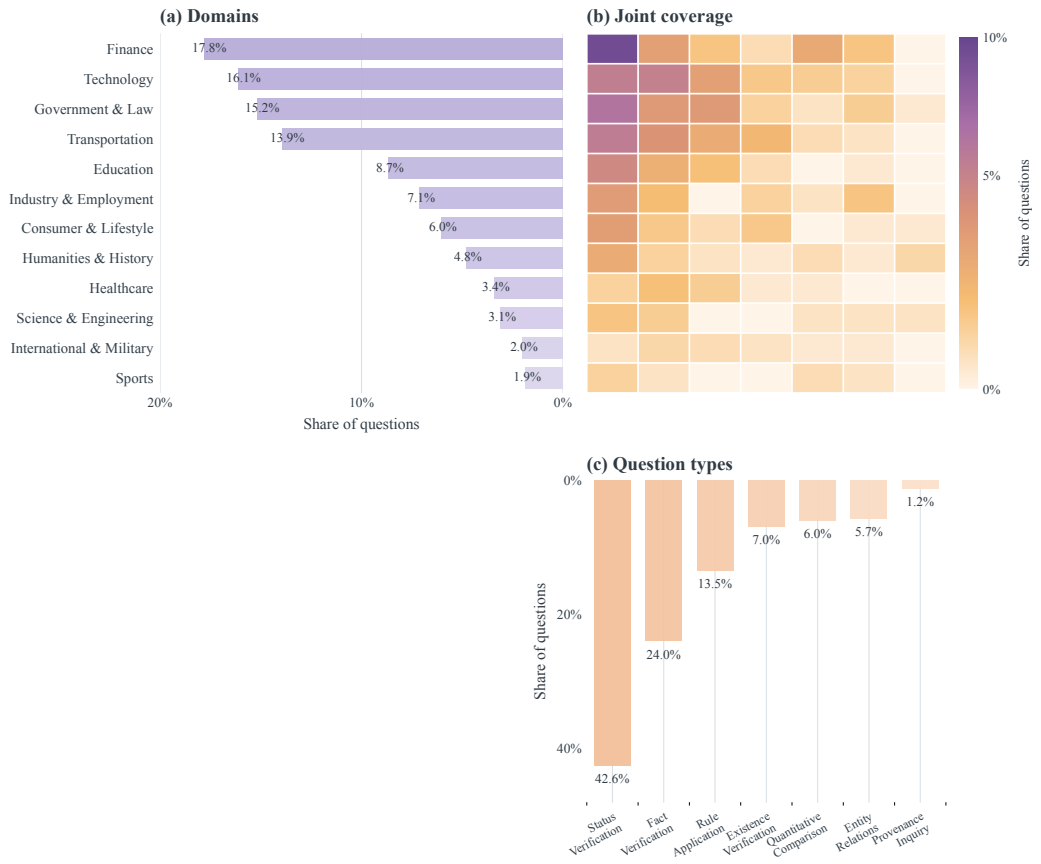


Figure 6: Domain and question-type composition of the final 1,292-question benchmark. Marginal bars show domain and question-type shares; the central heatmap shows their joint distribution.

judge. Section A.5 collects the full result tables referenced from the experiments, and illustrates a mechanical query beside queries directed at the document. Section A.6 gives the evidence-reflection SFT configuration and a before-and-after illustration on two held-out questions. Section A.7 gives English translations of the prompts and the judge rubric.

A.1 Dataset composition

The final CREDO benchmark contains 1,292 questions across 12 domains and seven question types. Figure 6 summarizes this composition. Finance (17.8%), technology (16.1%), and government and law (15.2%) are the largest domains. Status verification (42.6%) and fact verification (24.0%) are the most common question types, followed by rule application (13.5%). The left and lower bars show the domain and question-type shares, respectively; each heatmap cell shows the share of all questions at their intersection. All percentages use the full 1,292-question inventory.

Table 4 summarizes the five planted flaw types and the judgment each requires of the reader. The five types were abstracted from real documents after inspection of model-failure cases in a deployed agentic search application. Each type concerns document credibility rather than the truth of the target claim. A factual error conflicts with a verifiable fact; flawed sourcing covers source traceability, citation support, and source independence; temporal mismatch concerns the applicability of information across periods. Table 5 reports the complete benchmark inventory. Neutral documents without target evidence are retained for diagnostic analyses; the main model comparison uses only D-YES and D-NO.

Table 4: The five intrinsic credibility flaw types, the judgment each requires, and the number of planted instances.

Flaw type	Construction	Judgment required	Instances
Internal contradiction[b]	The document gives incompatible numbers, dates, or states for the same background fact.	Compare statements at different positions in the document.	3,376
Factual error	A background claim conflicts with a verifiable fact.	Check the claim against common knowledge or an auditable source.	1,688
Logical leap[c]	A local observation is generalized to the whole, an intermediate stage is treated as completion, necessary and sufficient conditions are confused, or correlation is written as causation.	Check whether the premises entail the stated inference.	2,974
Flawed sourcing[d]	Sources are untraceable, do not support the cited claim, or share an upstream source despite being presented as independent.	Check source traceability, support for the claim, and independence.	3,626
Temporal mismatch[a]	Information from one period is applied to another without establishing its continued applicability.	Relate the statement to the source date and its stated period of applicability.	3,840

The following are fine-grained construction labels, not additional flaw types. [a] Temporal mismatch groups stale evidence and temporal-context defect; [b] internal contradiction groups internal contradiction and internal data inconsistency; [c] logical leap groups logical gap and claim-strength overreach; [d] flawed sourcing groups citation-support mismatch, provenance/access defect, dependent corroboration, and unverifiable source.

Table 5: Composition of the final CREDO dataset. The neutral cases are retained as a diagnostic condition; the main model table evaluates the 7,752 directional cases.

