Evidence Over Memory: Do Long-Context LLMs Reason from What They Read?

Anonymous authors
Paper under double-blind review

Abstract

Long-context models can read an entire scientific paper, but do their conclusions actually depend on what they read? We introduce MainLogic, a benchmark that tests selective responsiveness: a model should change its judgment when decisive evidence changes, withhold judgment when evidence becomes insufficient, and remain stable when irrelevant content changes. Starting from human-verified scientific citations, we construct matched families that systematically vary the claim and its evidential basis while preserving the underlying scientific problem. Across ten frontier models, performance on natural claims substantially overstates evidence-grounded reasoning: models reaching over 90% accuracy on the original claims correctly solve at most 32.4% of complete evidence-intervention families. Moreover, when supplied evidence conflicts with a model's independently measured prior, accuracy drops by 6–51 points, and errors systematically drift toward that prior even when the evidence is insufficient. MainLogic reveals a gap between reading evidence and letting evidence determine the answer.

1 Introduction

Recent advances in language modeling have rapidly expanded the context windows of large language models, enabling them to process entire research papers, large document collections, and other evidence-rich inputs in a single interaction (Chen et al., 2024; Ding et al., 2024; Munkhdalai et al., 2024). This progress has already prompted extensive study of long-context capabilities, including whether models can retrieve relevant information, maintain coherence over long inputs, and reason across widely separated pieces of evidence (Shaham et al., 2022; Bai et al., 2024; Liu et al., 2024; Hsieh et al., 2024; Kuratov et al., 2024; Lin et al., 2026b). However, being able to process or retrieve information from a long context is not the same as having one's judgment actually determined by that evidence. When a model correctly endorses a certain claim, it is often unclear why it reached that conclusion. The model may have inferred the answer from the supplied sources, recalled the result from pretraining, or relied on a plausible prior association. When the supplied evidence happens to agree with the model's prior beliefs, all of these mechanisms can produce the same prediction(Neeman et al., 2023; Xie et al., 2024; Monea et al., 2024; Ming et al., 2025). Conventional accuracy therefore cannot determine whether the model's judgment is genuinely grounded in the provided evidence or would have remained unchanged even if that evidence were absent.

A stronger test is whether a model changes its judgment **only when the evidence warrants it**. If decisive evidence supporting method A over B is reversed, the judgment should flip from support to contradiction; if that evidence is removed, it should become undetermined. Yet edits to irrelevant wording should leave the judgment unchanged. We call this property *selective responsiveness*: sensitivity to evidence-changing interventions, invariance to irrelevant ones, and abstention when the evidence is no longer decisive. It rules out models that merely repeat familiar answers, react to any textual perturbation, or abstain indiscriminately. Genuine evidence grounding requires all three behaviours together. Selective responsiveness becomes especially challenging in long contexts, where the evidence governing a claim is often distributed across distant sections and must be integrated before its evidential force can be assessed. Scientific papers and their cited references provide a natural testbed for long-context reasoning: they are information-dense, logically structured, and contain explicit

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.

relationships between claims, methods, evidence, and conclusions distributed across long documents. This makes them particularly well suited for studying whether models can reason over context rather than merely retrieve isolated facts. Most scientific verification benchmarks use abstracts, selected passages, or pre-extracted claim–evidence pairs (Thorne et al., 2018; Wadden et al., 2020; 2022), removing the long-context challenge that general-purpose suites target (Shaham et al., 2022; Bai et al., 2024; Hsieh et al., 2024). Full-paper benchmarks restore this setting (Dasigi et al., 2021; Javaji et al., 2025), but still leave open whether model judgments are determined by the supplied evidence rather than by parametric knowledge or learned priors.

In this work, we introduce **MainLogic**, a benchmark for *anchor-matched, evidence-relative reasoning* in long-context language models. Each instance starts from a natural scientific citation: a claim is taken from a citing paper, while the full referenced works form the premises. Models must judge whether the claim is *supported*, *contradicted*, or *undetermined* by those premises, rather than assess its unrestricted real-world truth. Each anchor generates a matched *six-cell family*: two opposing claim orientations crossed with three premise states, original, opposing, and insufficient. The first two reverse the evidential relation, while the insufficient state deletes the located evidence for the claim so that neither orientation is warranted. A length-matched control deletes an equal volume of non-evidence text, leaving the evidential relation unchanged. Every anchor claim is manually human reviewed, and a sample of constructed variants is checked by a blinded human audit. We separately measure each model's evidence-free judgment, allowing us to test cases where evidence supports, contradicts, or is insufficient relative to its prior. We also ablate over three forms of input premise, unmarked documents, the same documents with the evidence highlighted in place, and the evidence passages alone, to separate the cost of finding evidence in a long input from the cost of reasoning from it.

Our contributions are threefold: **1. Evidence-guided scientific benchmark construction.** We ground matched interventions in manually reviewed natural claims and localized evidence from complete cited papers, with a staged protocol for checking the resulting relations. **2. A six-cell evaluation of selective evidence dependence.** We jointly test correct updating, correct stability, and appropriate withholding through relation-changing and relation-preserving variants derived from the same anchor. **3. A matched protocol for evidence, prior, and context analysis.** We combine independent prior measurement, evidence-access conditions and family-level metrics to examine when long-context judgments follow, or depart from the supplied evidence.

2 Related Work

Long-context understanding. A growing body of benchmarks evaluates whether language models can access and reason over increasingly long inputs. Early suites such as SCROLLS and LongBench cover long-document question answering, summarization, multi-document reasoning, and other realistic tasks (Shaham et al., 2022; Bai et al., 2024). Diagnostic benchmarks probe more specific limitations: Lost in the Middle studies positional sensitivity (Liu et al., 2024); RULER and BABILong extend needle-in-a-haystack evaluation toward multi-hop reasoning, aggregation, and distributed evidence (Hsieh et al., 2024; Kuratov et al., 2024); and ∞Bench evaluates reasoning beyond 100K-token contexts (Lin et al., 2026b). More recent benchmarks move further toward realistic understanding: LooGLE targets long-range dependencies in natural documents (Li et al., 2024), HELMET evaluates application-centered tasks at controlled lengths (Yen et al., 2025), LongBench v2 emphasizes difficult real-world problems requiring deeper reasoning (Bai et al., 2025), and NoLiMa removes lexical shortcuts between queries and relevant evidence (Modarressi et al., 2025). Scientific documents provide a particularly natural long-context setting: QASPER evaluates full-paper question answering (Dasigi et al., 2021), while CLAIM-BENCH studies claim–evidence extraction and validation across full research papers (Javaji et al., 2025). Collectively, these ask whether models can *locate, integrate, and reason over* information in long contexts. They do not, however, establish whether a resulting judgment is actually *determined by* that information rather than by knowledge or expectations the model already possesses.

Evidence-grounded reasoning. A complementary literature asks whether model judgments are supported by supplied evidence. FEVER and SciFact formulate verification in terms of support, contradiction, and insufficient evidence, with SciFact grounding scientific claims in abstract-level evidence (Thorne et al., 2018; Wadden et al., 2020). Contrastive benchmarks such as VitaminC and CrossAug test responsiveness to altered claim–evidence pairs (Schuster et al., 2021; Lee et al., 2021), while contextual-faithfulness work studies counterfactual or conflicting evidence and competition with parametric knowledge (Neeman et al., 2023; Xie et al., 2024; Monea et al., 2024; Ming et al., 2025). Evidence-sufficiency benchmarks further test whether models appropriately withhold answers when the supplied context is incomplete (Joren et al., 2025). MainLogic connects these questions to long-context scientific reasoning through controlled variants of the *same scientific problem*: decisive evidence can support, oppose, or become insufficient, while relation-preserving interventions leave the evidential conclusion unchanged. The resulting test is not simply whether models are sensitive to context, but whether their judgments move *when and only when* the evidence warrants it.

3 What Does It Mean to Reason from Evidence?

A correct answer to a natural scientific claim may be consistent with both the supplied evidence and the model's prior belief, making it unclear whether the judgment actually depends on what the model has read. MainLogic resolves this ambiguity through controlled, matched variants of the same scientific problem. A genuinely evidence-grounded model should update when the evidential relation changes, remain stable when it is preserved, and return *undetermined* when the available premises no longer justify either conclusion. These matched variants let us test evidence dependence directly across changes in evidence availability and conflict with measured prior beliefs.

Evidence-Relative Three-Way Judgments. An instance comprises a self-contained claim C and premises $P = (G, D_1, \ldots, D_k)$, where the D_j are complete cited documents and G is an optional explicitly supplied contextual premise. Let X denote the actual context shown to the model; the native full-document condition uses $X = P$. Extracted quotations are auxiliary annotations, not replacements for the complete premises when assigning the full-context label. A fixed annotation policy Π specifies the target assertion, its scope, conditions, population, and uncertainty, together with the background reasoning allowed in interpreting the sources. Let $w_\Pi(X, A)$ indicate whether X warrants assertion A under this policy. This is a source-relative evidential judgment, not a claim that an empirical result deductively proves an unrestricted scientific proposition. Define

$$R_\Pi(X, C) = \begin{cases} \mathsf{S}, & (w_\Pi(X, C), w_\Pi(X, \neg C)) = (1, 0), \\ \mathsf{C}, & (w_\Pi(X, C), w_\Pi(X, \neg C)) = (0, 1), \\ \mathsf{U}, & (w_\Pi(X, C), w_\Pi(X, \neg C)) = (0, 0), \end{cases} \tag{1}$$

where S, C, and U mean *supported*, *contradicted*, and *undetermined (insufficient evidence)*. The case $(1, 1)$ requires adjudication and is outside the three-way core. Support must cover the complete evaluated assertion: an unsupported detail alone does not establish contradiction. These final relations are distinct from the provisional citation-role labels used during collection. The prediction target is $R_\Pi(X, C)$, not the real-world truth of C. A true claim may be undetermined by the supplied source, and an internally coherent counterfactual source may support a globally false claim. In particular, U denotes an unresolved evidential relation, not low model confidence or low topical similarity. Linguistic and logical competence is permitted, but unprovided target-specific empirical facts cannot substitute for source evidence.

Six-Cell Evidence Dependence within a Natural Anchor. For a fully supported natural anchor (P_i^+, C_i^+), let C_i^- be a contrasting claim, P_i^- a counterfactual premise context, and P_i^0 a context that leaves both claim orientations unresolved. The matched core is

$$\begin{aligned} \mathcal{B}_i &= \{(P_i^r, C_i^b, Y_{rb}) : r \in \{+, -, 0\},\ b \in \{+, -\}\}, \\ \mathbf{Y} &= \big(R_\Pi(P_i^r, C_i^b)\big)_{r,b} = \begin{pmatrix} \mathsf{S} & \mathsf{C} \\ \mathsf{C} & \mathsf{S} \\ \mathsf{U} & \mathsf{U} \end{pmatrix}. \end{aligned} \tag{2}$$

Rows are ordered $(+, -, 0)$ and columns $(+, -)$. Superscript $-$ denotes a validated contrast, not blanket negation of a document. All six relations must hold against the complete supplied premises. If C_i^- is contrary to, rather than exactly the complement of, C_i^+, neither its label nor its undetermined status follows automatically from that of C_i^+. The first two rows test appropriate reversal: changing either decisive evidence or claim orientation reverses the label, while changing both restores support. The third tests appropriate withholding. For example, removing a necessary assumption may make a conclusion undetermined; the absence of that assumption does not establish the opposite conclusion. Relevant background may remain extensive in P_i^0. Let $\mathcal{T}_{ir}$ contain the identity and validated relation-preserving premise edits. These satisfy $R_\Pi(\tau(P_i^r), C_i^b) = Y_{rb} \quad \text{for every } \tau \in \mathcal{T}_{ir},\ b \in \{+, -\}$. Examples include rewriting introductory or concluding material and removing experiments that are nonessential to the target relation. Their status is defined by semantic effect, not by the operation's name. The extended family $\mathcal{F}_i = \{(\tau(P_i^r), C_i^b, Y_{rb}) : r \in \{+, -, 0\},\ b \in \{+, -\},\ \tau \in \mathcal{T}_{ir}\}$ retains the original anchor as its unit of analysis. A model is correct on this evaluated family when $f_\theta(X, C) = y$ for every $(X, C, y) \in \mathcal{F}_i$. An arbitrary prediction flip does not qualify, nor does an unchanged but incorrect prediction. Members are tested independently without sibling variants or augmentation metadata in the prompt. These constraints define observable behavior, not a certificate of an internal reasoning mechanism.