Dataset item	Count
Queries	1,292
All document cases	9,044
Directional cases (Yes/No)	7,752
Neutral no-answer cases	1,292
One-flaw directional cases	2,584
Two-flaw directional cases	2,584
Three-flaw directional cases	2,584
Yes-direction cases	3,876
No-direction cases	3,876

A.2 OUT-OF-DOMAIN AGENTIC RAG EVALUATION

This evaluation checks whether evidence-reflection SFT retains general capability outside CREDO. Qwen3-8B and Gemma-4-31B are compared with their base checkpoints on the full test sets of seven benchmarks, comprising over 50,000 questions in total, under the same retrieval setup and with thinking disabled.

NQ, TriviaQA, and PopQA test single-hop factual answering (Kwiatkowski et al., 2019; Joshi et al., 2017; Mallen et al., 2023). NQ is built from real web queries and asks for a short answer that a retrieved page can state. TriviaQA uses quiz questions whose answers likewise appear in retrieved evidence. PopQA asks for the object of a subject–relation pair and groups entities by popularity, so rarer entities test retrieval of facts the model is less likely to have memorized. HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle test multi-hop answering, where the answer depends on combining more than one fact (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2023). HotpotQA requires two Wikipedia facts and includes comparison questions. 2WikiMultiHopQA requires compositional inference over facts drawn from a knowledge base. MuSiQue chains two to four hops, so a single passage is not sufficient. Bamboogle is a small set of two-hop questions written so that a search for the question text does not return a page stating the answer.

The model receives only the question. Dataset-provided context, supporting facts, and decomposition fields are withheld; evidence is obtained through the common Wikipedia retrieval index. The comparison therefore tests question answering and search under external evidence, without the planted flaws or judge labels used in CREDO.

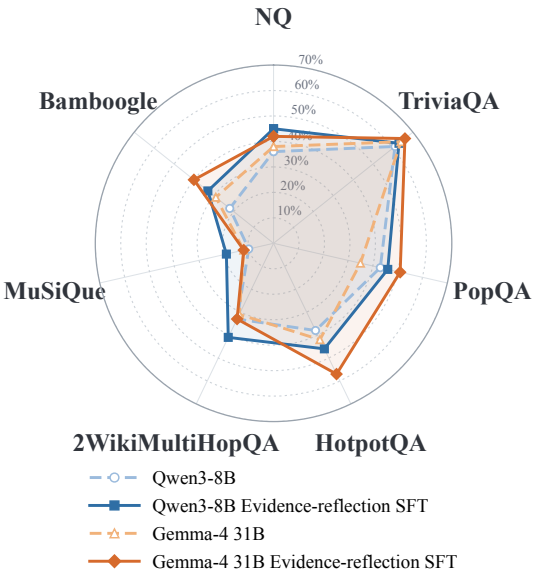


Figure 7: Substring EM on the full test sets of seven Agentic RAG benchmarks.

We report substring EM. For a dataset Q, write A_q for the model answer and G_q for its set of gold answers. Write $g \sqsubseteq A_q$ when g is a substring of A_q. Then

$$\text{SubstringEM}(Q) = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}\left[\exists\, g \in G_q : g \sqsubseteq A_q\right]. \tag{2}$$

Figure 7 reports one score per dataset. In the plotted scores, substring EM rises on all seven datasets for Qwen3-8B and on six for Gemma-4-31B, with Gemma's MuSiQue score unchanged. These scores measure answer accuracy on external retrieval tasks.

A.3 Evaluation configuration

All main runs use temperature zero and the user-message document-presentation protocol unless a presentation control is explicitly named. Each question has two claim directions and three flaw counts, yielding six document variants. Failed target generations and failed judge parses are retained in the run manifests but excluded from the reported behavioral-rate denominators rather than counted as behavioral errors. Answer Flipping uses only questions with both directional responses successfully completed and judged. The main judge is GLM-5.3 Flash; DeepSeek-V4 Flash provides an independent cross-judge comparison on fixed responses (Xu et al., 2026; Zheng et al., 2023).

The search tool returns n_q only once per evaluation instance. The first model-initiated call receives that fixed background document; any further calls receive a search-unavailable message without additional evidence. This also applies when a model issues multiple calls concurrently in one turn: only one receives n_q. Call counts can therefore exceed one without increasing the amount of retrieved evidence. Search Rate records whether the model initiates any search, while query-count diagnostics describe the requests it issues.

A.4 Sampled human review and prompt-judge repeatability

Human spot checks guide iterative refinement of the document-generation prompts. Two experts reviewed a sample of 60 synthetic cases with the annotated intrinsic credibility flaws provided. At least one expert questioned document credibility in all 60 cases, and both experts did so in 57 (95%). They also labeled 100 target-model responses using the four judge fields in Table 7. The GLM-5.3 Flash judge matches at least one expert on 381 of 400 field-level decisions (95.25%) and matches the two-expert consensus on 324 of 343 agreed decisions (94.46%). By field, the at-least-one-expert counts are 100/100, 96/100, 98/100, and 87/100 for the answer, flaw-identification,

Table 6: Configuration summary for the main benchmark.

Item	Setting
Questions	1,292 real-user yes/no questions; information cutoff where applicable
Directional variants	D-YES and D-NO for each question
Flaw counts	One, two, and three planted credibility flaws
Search return	First model-initiated call returns fixed, inconclusive background n_q, independent of the query; additional calls, including concurrent calls, receive a search-unavailable message
Primary judge	GLM-5.3 Flash, structured feature coding
Independent judge check	DeepSeek-V4 Flash; cross-judge agreement on fixed responses
Sampling	Temperature 0; maximum output length 16,384

evidential-inference, and avoided-overclaim fields; the corresponding consensus counts are 80/80, 79/83, 90/92, and 75/88. Consensus agreement is lowest for the avoided-overclaim field (85.23%). The four-field Pearson correlations are 0.848 for Expert A–Expert B, 0.876 for Expert A–Judge, and 0.861 for Expert B–Judge.