Evidence over Memory under Long-Context Access. Counterfactual status alone does not establish conflict with a particular model's prior. Let $b_\theta(C) \in \{1, 0, \perp\}$ denote a stable evidence-free factual judgment, measured using separate probes, with $\perp$ indicating no stable eligible judgment. Define $h_\theta(C) = \mathsf{S}$ when $b_\theta(C) = 1$ and $h_\theta(C) = \mathsf{C}$ when $b_\theta(C) = 0$. On prior-eligible instances, the evaluation distinguishes

$$\begin{aligned} \text{supporting:} \quad & R_\Pi(X, C) = h_\theta(C), \\ \text{contradicting:} \quad & R_\Pi(X, C) \in \{\mathsf{S}, \mathsf{C}\} \setminus \{h_\theta(C)\}, \\ \text{prior available, undetermined:} \quad & R_\Pi(X, C) = \mathsf{U}. \end{aligned} \tag{3}$$

The last condition asks whether a prior belief is imported into a context that warrants neither conclusion; it is not assigned a factual truth value. Here, "memory" is operationalized by a measured behavioral prior, not direct access to knowledge stored in parameters. All valid instances remain in the general evaluation, irrespective of prior eligibility. Evidence availability is a separate axis. For an instance (X, C, y), let $\rho_v(X, C)$ render its premise under condition v: the full documents, the full documents with evidence highlighted in place, or the evidence passages alone (oracle). A matched comparison requires $R_\Pi(\rho_v(X, C), C) = y$, for every evaluated v. Thus, interpretability and the target relation remain fixed while access and integration demands change. The central question is which of these demands, locating evidence or integrating it, weakens correct evidence dependence, particularly when the source conflicts with a measured prior or leaves a previously believed claim undetermined. Performance differences characterize these behaviors; they do not by themselves identify causal memory override.

4 MainLogic: Evidence-Guided Benchmark

4.1 Constructing the Benchmark

Construction proceeds from natural citation instances, which are augmented to construct *supported*, *contradicted* and *undetermined* logical propositions. Source collection and claim screening establish candidate premise–claim pairs; full-document evidence annotation and human review establish the natural anchors and guide interventions. The full cited texts remain the premises throughout, including after augmentation.

Step 1: Sources. From a configured list of arXiv identifiers we download each citing paper's LaTeX, inline its includes, and recover every citation with its position in the text. Bibliography entries are matched to Semantic Scholar records by arXiv identifier, DOI, or title, and the cited full text is retrieved from open-access sources (Appendix D.1). A cite key with no open-access full text is rejected.

Step 2: Anchor claims. Anchor claims are extracted using two methods. A filtering based method screens every citation-bearing sentence with two LLM filtering stages, one for keeping citations which strictly make logical claims (Cohan et al., 2019), and one for decontextualisation (Choi et al., 2021), keeping claims interpretable without additional context from the main paper. More details can be found on claim requirements in Appendix D.1. An alterative method for claim retrieval is direct human or LLM extraction, under a brief requiring standalone, logical claims (as in the filtering step).

Step 3: Human Review. A full human review triages all collected claims (P_i^+, C_i^+), ensuring they are all decontextualised, and logical claims following the claim specifications in Appendix D.1.

Step 4: Augmentation. A requirement for augmented premises P^-, is they must establish the contrasting claim C^-, while contradicting C^+. We therefore construct C^- first, under the negation requirements of Appendix D.2. The premise document P^+ is then edited to support C^-. For the implication to hold, the *coverage* of the $P^+ \rightarrow P^-$ edits must be sufficient that no remaining statement in P^- supports the original C^+.

A retrieval model, `GPT-5.6 Sol`, reads each cited document and extracts passages relating to C^+. An augmentation model, `GPT-6 Astra`, labels each as an *assertion* (it states the claim), a *premise* (the claim follows from it) or *unrelated*. It then chooses the proposition to change, writes C^- as per Appendix D.2, and rewrites the fewest passages needed for the augmented implication to hold. We enforce deterministic checks, such as length bound and no added sentences to drop poor edits. P^0 removes every assertion and premise passage from the original documents, so that neither claim is settled. A control $P^=$ removes a matched volume of non-evidence text at similar positions and leaves C^+ supported. Further details on augmentation process are in Appendix F.

Step 5: Human audit and release. A blinded audit samples rows by hash and walks each through five steps, described in Appendix E. The auditor answers with premises and claim only, is shown the reference label, then reviews the claim quality, the negation quality, and the coverage of the edit. The entire anchor is labeled *verified* if every audited row in it passes and *failed* if any audited row finds a defect. Over 48 rows from 40 anchors we found 2 bad claims, 3 weak negations and 1 row with competing evidence, leaving 34 anchors verified (258 of the 812 instances, 31.8%) and 6 failed.

4.2 MainLogic Dataset

As shown in Table 1, each instance pairs a one-sentence *conclusion* from a scientific paper with the *full text* of the works it cites as evidence, and asks for SUPPORTED, CONTRADICTED, or UNDETERMINED. Nothing is excerpted: annotated evidence spans run a median of 35 words inside 1–4 complete papers (mean 1.3). Premises reach a median 16,078 tokens (mean 24,640, p95 75,880, max 120,097; `cl100k_base`), with 51.2/24.6/8.4% of instances above 16K/32K/64K and 20.0M premise tokens in total; conclusions average 38.7 tokens. Long-document retrieval and aggregation are part of the task, not preprocessing.

The 812 instances come from 108 *anchors* (original (P^+, C^+) pairs) drawn from 27 source papers citing 121 works (465 distinct premise texts). Each anchor is expanded over four premise conditions $\times$ two conclusion polarities, and the cell fixes the label (Table 1); no annotator assigns one. 82 anchors admit all eight cells, 26 (no removable redundancy) only the core six: $82{\times}8 + 26{\times}6 = 812$, giving 298 supported, 298 contradicted, 216 undetermined. An UNDETERMINED class invites a shortcut: if abstention is required only where evidence was deleted, a model can abstain on the cue that text is missing. We remove that cue by construction. Over the 82 anchors carrying both, P^0 and $P^=$ delete indistinguishable fractions of the premise (Table 1), yet P^0 is UNDETERMINED while $P^=$ retains the unperturbed label. They differ only in *which* text was removed, so abstention cannot be produced by a "the premise got shorter" heuristic. $P^=$ is verifiably meaning-preserving: all 1,870 annotated evidence spans reappear verbatim at resolving offsets, whereas under P^- they differ from P^+ in 108/108 anchors.