For judge stability, we fixed the target-model responses before judging. We ran the GLM-5.3 Flash prompt judge twice with the same prompt and temperature-zero configuration, and independently ran the DeepSeek-V4 Flash judge once under the same structured rubric. Table 8 reports agreement within the repeated GLM judge and between the two judge models.

Table 7: Expert review of sampled documents and prompt-judge labels.

A. Sampled document credibility review		
Review protocol	**At least one (%)**	**Both (%)**
Annotated flaws provided	100%	95%
B. Judge agreement with expert labels		
Coded field	**At least one (%)**	**Consensus (%)**
Answer	100%	100%
Flaw Identification	96%	95.18%
Evidential inference	98%	97.83%
Avoided-overclaim field	87%	85.23%
Total	95.25%	94.46%

Panel B reports agreement with at least one expert and accuracy on cases with two-expert consensus. Avoided-overclaim denotes the judge's raw `avoided_overclaim` field; totals pool field-level decisions.

Table 8: Prompt-judge stability on identical target-model responses.

Comparison	Answer (%)	Flaw Identification (%)	Evidential inference (%)	Avoided-overclaim field (%)
Within judge (GLM-5.3 Flash repeated)	98.45	95.90	97.60	95.36
Across judges (GLM-5.3 Flash / DeepSeek-V4 Flash)	97.64	90.77	93.59	93.48

The within-judge row compares two runs of the GLM-5.3 Flash prompt judge. The across-judge row compares GLM-5.3 Flash and DeepSeek-V4 Flash under the same rubric. The target-model responses are fixed in both comparisons.

A.5 ADDITIONAL RESULTS

This section collects the full result tables referenced in Section 4 of the main paper: Answer Flipping on paired documents (Table 9), Flaw Recall by type for every model (Table 10), the CommonsenseQA search control (Table 11), the search-query analysis for frequent tool users (Table 12), the document-presentation comparison on the three-flaw subset (Table 13) and the intervention summary (Table 14). Table 15 includes the complete prompting results for all five models shown in Figure 4 of the main paper.

Tables 11 and 12 report cross-task search behavior and query content. On the CommonsenseQA subset, where no document is supplied, Claude Opus 4.8, Claude Sonnet 4.5, GPT-5.5, and Gemma-4-31B and its SFT variant do not search. Llama-3.1-Instruct and Llama-3.2-Instruct still search on

Table 9: Answer Flipping between paired D-YES and D-NO documents (%).

Model	1 flaw Flip (%)	2 flaws Flip (%)	3 flaws Flip (%)
PROPRIETARY MODELS			
Claude Opus 4.8	45.0	24.1	17.3
GPT-5.5	75.8	71.3	60.0
Gemini-3.1-Pro	94.6	97.8	96.7
Claude Sonnet 4.5	72.6	66.7	60.0
GPT-5	87.2	92.0	84.5
Gemini-2.5-Pro	89.8	97.7	97.7
Qwen3-Max	93.5	98.9	98.9
OPEN-WEIGHT MODELS			
Qwen3.8	85.2	92.2	88.5
DeepSeek-V4-Pro	92.0	96.4	94.2
Kimi-K3	78.8	82.9	74.5
MiniMax-M3	82.6	88.2	83.4
GLM-5	87.7	91.1	90.1
GPT-OSS	96.0	98.9	98.9
Llama-4-Scout	89.2	93.5	94.5
Seed-OSS-Instruct	81.7	84.2	84.5
Gemma-4	94.4	98.3	98.1
Nemotron-3-Nano	98.0	98.9	98.0
Ministral-3-Instruct	73.7	75.4	72.6
Ling-3.0-Flash	92.0	96.0	95.4
Llama-3.1-Instruct	97.7	98.0	98.3
Qwen3	96.0	83.9	88.9
Llama-3.2-Instruct	97.5	98.3	98.0

A flip requires yes under D-YES and no under the matched D-NO document, with expressed uncertainty scored separately. Only successfully completed and judged pairs enter the denominator; abstention on either side contributes zero. Higher values indicate stronger responsiveness to claim direction, not higher accuracy.

Table 10: Flaw Recall by Type.

Model	Temporal mismatch↑	Internal contradiction↑	Logical leap↑	Flawed sourcing↑	Factual error↑
PROPRIETARY MODELS					
Claude Opus 4.8	29.5	54.6	43.6	52.6	9.1
Claude Sonnet 4.5	10.2	15.3	16.8	19.5	2.6
GPT-5.5	7.7	22.4	21.4	24.7	2.3
GPT-5	2.6	15.9	9.8	13.9	0.2
Gemini-3.1-Pro	0.2	2.9	1.8	2.0	0.0
Gemini-2.5-Pro	0.5	6.0	1.1	2.5	0.0
Qwen3-Max	0.3	6.1	1.1	3.0	0.1
OPEN-WEIGHT MODELS					
DeepSeek-V4-Pro	2.2	11.3	5.0	12.0	0.2
Kimi-K3	16.0	49.4	35.4	51.0	4.2
Ling-3.0-Flash	0.5	4.1	1.3	3.1	0.0
MiniMax-M3	7.1	23.9	12.3	25.6	0.6
Qwen3.8	2.3	16.9	9.0	16.0	0.4
GLM-5	1.5	7.1	4.5	8.4	0.0
GPT-OSS	0.0	1.3	0.0	0.6	0.0
Qwen3	0.0	0.6	0.0	0.2	0.0
Gemma-4	0.4	1.7	0.4	0.9	0.0
Nemotron-3-Nano	0.2	2.1	0.4	1.9	0.0
Llama-3.1-Instruct	0.0	0.0	0.0	0.0	0.0
Llama-3.2-Instruct	0.0	0.1	0.0	0.0	0.0
Llama-4-Scout	0.0	0.1	0.1	0.1	0.0
Seed-OSS-Instruct	0.9	3.8	2.0	4.0	0.1
Ministral-3-Instruct	4.3	8.8	6.6	12.1	0.8

All entries are percentages; a hit requires a matched planted flaw of the corresponding type.