Table 1: The cell determines the label and the instance count. P^0 and $P^=$ are the length-matched pair: equal-sized deletions with opposite label consequences, differing only in whether the deleted span carried the evidence ($P^=$ leaves every evidence span verbatim).

			Label (n)	
Premise	Edit to the cited text	Median edit	C^+ original	C^- negated (1 word)
P^+	cited papers verbatim	—	supported (108)	contradicted (108)
P^-	evidence reversed	34 words	contradicted (108)	supported (108)
P^0	evidence deleted	−7.0% chars	undetermined (108)	undetermined (108)
$P^=$	redundancy deleted	−7.1% chars	supported (82)	contradicted (82)

5 Evaluation Protocol

We structure the evaluation around three research questions, with the **natural anchor** (P^+, C^+) as the basic unit of analysis. We start by investigating accuracy within anchors to understand how the model tracks evidential relations, followed by a deep analysis on the models tendency to rely on prior belief, ending with ablations which explore the impact of evidence based reasoning under long-horizon. Each model is given the same system prompt, given we test model prior knowledge, we ablate performance across prompt strictness on using world-knowledge in Appendix H, finding that performance is mostly within noise for varying levels of prompt strictness. Confidence intervals use paired anchor-level resampling.

5.1 RQ1: Do Judgments Track Evidential Relations?

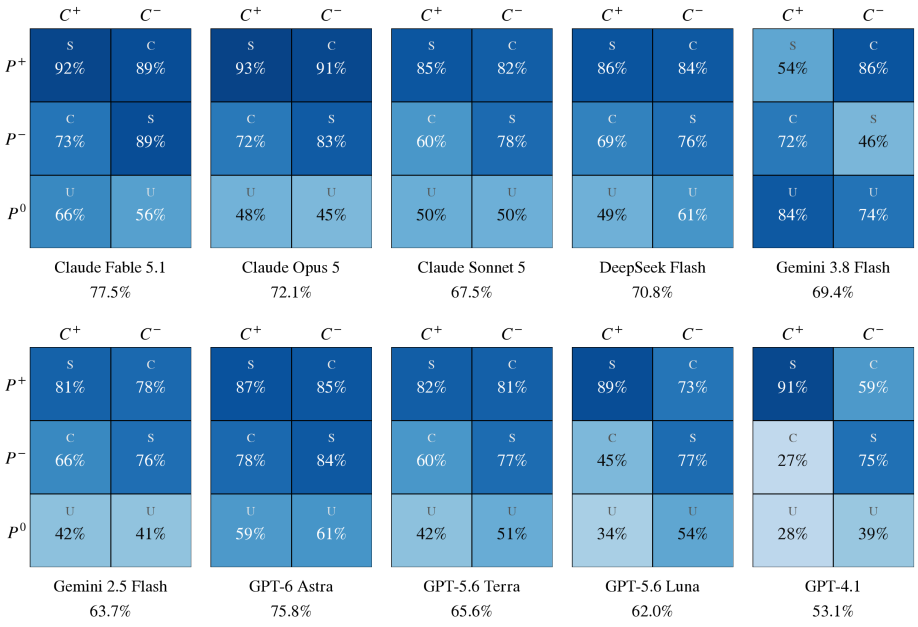


Figure 1: The six-cell core for each model, premise state down the side and claim orientation across the top. Each cell carries its accuracy, its reference label and its denominator.

We evaluate all six cells of every anchor and score them three ways. Holding the premise documents fixed and flipping the claim tests whether a judgment read the claim. Holding the claim fixed and varying the premise tests whether it followed the evidence. Requiring every cell of an anchor to be correct tests both at once. A fourth test asks whether a judgment tracks what a premise *contains* or only its *shape*. P^0 and $P^=$ delete the same volume from the same documents, and only P^0 deletes the evidence, so a system that answers U to both is reacting to the deletion rather than to what was deleted.

Let $a_{irb} \in \{0, 1\}$ indicate that the judgment on cell (r, b) of anchor i matches its reference label. Acc_6 is the mean of a_{irb} over all cells, and we report macro-F1 beside it. Three grouped scores demand correctness on a set of cells at once. $\text{GC}_{\text{premise}}$ is the fraction of rows on which both claims are correct, over $3N$ rows. GC_{claim} is the fraction of columns on which all three premises are judged correctly, over $2N$ columns. GC_6 is the fraction of anchors on which every cell is correct, $\frac{1}{N}\sum_i \prod_{r,b} a_{irb}$. The structure test is paired correctness on $(P^0C^+, P^{=}C^+)$ over the 82 anchors that carry both, which requires the right answer on each.

Results. Eight of ten models score above 82% on the natural (P_i^+, C_i^+) anchor (Figure 1). For those eight, Acc_6 falls by between 11 and 38 points once we perform augmentation (Table 2). GPT-4.1 is within two points of the best system on the natural anchor and last on Acc_6, where it must reason from evidence, by nine. Gemini 3.8 Flash is the sole system which struggles on the natural anchors themselves. The grouped scores ask a stricter question than Acc_6. A single cell can be right by disposition, but a row or a column is right only if the judgment moved with the evidence, and a core only if it did so in every direction at once. On this axis the benchmark is far from saturated. The best system completes fewer than half of its columns and under a third of its cores, and the weakest completes one core in thirty-six. The same measures that leave the strongest models room to improve separate strong from weak tenfold on GC_6, where Acc_6 separates them by twenty-four points.

Table 2: Performance within matched natural-anchor families. Nat. is accuracy on the natural cell (P_i^+, C_i^+) alone. A row is correct when both claims are, a column when all three premises are, a core when every cell is.

System	Nat.	Acc_6	Macro-F1	Grouped correctness		
				$\text{GC}_{\text{premise}}$	GC_{claim}	GC_6
Claude Fable 5.1	91.7	**77.5**	**77.0**	**68.2**	**43.5**	28.7
GPT-6 Astra	87.0	75.8	75.3	66.4	40.7	**32.4**
Claude Opus 5	**92.6**	72.1	70.8	62.3	31.5	21.3
DeepSeek Flash	86.1	70.8	70.4	57.1	33.3	19.4
Gemini 3.8 Flash	53.7	69.4	69.0	54.6	31.5	20.4
Claude Sonnet 5	85.2	67.5	66.8	55.0	28.3	17.9
GPT-5.6 Terra	82.4	65.6	64.8	51.5	24.1	13.0
Gemini 2.5 Flash	80.6	63.7	62.6	49.4	25.0	12.0
GPT-5.6 Luna	88.9	62.0	61.1	44.8	23.1	11.1
GPT-4.1	90.7	53.1	51.3	31.8	9.7	2.8

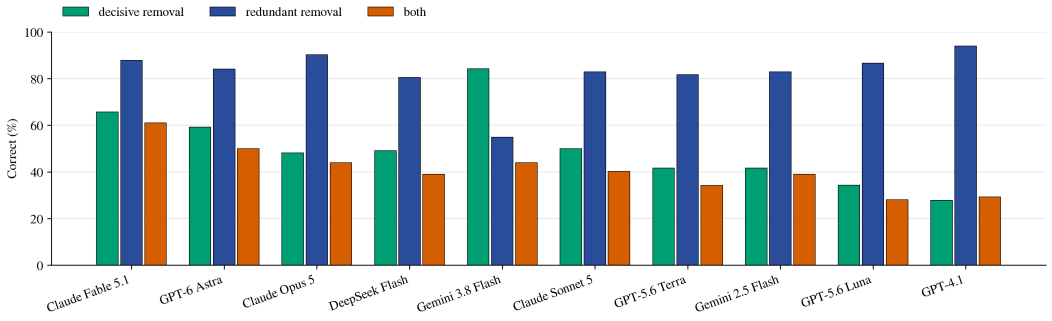


Figure 2: *Decisive removal* is P^0C^+, with the localized evidence cut and reference label U. *Redundant removal* is $P^{=}C^+$, with the same volume of non-evidence text cut from the same documents and reference S. *Both* is paired correctness, which requires both answers to be correct.

Structure against content. Nine systems read redundant removal better than decisive removal, most by a wide margin, which is the expected direction (Figure 2). Gemini 3.8 Flash inverts it, leading the benchmark where the evidence is genuinely gone and coming last where the deleted text was irrelevant. That is the signature of a judgment made on whether the evidence looks complete rather than on what it says, and no aggregate records it. The same system is the outlier in every other decomposition of the core. Appendix F.5 further explores how Gemini 3.8 Flash differs from the other models in how it trades commitment against hedging.

5.2 RQ2: When Do Prior Beliefs Compete with Evidence?

RQ1 evaluates judgments against the supplied evidence, but cannot distinguish evidence use from prior knowledge when the two agree. We therefore use evidence-free probes to separate alignment, conflict, and undetermined cases. Alignment and conflict are reported separately because pooling can hide strong but opposing prior effects; undetermined cases instead test whether models import a prior when the evidence supports neither conclusion. These prior-conditioned analyses are model-specific, reported with their denominators, and complement rather than replace the common-set RQ1 evaluation.

All probabilities use the anchor-balanced weighting defined above. On prior-eligible instances with determinate reference labels, write $\alpha = \mathbf{1}\{y = h_\theta(C)\}$ for the supporting and contradicting conditions of equation 3. We report three complementary rates,

$$\begin{aligned} \mathrm{Gap}_\theta &= \Pr\big(\widehat{y} = y \mid \alpha = 1\big) - \Pr\big(\widehat{y} = y \mid \alpha = 0\big), \\ \mathrm{ErrDir}_\theta &= \Pr\big(\widehat{y} = h_\theta(C) \mid \alpha = 0,\ \widehat{y} \neq y\big), \\ \mathrm{DirComp}_\theta &= \Pr\big(\widehat{y} = h_\theta(C) \mid y = \mathsf{U},\ \widehat{y} \in \{\mathsf{S}, \mathsf{C}\},\ b_\theta(C) \neq \perp\big). \end{aligned} \tag{4}$$

Gap_θ is the matched alignment gap: the accuracy difference between aligned and conflicting versions of the same claim, with the model and presentation fixed. It therefore measures, in percentage points, how much agreement with the model's own prior shifts accuracy. ErrDir_θ characterizes errors under conflict. Conditional on being wrong, the model either follows its prior or abstains to U. This separates error direction from error frequency: prior-directed errors indicate the prior overriding evidence, whereas U errors indicate conflict-induced abstention. $\mathrm{DirComp}_\theta$ measures whether determinate answers under insufficient evidence follow the model's prior. Conditioning on commitment removes variation in abstention propensity; 1/2 is the direction-uniform reference, with larger values indicating stronger prior-directed completion.

Memory versus label disposition. The matched contrast alone cannot separate prior reliance from a bare preference for one response label. A tendency to answer S lands on the correct label disproportionately in the aligned arm of claims the model believes true and in the conflicting arm of claims it believes false, meaning a fixed label preference can masquerade as an alignment gap wherever stable priors are imbalanced across P^+ and P^-. Table 3 separates the two explanations by recomputing the gap within each belief-sign group. A fixed label preference enters the two group gaps with opposite signs, the aligned arm of believes-true claims and the conflicting arm of believes-false claims share the same gold label, whereas prior reliance enters both with the same sign; their average, therefore, isolates memory, and their half-difference exposes the disposition. Across models, the memory share is uniformly positive while disposition shares scatter near zero with mixed signs, so the gaps of Table 4 generally track each model's own prior rather than a generic label preference.

Table 3: Alignment gaps of equation 4 recomputed within the n^+ claims the model believes true (prior S) and the n^- claims it believes false (prior C); memory and disposition are their half-sum and half-difference (50/50 group weighting). Intervals appear in Appendix I.5.

Model	n^+	true	n^-	false	memory	disposition
GPT-4.1	173	0.53	35	0.30	0.41	0.11
GPT-5.6 Luna	135	0.41	66	0.14	0.28	0.14
GTP-5.6 Terra	115	0.27	89	0.15	0.21	0.06
Claude Sonnet 5	116	0.23	65	0.12	0.18	0.06
Gemini 3.8 Flash	61	−0.14	80	0.39	0.12	−0.26
Deepseek Flash	90	0.13	60	0.19	0.16	−0.03
Claude Opus 5	97	0.15	96	0.09	0.12	0.03
Gemini 2.5 Flash	99	0.13	42	0.13	0.13	0.00
Claude Fable 5.1	94	0.14	89	0.03	0.09	0.06
GPT-6 Astra	91	0.11	91	0.01	0.06	0.05

Adjusted estimate. A logistic regression of correctness on alignment over prior-eligible determinate instances, with anchor-cluster-robust error terms and paired anchor-level resampling, finds the alignment coefficient positive for all ten models and significant for nine (Appendix I.5).