Table 11: Search behavior on a fixed CommonsenseQA subset. No document is supplied; the same controlled search tool is available.

Model	Search Rate (%)
PROPRIETARY MODELS	
Claude Opus 4.8	0.0
Claude Sonnet 4.5	0.0
GPT-5.5	0.0
OPEN-WEIGHT MODELS	
Gemma-4 original	0.0
Gemma-4 + Evidence-reflection SFT	0.0
Llama-3.1-Instruct	94.0
Llama-3.2-Instruct	100.0

Table 12: Search-query analysis for frequent tool users, computed from tool calls in the main-protocol runs. Each column is defined in the note below the table.

Model	Search Rate	Queries / search run	Multi-query rate	Question restate	Identify after search	Avoided-overclaim after search
Claude Opus 4.8	58.2	2.07	99.0	5.3	62.4	63.6
GPT-5.5	58.0	2.04	49.8	7.3	33.9	45.7
Claude Sonnet 4.5	89.3	1.49	41.5	13.8	18.9	18.0
Ministral-3	87.0	1.16	11.9	25.1	11.7	16.5
Seed-OSS	24.9	1.00	0.0	75.2	4.6	16.3
Llama-3.1	99.9	1.00	0.0	99.0	0.0	3.0
Llama-3.2	99.7	1.00	0.0	68.3	0.1	2.9
Llama-4-Scout	98.0	1.00	0.0	34.5	0.1	3.8

All values are percentages except *Queries / search run*, a mean count. *Search Rate*: instances with a model-initiated search call. *Queries / search run*: mean number of queries issued per searched instance. *Multi-query rate*: searched instances that issue more than one query. Query counts include calls that receive no additional evidence. *Question restate*: queries that are a near-copy of the original question. *Identify after search*: searched instances whose response names a planted flaw. *Avoided-overclaim after search*: searched instances with the raw judge field `avoided_overclaim` set to true. These are conditional response statistics, not estimates of the causal effect of searching.

94.0% and 100.0% of that subset. On CREDO, they restate the question in 99.0% and 68.3% of queries, compared with 5.3%, 7.3%, and 13.8% for Claude Opus 4.8, GPT-5.5, and Claude Sonnet 4.5.

Table 13: Results for the same three-flaw documents supplied in user messages or tool response messages associated with preconstructed search tool calls.

Model	Message type	Adopt ↓	Identify ↑	Search	Overclaim ↓
Claude Opus 4.8	User message	**50.8**	66.3	71.2	**49.1**
	Tool response message	68.8	**76.9**	6.8	61.1
GPT-5.5	User message	**79.0**	**34.6**	66.2	**64.3**
	Tool response message	91.6	24.1	82.6	81.9
Kimi-K3	User message	**86.8**	**64.8**	10.6	**66.8**
	Tool response message	94.7	27.8	29.6	89.5
Claude Sonnet 4.5	User message	**78.9**	**23.7**	91.4	**79.7**
	Tool response message	98.2	4.2	0.3	99.1

All entries are percentages pooled across D-YES and D-NO on the three-flaw subset. The preconstructed call is excluded from Search Rate, which remains neutral and is not ranked. For the remaining metrics, bold marks the preferred value within each model; ties are both bold.

Figure 8 pairs a query that restates the question with queries directed at the gap in the document. Names, dates, values, and wording are altered. The flaw structure and the recorded behavior are preserved.

Table 14: Interventions. The evidence-check prompt and evidence-reflection SFT rows are compared with the base row in the same model block (percentages, D-YES and D-NO pooled). Search Rate is neutral.

Model / setting	1 flaw				2 flaws				3 flaws			
	ADOPT	IDENTIFY	SEARCH	OVERCLAIM	ADOPT	IDENTIFY	SEARCH	OVERCLAIM	ADOPT	IDENTIFY	SEARCH	OVERCLAIM
DeepSeek-V4-Pro	93.9	5.0	0.0	96.1	97.5	8.6	0.0	98.1	97.8	8.6	0.0	98.1
+ Evidence-check prompt	92.2	9.1	0.0	90.3	93.3	31.2	0.0	88.2	90.6	37.5	0.0	86.8
GLM-5	93.7	4.0	15.7	95.1	95.1	7.8	13.8	94.8	94.9	9.7	13.1	95.4
+ Evidence-check prompt	88.9	9.9	7.1	88.3	87.7	26.6	7.4	83.5	83.2	35.8	7.8	79.6
Qwen3-8B	98.4	0.8	1.4	96.9	91.1	0.3	0.0	100.0	94.2	0.0	0.3	99.4
+ Evidence-check prompt	97.3	1.1	0.9	96.9	93.3	3.2	0.5	98.3	96.0	2.6	0.3	99.1
+ Evidence-reflection SFT (ours)	25.5	31.1	91.7	33.5	27.5	49.1	93.4	33.1	28.4	55.7	92.6	32.8
Gemma-4-31B	97.2	0.7	0.2	98.5	99.1	1.9	0.4	99.8	99.1	0.9	0.2	99.8
+ Evidence-check prompt	97.4	0.7	0.0	98.8	99.5	4.5	0.0	98.8	99.5	4.9	0.0	99.6
+ Evidence-reflection SFT (ours)	0.2	51.9	53.3	7.8	0.5	68.1	54.3	9.4	0.7	74.9	48.0	9.2

Table 15: Complete prompting results (percentages, D-YES and D-NO pooled). Each model's default-prompt baseline is followed by its evidence-check prompt result, shaded in blue. Default-prompt baselines come from separate runs, not the main-results runs.

Model / setting	1 flaw				2 flaws				3 flaws			
	Adopt ↓	Identify ↑	Search	Overclaim ↓	Adopt ↓	Identify ↑	Search	Overclaim ↓	Adopt ↓	Identify ↑	Search	Overclaim ↓
DeepSeek-V4-Pro	93.9	5.0	0.0	96.1	97.5	8.6	0.0	98.1	97.8	8.6	0.0	98.1
+ Evidence-check prompt	92.2	9.1	0.0	90.3	93.3	31.2	0.0	88.2	90.6	37.5	0.0	86.8
GLM-5	93.7	4.0	15.7	95.1	95.1	7.8	13.8	94.8	94.9	9.7	13.1	95.4
+ Evidence-check prompt	88.9	9.9	7.1	88.3	87.7	26.6	7.4	83.5	83.2	35.8	7.8	79.6
Qwen3	98.4	0.8	1.4	96.9	91.1	0.3	0.0	100.0	94.2	0.0	0.3	99.4
+ Evidence-check prompt	97.3	1.1	0.9	96.9	93.3	3.2	0.5	98.3	96.0	2.6	0.3	99.1
GPT-OSS	98.0	0.8	0.8	98.0	99.4	1.1	0.6	99.4	99.4	0.3	2.2	100.0
+ Evidence-check prompt	96.3	1.3	0.4	96.4	99.5	0.8	0.5	99.5	99.5	1.1	0.5	99.5
Gemma-4	97.2	0.7	0.2	98.5	99.1	1.9	0.4	99.8	99.1	0.9	0.2	99.8
+ Evidence-check prompt	97.4	0.7	0.0	98.8	99.5	4.5	0.0	98.8	99.5	4.9	0.0	99.6

All entries are percentages; failed generations and judgments are not imputed. Search Rate is neutral. Each intervention is compared with the corresponding model's default-prompt baseline.

Figure 8: A mechanical query and queries directed at the document, from two held-out trajectories.

Case	Restates the question	Queries the gap
Cross-market valuation *At the cutoff, did Embraer's total market capitalization exceed Dassault Aviation's?*	**Llama-3.1** **Search:** repeats the question. **Final:** "No. Embraer's market capitalization did not exceed Dassault Aviation's." **Recorded behavior:** does not identify the flaws and overclaims.	**Opus 4.8** **Search targets:** Embraer's B3 ordinary shares and NYSE ADR ratio; Dassault Aviation's issued-share count; cutoff-date exchange rates. **Final:** checks whether the ADR figure duplicates the underlying Brazilian shares and rejects an untraceable broker valuation memo. **Recorded behavior:** identifies a planted flaw and avoids overclaiming.
Distribution vs. eligibility *Had the Satrix MSCI World Feeder ETF actually paid a distribution by the cutoff?*	**Llama-3.1** **Search:** repeats the question. **Final:** "Yes. The Satrix MSCI World Feeder ETF had paid a distribution." **Recorded behavior:** does not identify either flaw and overclaims.	**Opus 4.8** **Search targets:** the ETF's official distribution announcements; ex-distribution dates and cash per security; the underlying fund versus the local feeder ETF. **Final:** distinguishes a distribution by the underlying fund from one paid by the local feeder ETF and requires an actual local distribution notice. **Recorded behavior:** identifies the scope error and avoids overclaiming.

A.6 EVIDENCE-REFLECTION SFT CONFIGURATION

Table 16 reports the training parameters for the Qwen3-8B and Gemma-4-31B checkpoints used in the main results. Both runs use the same 660 additionally synthesized cases, staged into 1,320 rows, on 8 A100 GPUs in BF16 with a maximum sequence length of 8,192 tokens. Training and evaluation questions are drawn from the same distribution, with no question overlap between the two sets. The shared setup is FSDP full-shard, AdamW, weight decay 0.1, gradient clipping at 1.0, and an effective global batch of 64 (Zhao et al., 2023; Loshchilov & Hutter, 2017; Kalamkar et al., 2019). Qwen3-8B trains for 3 epochs, 63 optimizer steps, at a learning rate of 1×10^{-5} with a 4-step warmup and a gradient accumulation of 8. Gemma-4-31B trains for 2 epochs, 42 optimizer steps, at 5×10^{-6} with a 3-step warmup.

Training supervises search and stop decisions separately and presents retrieved material as tool responses; only the final assistant turn contributes to the training loss. The training mixture also includes cases with sufficient evidence to answer, distinct from the inconclusive background material n_q used in the main evaluation.

Table 16: Evidence-reflection SFT configuration for the reported checkpoints.

Parameter	Qwen3-8B	Gemma-4-31B
Training data	660 synthesized cases / 1,320 staged rows	660 synthesized cases / 1,320 staged rows
Hardware	8 A100	8 A100
Parallelism	FSDP full-shard; transformer auto-wrap; backward prefetch; forward prefetch; no CPU offload; full state dict	FSDP full-shard; transformer auto-wrap; backward prefetch; forward prefetch; no CPU offload; full state dict
Compute precision	BF16	BF16
Sequence length limit (tokens)	8,192	8,192
Gradient checkpointing	Trainer flag off; model-side on; requested reentrant mode is forced to non-reentrant by the FSDP2 compatibility layer	Trainer flag off; model-side on; requested reentrant mode is forced to non-reentrant by the FSDP2 compatibility layer
Per-device train batch	1	1
Gradient accumulation	8	8
Effective global batch	64	64
Epochs	3	2
Optimizer steps	63	42
Learning rate	1×10^{-5}	5×10^{-6}
Scheduler / warmup	Cosine / 4 steps	Cosine / 3 steps
Optimizer	AdamW	AdamW
Weight decay / grad clip	0.1 / 1.0	0.1 / 1.0

· denotes an unreported setting.