Results. Accuracy is higher on prior-aligned than prior-conflicting instances for all ten models, with gaps of 6–51 percentage points, statistically significant for nine (Table 4).

Models differ most under conflict: accuracy ranges from 39–83%, compared with 74–92% under alignment. When wrong under conflict, nine models more often repeat their measured prior than abstain. This tendency persists when evidence is insufficient: among determinate commitments, all ten models favor their prior, with confidence intervals excluding 1/2.

Notably, Fable and Astra have the smallest alignment gaps, yet 97% and 93% of these commitments, respectively, follow their priors. Thus, comparatively strong performance against conflicting evidence can coexist with strongly prior-directed errors when evidence warrants neither conclusion. These behavioral contrasts do not establish a mechanism of memory override; additional statistical details and model-specific analyses appear in Appendix I.5.

Table 4: Per-model prior-conditioned measures of equation 4 on n prior-eligible determinate instances. Intervals ($B = 2000$) appear in Appendix I.5.

Model	n	Gap_θ	ErrDir_θ	$\mathrm{DirComp}_\theta$
GPT-4.1	416	0.51	0.84	0.84
GPT-5.6 Luna	402	0.30	0.70	0.79
GPT-5.6 Terra	408	0.23	0.70	0.85
Claude Sonnet 5	363	0.18	0.84	0.80
Gemini 3.8 Flash	282	0.17	0.41	0.91
Deepseek Flash	300	0.16	0.73	0.78
Claude Opus 5	386	0.14	0.87	0.94
Gemini 2.5 Flash	282	0.14	0.74	0.71
Claude Fable 5.1	366	0.09	0.72	0.97
GPT-6 Astra	364	0.06	0.70	0.93

5.3 RQ3: How Does Long Context Change Evidence Dependence?

Design. We compare three renderings of each instance: (i) *native*, the full cited documents; (ii) *highlighted*, the same documents with the located evidence marked in place, which removes the search but keeps the length; and (iii) *oracle*, the located evidence alone. If highlighting helps, the model was failing to find the evidence under long-context; if neither helps, the cost lies in using it. We report paired differences from native with anchor-level bootstrap intervals. Table 5 summarizes counterfactual premises (P^-), pooled over both claim orientations; original-premise and cell-level comparisons appear in Appendix I.5.

Results. *Easier access helps selectively, not universally.* On P^-, highlighting significantly improves GPT-4.1, GPT-5.6 Luna and Claude Fable 5.1, while remaining models benefit less from simplifying the information location axis. Consistent with Appendix F.5, Gemini 3.8 Flash loses performance on oracle, likely because shorter premise pushes it toward default behavior.

Table 5: Counterfactual-row accuracy under three renderings (H: highlighted, O: oracle), pooled over both claims. *Prior* is evidence-free support rate; gains are relative to native, with **bold** as bootstrap intervals excluding zero.

Model	Prior	Accuracy on P^-			Gain over native	
	says S	native	H	O	H	O
GPT-4.1	91.7	50.9	56.9	59.3	+**6.0**	+**8.3**
GPT-5.6 Luna	86.1	61.1	68.1	68.1	+**6.9**	+**6.9**
GPT-5.6 Terra	86.1	68.5	71.8	71.8	+3.2	+3.2
Claude Opus 5	83.3	77.8	82.4	82.9	+**4.6**	+**5.1**
Claude Fable 5.1	81.5	81.0	81.0	83.8	+0.0	+2.8
GPT-6 Astra	77.8	81.0	84.7	82.4	+3.7	+1.4
Claude Sonnet 5	75.0	68.8	67.0	69.3	−1.9	+0.5
DeepSeek Flash	59.3	72.2	74.1	73.6	+1.9	+1.4
Gemini 2.5 Flash	53.7	70.8	71.8	66.7	+0.9	−4.2
Gemini 3.8 Flash	50.0	59.3	61.6	44.4	+2.3	−**14.8**

6 Conclusion

Long-context models can read the evidence without necessarily letting it govern the answer. MainLogic makes this distinction explicit by testing whether models *update when evidence changes, remain stable when it does not, and withhold judgment when evidence becomes insufficient.* Across ten models, high accuracy on natural scientific claims masks a much weaker ability to satisfy these requirements jointly: judgments degrade when evidence conflicts with measured priors, and easier access to decisive passages only partially closes the gap. The central lesson is that *having evidence in context is not the same as reasoning from it.* Reliable long-context reasoning therefore requires more than finding the right passage or producing the right answer, it requires conclusions whose behavior is systematically determined by what the evidence actually warrants.

AI use statement

In this work, we used generative AI tools to edit and review our writing, and coding agents to help build the benchmark's construction pipeline and evaluation code. We did not use generative AI tools to come up with the research questions, the benchmark design or the analysis. Separately, language models are part of the method itself: they locate evidence and build the counterfactual premises (Section 4.1). We reviewed all AI-assisted work. The authors checked every suggested wording change against the intended claims before accepting it. The authors also read, tested and ran the agent-written code, and checked its outputs against a blinded human audit of sampled rows (Appendix E.2). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

We do not identify any ethical concerns with this work. MainLogic is built only from scientific papers that are freely accessible: the citing papers come from arXiv, and the works they cite are retrieved only when openly available, via arXiv, Semantic Scholar and other open-access sources.

Reproducibility statement

We release everything needed to rebuild MainLogic and reproduce every reported number. The full benchmark (all 812 instances with premise documents, claims, cell assignments, reference labels, the 15,252 located evidence spans with verified offsets, construction route, and audit status) is available at `https://huggingface.co/datasets/anonymised-xxx-temp/mainlogic`, and the construction pipeline, evaluation harness, and analysis scripts are at `https://anonymous.4open.science/r/MainLogic-9830`. Full implementation details are given in the appendix: task formalization and proofs (App. C), claim and negation rules (App. D), human review protocol (App. E), the construction pipeline (App. F), dataset statistics (App. G), prompts (App. H), and statistical analyses (App. I).