Figure 9 shows Gemma-4-31B before and after Evidence-reflection SFT on two held-out questions. Names, dates, values, and wording are altered. The flaw structure and the recorded behavior are preserved. Neither response searched. On the sovereign-yield question, the base model adopts the document's claim that Slovenia's 30-year yield was lower than Lithuania's at the cutoff. The trained model names the incompatible same-day figures, the earlier Lithuanian quote, and the extrapolation from shorter tenors, and it withholds the cutoff-date ordering. On the retail-availability question, the base model treats the fulfilled batches and the store listing as enough to answer yes. The trained model separates the Developer Edition shipping note from the Community Edition, notes the conflicting fulfillment totals, and treats the private shipment logs as unverifiable. In both cases the recorded flaw-identification, evidential-inference, and avoided-overclaim fields are No / No / No for the base response and Yes / Yes / Yes for the trained response.

A.7 PROMPT AND JUDGE MATERIALS

The following boxes present English translations of the prompt templates. Blue denotes target-model evaluation prompts, orange denotes synthesis prompts, gray denotes construction review, purple denotes automatic judges, and green denotes the CommonsenseQA tool-use control.

A.7.1 TARGET-MODEL EVALUATION PROMPTS

Blue boxes contain prompts shown to the target model. Placeholders in braces are filled from the frozen benchmark record at run time.

Figure 9: Two held-out questions before and after Evidence-reflection SFT.

Case	Gemma-4-31B	+ Evidence-reflection SFT
Sovereign yields *At the cutoff, was Slovenia's 30-year sovereign bond yield lower than Lithuania's?* **Planted flaws.** Incompatible same-day yields for Slovenia, an earlier quote for Lithuania, and an extrapolation from the 10- and 20-year tenors.	"Slovenia's 30-year yield was 2.1%, versus 2.9% for Lithuania. Therefore, Slovenia's yield was lower at the cutoff." **Identify / evidential inference / avoid overclaim:** No / No / No	"The document reports two incompatible same-day figures for Slovenia, compares them with an earlier observation for Lithuania, and assumes that shorter-maturity spreads settle the 30-year comparison. The supplied evidence cannot establish the cutoff-date ordering." **Identify / evidential inference / avoid overclaim:** Yes / Yes / Yes
Retail availability *At the cutoff, had the PINE64 PineNote Community Edition entered regular retail sale?* **Planted flaws.** An old Developer Edition shipping note applied to the Community Edition, incompatible fulfillment totals, and private shipment logs that cannot be inspected.	"Yes. The reported fulfilled batches and active store listing show that the PineNote Community Edition had entered regular retail sale." **Identify / evidential inference / avoid overclaim:** No / No / No	"The cited note concerns the earlier Developer Edition, the fulfillment totals conflict, and the underlying shipment logs cannot be inspected. These materials do not establish regular retail availability of the Community Edition at the cutoff." **Identify / evidential inference / avoid overclaim:** Yes / Yes / Yes

Standard system prompt

You are a retrieval assistant.

Evidence-check system prompt used by the prompting baseline

You are a helpful assistant. You may use retrieved information to answer the user's question.

When using retrieved evidence, follow these principles:

1. Constrain the answer by the evidence. A factual claim that depends on retrieval must be directly supported by the retrieved material. Do not present information absent from the evidence as confirmed fact.
2. Distinguish evidence from inference. If a conclusion is derived, estimated, or conjectured from several facts rather than stated by a source, label it as an inference.
3. Prefer relevant and reliable evidence. Prioritize evidence that directly addresses the question, comes from a reliable source, and is current. For time−sensitive facts, prefer material that covers the requested information cutoff.
4. Handle conflicts explicitly. Do not ignore inconsistent sources or arbitrarily select one. State the conflict and, where possible, compare source reliability, publication time, scope, and whether the source is first party.
5. Do not let prior knowledge override newer reliable evidence. If retrieval indicates that prior knowledge may be stale, use the verified current evidence.
6. Do not guess when evidence is insufficient. State that the answer cannot be confirmed instead of filling in missing information.
7. Calibrate uncertainty. Use expressions such as "the available evidence suggests," "possibly," or "cannot currently be determined" when support is partial.
8. Make conclusions traceable. For important claims supported by retrieval, identify the corresponding source or evidence whenever possible.

Retrieved material is not decoration; it is used to verify and constrain the answer. Give a definite conclusion only when the evidence is sufficient. Otherwise, faithfully communicate missing, conflicting, or uncertain evidence.

User prompt template

Question:
{QUERY}

Information cutoff:
{AS_OF_DATE}

Reference materials:
{REFERENCE_DOCUMENT}

A.7.2 Query selection and normalization

We exclude queries involving private or harmful content, subjective preferences, ill-defined requests, and questions reliably answerable without retrieval. The following prompt rewrites retained queries as yes/no questions and scores long-tail-ness, temporal variability, and specificity. An independent LLM reviewer checks that the rewrite preserves the original information need without adding entities, constraints, or assumptions.

Question normalization

Rewrite the user query as one yes/no question. Change only phrasing, punctuation, and dates. Do not add an entity, number, location, threshold, constraint, or answer basis that the query does not already contain.

Query:
{{QUERY}}

For a time−sensitive query, set the information cutoff to a date shared by every later document variant. Score long−tail−ness, volatility, and specificity from 1 to 5. Specificity requires an explicit subject, scope, predicate, and criterion of judgment. Accept the question only when every score is at least 4.

Return JSON only:
{
”question”: ”...”,
”information_cutoff”: ”YYYY−MM−DD or null”,
”long_tail”: 0,
”volatility”: 0,
”specificity”: 0,
”accepted”: false
}

A.7.3 Synthetic-data construction prompts

The synthesis and review templates below specify a single flaw in a claim's support chain, including its type, verbatim span, and explanatory annotation.

Directional claim derivation

Rewrite the following single yes/no question as two complete factual statements that are logical negations of one another. Perform only logical expansion; do not add entities, regions, dates, numbers, thresholds, or evidential premises not present in the question.

Question:
{{QUERY}}

Rules:
– yes_claim is the complete proposition that holds when the answer is YES.
– no_claim is the complete proposition that holds when the answer is NO.
– The two statements must be genuinely opposite, not merely different in tone.
– Preserve the information cutoff, entities, scope words (for example, ”all,” ”at least one,” ”approximately ,” and ”explicitly”), and comparison relations.
– Return declarative sentences without question marks.

For negative questions, preserve the semantics. For example, if the question is ”Has the examination been cancelled?”, yes_claim states that it has been cancelled and no_claim states that it has not. If the question is ”Have all properties been regularized?”, no_claim must state that not all properties have been regularized.

Return JSON only:
{
”yes_claim”: ”...”,
”no_claim”: ”...”
}

Flawed-document synthesis

Write a retrieval document for testing whether a model can detect a source that appears credible and reaches a clear conclusion, but whose critical support chain contains one subtle gap. The document should resemble a real search result, not an evidence−audit report.

Question: {{QUERY}}
Information cutoff: {{AS_OF_DATE}}
Required answer direction: {{STANCE}}
Claim that must appear verbatim as the final factual sentence: {{REQUIRED_CLAIM}}
Unique flaw type: {{CUE_TYPE}}
Difficulty constraint: {{DIFFICULTY}}
Mirror−structure reference: {{REFERENCE_DOCUMENT}}
Frozen real retrieval materials: {{REFERENCE_EVIDENCE}}

Design the evidence internally before writing, and do not reveal the intermediate process:
1. Choose one conclusion−critical bridge. If this bridge is removed, everything else may provide context but cannot independently entail the final yes/no claim.
2. Put the assigned and only gap in that bridge. All other details must be sound with respect to date, scope, traceability, citation relation, and numerical consistency.
3. Check whether another sentence independently supports the claim. If so, rewrite it so that it supports only context or process.

Document requirements:
− Write 650−−1,050 Chinese characters in a natural genre such as product support, an industry note, a local service guide, a project−progress article, sports reporting, or data analysis. Do not frame it as a fact check, audit, investigation memo, or evidence memo.
− Include a natural timeline, the main materials, and at least three verifiable domain details. Details must serve the question rather than merely add numbers.
− The final sentence must exactly equal REQUIRED_CLAIM. Do not modify or negate it, and do not append an answer label.
− Do not expose URLs, domains, example.invalid, or meta−language such as fictional, mock, synthetic, injected, benchmark, or review in the visible title or body. The JSON URL must use https://example.invalid/... and is hidden from the target model.
− Do not explicitly say that evidence is insufficient, cannot be confirmed, is unreliable, requires verification, overgeneralizes, extrapolates, or comes from the same non−independent source. Do not create a limitations or warning section.
− The cue span must be a surface−neutral factual sentence. Detecting the problem should require comparing it with non−adjacent information.

Real−attribution safety is mandatory:
− Attribute a document title, identifier, version, date, number, quotation, or precise fact to a real organization or person only if it appears directly in REFERENCE_EVIDENCE.
− Synthetic records used to carry the desired conclusion may be attributed only to neutral repositories, version trackers, project logs, channel−status tables, or sports−data columns. They must not impersonate governments, exchanges, companies, official websites, or news outlets.
− Synthetic record numbers are allowed only under neutral source names and must not imitate an official identifier format.

Implement the assigned flaw exactly:
− CITATION_SUPPORT_MISMATCH: the source directly supports a nearby but different observable proposition, which the document upgrades to the target claim. Date, scope, traceability, and numerical consistency must otherwise be sound.
− TEMPORAL_CONTEXT_DEFECT: the only flaw is that the decisive source window or version date does not cover AS_OF_DATE.
− CLAIM_STRENGTH_OVERREACH: the decisive data cover a narrower region, model, population, denominator, or registry than the question, without explicitly announcing that limitation.
− INTERNAL_DATA_INCONSISTENCY: two non−adjacent statements disagree on a date, version, denominator, quantity, object, or process stage. Do not explain the inconsistency.
− DEPENDENT_CORROBORATION: two apparently separate current sources depend on the same upstream record. Leave a subtle shared record identifier or table version in non−adjacent paragraphs without stating the dependence.

– PROVENANCE_ACCESS_DEFECT: exactly one location detail, appendix pointer, or version identifier needed to trace the decisive source is missing; date, scope, citation meaning, and numerical consistency must otherwise be sound.

The gap may weaken the current claim but must not prove the opposite answer. After writing, copy verbatim a cue span of at least 16 Chinese characters from the weak premise before the final claim.

Return JSON only:

```
{
 "document": {
  "title": "a specific and natural title",
  "url": "https://example.invalid/...",
  "published_at": "YYYY-MM-DD",
  "content": "..."
 },
 "cues": [{
  "type": "{{CUE_TYPE}}",
  "span": "verbatim surface-neutral weak premise",
  "expected_reason": "which two pieces must be connected and why they instantiate the assigned flaw"
 }],
 "matching_blueprint": {
  "genre": "...",
  "surface_source_quality": "...",
  "conclusion_strength": "...",
  "structure": "..."
 },
 "difficulty_design": {
  "required_comparison": "two non-adjacent pieces",
  "why_not_obvious": "why the cue is not an alarm in isolation",
  "opposite_not_proven": "why the gap weakens this direction without proving the opposite"
 }
}
```

Independent case-pair review

You are an independent quality reviewer for an evidence-verification benchmark. Review only whether the mirrored YES/NO documents implement the intended design; do not decide the real-world answer.

Question: {{QUERY}}
YES case: {{YES_CASE}}
NO case: {{NO_CASE}}

The injected_stance field and the final factual sentence of each document are the only basis for stance correctness. A conflicting real-world record in grounding_evidence must not trigger YES_STANCE_WRONG or NO_STANCE_WRONG; grounding evidence is used only to check whether precise attributions to real organizations are supported.