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks, 2025. URL `https://arxiv.org/abs/2412.15204`.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models, 2024. URL `https://arxiv.org/abs/2309.12307`.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461, 2021. doi: 10.1162/tacl_a_00377. URL `https://aclanthology.org/2021.tacl-1.27/`.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. 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. 3586–3596, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1361. URL `https://aclanthology.org/N19-1361`.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

John Dougrez-Lewis, Mahmud Elahi Akhter, Federico Ruggeri, Sebastian Löbbers, Yulan He, and Maria Liakata. Assessing the reasoning capabilities of LLMs in the context of evidence-based claim verification. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20604–20628, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1059. URL https://aclanthology.org/2025.findings-acl.1059/.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Shashidhar Reddy Javaji, Yupeng Cao, Haohang Li, Yangyang Yu, Nikhil Muralidhar, and Zining Zhu. Can AI validate science? benchmarking LLMs on claim →evidence reasoning in AI papers. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2355–2379. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.127. URL https://aclanthology.org/2025.ijcnlp-long.127/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of llms with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106519–106554, 2024.

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. CrossAug: A contrastive data augmentation method for debiasing fact verification models. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 3181–3185. Association for Computing Machinery, 2021. doi: 10.1145/3459637.3482078. URL https://arxiv.org/abs/2109.15107.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16304–16333. Association for Computational Linguistics, 2024.

Tianwei Lin, Zuyi Zhou, Xinda Zhao, Chenke Wang, Xiaohong Li, Yu Chen, Chuanrui Hu, Jian Pei, and Yafeng Deng. Beyond the needle's illusion: Decoupled evaluation of evidence access and use under semantic interference at 326M-token scale, 2026a. URL https://arxiv.org/abs/2601.20276. arXiv preprint.

Tianwei Lin, Zuyi Zhou, Xinda Zhao, Chenke Wang, Xiaohong Li, Yu Chen, Chuanrui Hu, Jian Pei, and Yafeng Deng. Beyond the needle's illusion: Decoupled evaluation of evidence access and use under semantic interference at 326m-token scale, 2026b. URL https://arxiv.org/abs/2601.20276.

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/.

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", 2025. URL https://arxiv.org/abs/2410.03727.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. NoLiMa: Long-context evaluation beyond literal matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44554–44570. PMLR, 2025. URL https://proceedings.mlr.press/v267/modarressi25a.html.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6828–6844, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.369. URL https://aclanthology.org/2024.acl-long.369/.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention, 2024. URL https://arxiv.org/abs/2404.07143.

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

Jingxi Qiu, Zeyu Han, and Cheng Huang. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2605.03534. arXiv preprint.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3419–3425, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1341. URL https://aclanthology.org/D19-1341/.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52/.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized comparison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 12007–12021. Association for Computational Linguistics, 2022.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347/.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13300.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context language models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f5332c8273d02729730a9c24dec2135e-Abstract-Conference.html.

Hantian Zhang and Wentai Wu. Do LLMs know when evidence is insufficient? an evidence sufficiency benchmark for answer-abstention calibration in retrieval-augmented generation. *Computers, Materials & Continua*, 89(1):69, 2026. doi: 10.32604/cmc.2026.086343. URL https://www.techscience.com/cmc/v89n1/68467.

A Limitations

MainLogic currently draws on 108 anchors from 27 open-access arXiv papers, mostly in machine learning; because the pipeline is automated end to end, extending it to other fields and larger corpora is straightforward. Counterfactual premises are built by LLMs, and the blinded human audit covers 40 of the 108 anchors (34 verified). Priors are measured behaviourally, so prior-consistent errors show an association rather than a mechanism of memory override.

B A more detailed comparison of our work with related works.

FEVER formalized large-scale verification with support, refutation, and insufficient-information labels and accompanying evidence (Thorne et al., 2018). SciFact brought this formulation to scientific claims, using expert-written claims derived from citation sentences and abstract-level evidence with annotated rationales (Wadden et al., 2020); SciFact-Open extended verification to a substantially larger abstract collection (Wadden et al., 2022). These datasets establish natural citation-based claims and source-relative judgments as foundations for scientific evidence evaluation. Beyond abstracts, QASPER evaluates evidence-grounded question answering over research papers (Dasigi et al., 2021), while CLAIM-BENCH evaluates the extraction and validation of claim–evidence relationships within full-length scientific papers (Javaji et al., 2025). RECV additionally analyzes the atomic reasoning requirements of evidence-based claim verification (Dougrez-Lewis et al., 2025). MainLogic focuses on claims made by citing papers and the complete works they reference. Its analysis follows each natural claim through matched changes to its evidential basis, including assumptions, theoretical connections, and experimental comparisons, rather than treating the original verification decision as the sole outcome.

Correct predictions can coexist with reliance on dataset-specific cues: Schuster et al. (2019) studied claim-only biases and constructed a symmetric fact-verification test set. VitaminC uses contrastive evidence, including factual Wikipedia revisions, to test sensitivity to small changes that alter whether a claim is supported (Schuster et al., 2021). CrossAug is particularly close to our augmentation: it generates contrasting claims and evidence and cross-combines them with the originals to produce supporting and refuting training pairs (Lee et al., 2021). The broader principle of jointly testing directional changes and invariance is also central to CheckList (Ribeiro et al., 2020). Building on these ideas, MainLogic uses a validated six-cell evaluation family: two claim orientations crossed with supporting, opposing, and insufficient premise states. It additionally checks correctness under relation-preserving rewrites and nonessential deletions. The focus is joint behavior on complete scientific premises, not cross-pairing itself; paired correctness distinguishes appropriate evidence sensitivity from indiscriminate reactions to editing.

DisentQA separates contextual and parametric answers through counterfactual question answering (Neeman et al., 2023). Xie et al. (2024) elicit model-specific parametric knowledge and construct conflicting evidence to investigate receptivity and confirmation bias. Fakepedia combines counterfactual contexts with analyses of where grounding occurs inside a model (Monea et al., 2024). FaithEval broadens contextual-faithfulness evaluation to unanswerable, inconsistent, and counterfactual inputs, with model-assisted and human validation (Ming et al., 2025). Building on these investigations of contextual faithfulness, MainLogic asks how measured prior conflict interacts with evidence sufficiency and access *within the same scientific problem*. Separate evidence-free factual probes identify model-specific alignment; counterfactual status alone is not treated as proof of a conflicting prior. We distinguish failure to follow opposing evidence from importing a prior into an unresolved context, and interpret both as behavioral outcomes rather than direct identification of an internal memory mechanism.