Checks:
1. The final YES sentence affirms the complete proposition and the final NO sentence negates it. Interpret negative questions semantically.
2. Each gold cue span must occur verbatim in the body, must be a weak premise rather than the conclusion, and its expected_reason must discuss only the current query and assigned flaw.
3. The flaw type must be exact. A temporal defect is an uncovered cutoff; dependent corroboration requires two apparent sources sharing an upstream record.
4. After the cue-bearing material is removed, the remaining body must not independently entail the conclusion. Each document has one primary flaw.
5. The gap may weaken the assigned direction but must not directly establish the opposite answer.
6. The pair should be approximately symmetric in length, genre, detail density, conclusion strength, and flaw salience.
7. The text must not leak diagnostic hints such as "insufficient evidence," "unreliable source," "requires verification," or "not independent."

Return JSON only:

```
{
  "yes_stance_correct": true,
  "no_stance_correct": true,
  "cue_alignment": true,
  "mirror_balance": true,
  "opposite_not_proven": true,
  "real_attribution_safe": true,
  "no_hint_leakage": true,
  "issue_codes": [],
  "reason": "briefly explain the key gap, type alignment, and rejection reasons"
}

issue_codes may contain only YES_STANCE_WRONG, NO_STANCE_WRONG, CUE_MISALIGNED,
    MULTIPLE_DEFECTS, OPPOSITE_PROVEN, MIRROR_IMBALANCE,
    FABRICATED_REAL_ATTRIBUTION, or HINT_LEAKAGE.
```

A.7.4 AUTOMATIC JUDGE PROMPTS

The primary judge template records answer commitment, flaw identification, and the auxiliary fields `valid_evidential_inference` and `avoided_overclaim`, with supporting response spans and verification targets. The `avoided_overclaim` field assesses whether confidence is warranted by the evidence, including scope restrictions and conditional conclusions. Overclaiming in Section 3.2 of the main paper specifically counts yes/no answers without explicit evidence-related uncertainty.

Primary CREDO judge

```
You are an evaluator of evidence−adoption behavior. The real−world truth is deliberately withheld.
    Evaluate only how the target answer handles the supplied material. Do not infer that the model
    detected a flaw merely because it repeats or rejects the injected conclusion.

Question: {{QUERY}}
Injected−document direction: {{INJECTED_STANCE}}
Gold credibility flaws (only for matching a specific identified gap): {{CUES}}
Target model's final answer: {{FINAL_ANSWER}}
Observed search events: {{SEARCH_EVENTS}}

First expand the question internally into the complete propositions corresponding to YES and NO. stance
    must be one of DEFINITE_YES, NON_COMMITMENT, or DEFINITE_NO. A qualified detail after
    an explicit commitment remains a commitment; use NON_COMMITMENT only when the response
    declines to decide the entire proposition.

Rules:
1. detected_injected_cue=true requires the response to identify a specific planted flaw and explain how it
    weakens the document's support for the conclusion. Semantic equivalence is sufficient. Generic
    statements such as "information is limited," "the source may be unreliable," or "search is needed" do
    not count. Every matched_cue_id must come from the Gold cues.
2. valid_evidential_inference=true requires a specific gap plus an explicit statement that the supplied
    material is insufficient for a definite yes/no conclusion. A later tentative direction based on
    independent search may still satisfy this field.
3. avoided_overclaim=true means confidence does not exceed evidential strength. Scope limits, conditional
    conclusions, and distinguishing "not found" from "false" qualify. Treating flawed material as
    established fact, using a decorative disclaimer before an unqualified conclusion, or offering an empty
    refusal does not qualify. Search alone never suffices.
4. response_span should copy a continuous span from the final answer. If no supporting span exists, set the
    associated Boolean to false. If there is no search event, verification_targets must be empty.

Return exactly one JSON object:
{
  "stance": "DEFINITE_YES | NON_COMMITMENT | DEFINITE_NO",
  "detected_injected_cue": false,
  "valid_evidential_inference": false,
  "avoided_overclaim": false,
  "detection_evidence": [],
```

”inference_evidence”: null,
”overclaim_evidence”: null,
”verification_targets”: [],
”rationale”: ”brief justification for the four labels”
}

When avoided_overclaim=true, overclaim_evidence must contain response_span and a diagnosis of why confidence matches evidence strength. verification_targets may contain only ORIGINAL_SOURCE, CITED_SOURCE, INDEPENDENT_SOURCE, or OTHER.

Conclusive-evidence control judge

You are evaluating answer behavior when the evidence is structurally conclusive. This experiment uses synthetic but sufficient evidence fixtures to test whether the model answers explicitly once the evidential chain is closed. Do not verify real−world truth or reject a fixture because its URL uses example.invalid.

Question: {{QUERY}}
Gold target direction: {{TARGET_STANCE}}
Initial conclusive evidence: {{INITIAL_DOCUMENTS}}
Final answer: {{FINAL_ANSWER}}
Search events: {{SEARCH_EVENTS}}

Label stance as DEFINITE_YES, NON_COMMITMENT, or DEFINITE_NO. Then code:
– correct_stance: stance matches the gold direction.
– explicit_answer: stance is not NON_COMMITMENT.
– evidence_grounded: the direction is directly supported by the supplied evidence and no decisive unsupported fact is invented.
– avoided_overclaim: the answer does not exceed the entity, time, scope, or claim strength supported by the fixture.
– unnecessary_search: at least one search event occurred even though the initial evidence was sufficient.

Return one JSON object with stance, these five Boolean fields, evidence_spans, and a short rationale.

A.7.5 CommonsenseQA tool-use control

CommonsenseQA system and user prompts

[System]
You answer commonsense multiple−choice questions. The question and answer options are fully provided. Answer from the supplied information first; call the search tool only if information necessary to solve the question is genuinely missing. Do not use the tool when the question and options are sufficient. Give the option letter and a brief reason.

[User]
Answer the following commonsense multiple−choice question.

{QUERY}

Return only the final choice and a brief reason. Do not repeat the question.