The Attributable to Identified Sources framework evaluates whether generated statements can be verified against provided sources, separating attribution from unrestricted factual correctness (Rashkin et al., 2023). Sufficient Context distinguishes missing information from

failure to use available information and studies selective answering (Joren et al., 2025). More recent work also directly addresses these boundaries: the SURE-RAG preprint formulates set-level support, refutation, and insufficiency for selective answering (Qiu et al., 2026), and the Evidence Sufficiency Benchmark evaluates answer–abstention behavior across supportive, partial, irrelevant, absent, and conflicting contexts (Zhang & Wu, 2026). MainLogic retains the three-way source-relative distinction rather than treating uncertainty as a model confidence score. Its matched necessity controls compare deletion of indispensable information with deletion of redundant content; validated shuffled pairs provide a separate insufficiency source. This design asks whether a model recognizes a missing evidential dependency while remaining correct when only nonessential text is removed, rather than rewarding blanket abstention after any deletion.

Lost in the Middle investigates sensitivity to the position of relevant information within long inputs (Liu et al., 2024). HELMET provides application-centered evaluation with controlled context lengths (Yen et al., 2025), while LongBench v2 emphasizes deeper understanding and reasoning across realistic long-context tasks (Bai et al., 2025). NoLiMa reduces literal matching between questions and evidence, requiring latent associations to locate relevant information (Modarressi et al., 2025). The recent EverMemBench-S preprint further separates evidence access from end-to-end use under semantic interference (Lin et al., 2026a). These benchmarks already extend beyond simple span retrieval. MainLogic complements them by comparing the same validated claim relations under oracle evidence, complete cited documents, and relation-preserving length or layout conditions. The resulting question is not merely whether performance declines with length, but whether reduced evidence accessibility changes appropriate updating, appropriate withholding, and errors associated with independently measured priors.

C Semantic Details and Properties of Matched-Family Evaluation

This appendix clarifies the boundary conditions of Section 3 and derives elementary properties of its matched evaluation. The results are conditional on valid reference relations; they do not establish the validity of constructed documents or identify an internal reasoning mechanism.

C.1 Semantic and Construction Requirements

The domain of the three-way relation. The relation R_Π is defined on the adjudicated three-way core. If a context warrants both C and $\neg C$, the case requires adjudication under the fixed policy Π. A conflict that remains unresolved is excluded rather than mapped to U: incompatible warranted assertions are distinct from the absence of warrant for either assertion. Negation applies to the same scoped claim, including its conditions, population, and uncertainty. An observed experimental comparison and an unrestricted population-level assertion need not have the same evidential requirements.

Contrary claims and six-cell eligibility. equation 2 is a validation requirement, not a consequence of generating texts called "opposing" or "insufficient." For noncomplementary claim orientations, all six relations must be established separately. For example, let C^+ assert $a > b$ and C^- assert $b > a$ for the same precisely specified measured scores. Evidence establishing $a = b$ contradicts both claims; it does not make both undetermined. Thus, the absence of support for both orientations does not by itself establish the third row of the matrix.

Complete-premise validation. Every intended relation is evaluated against the complete resulting premise bundle, including G and every cited document. Removing a result from one location does not create insufficiency if another retained passage or cited work still warrants the conclusion. Reversing a result likewise requires checking for retained evidence that contradicts the intended new relation. These requirements apply to the source contents, not merely to the extracted evidence annotations.

Table 6: **How MainLogic relates to existing evaluation designs.** Prior work already supplies many individual components. MainLogic's focus is their joint application within the same natural scientific premise–claim family. Entries describe the cited protocols and our specified design, not a ranking of empirical performance.

Dimension	What prior work already provides	MainLogic's matched design
Scientific evidence	SciFact(-Open) verifies claims against abstracts; QASPER and CLAIM-BENCH use full research papers (Wadden et al., 2020; 2022; Dasigi et al., 2021; Javaji et al., 2025).	Citing claims paired with complete cited works. Original claims are manually reviewed; localized quotations annotate, rather than replace, the full premises.
Relation-changing tests	VitaminC uses factual revisions; CrossAug crosses original and edited claims and evidence for debiasing (Schuster et al., 2021; Lee et al., 2021).	Two fixed claim orientations crossed with three premise states: a validated six-cell supported/contradicted/undetermined core for each natural anchor.
Evidence sufficiency	Sufficient Context analyzes sufficiency; SURE-RAG tests set-level support, refutation, missing hops, and grouped evidence swaps (Joren et al., 2025; Qiu et al., 2026).	Remove necessary assumptions, theoretical connections, or comparisons; contrast with redundant deletions. Keep validated shuffled insufficiency cases separately identified.
Relation-preserving controls	CheckList includes invariance and directional tests; VitaminC also studies factual versus non-factual revisions (Ribeiro et al., 2020; Schuster et al., 2021).	Check paired correctness after relation-preserving introduction/conclusion rewrites or nonessential experiment removal, alongside decisive changes.
Evidence versus prior	Knowledge-conflict studies elicit model-specific priors; FaithEval also compares closed-book and counterfactual QA (Xie et al., 2024; Ming et al., 2025).	Measure an evidence-free behavioral prior; distinguish alignment, conflict, and prior-available but undetermined cases for the same claim.
Long-context access	HELMET controls length, NoLiMa reduces literal matching, and EMB-S separates evidence access from end-to-end use (Yen et al., 2025; Modarressi et al., 2025; Lin et al., 2026a).	Preserve each claim's relation across oracle evidence, complete premises, and matched length/layout views; examine prior-conflict costs under these access conditions.
Evaluation unit	Contrastive pairs and same-question grouped conditions already appear in VitaminC and SURE-RAG (Schuster et al., 2021; Qiu et al., 2026).	Aggregate within the original anchor: six-cell joint correctness, correct relation changes, correct invariance, and prior-by-context comparisons.

Missing exposition versus missing premises. Deletion is relation-changing only through its semantic effect under Π. If the retained premises include an assumption A and a rule $A \Rightarrow C$, and the policy permits the corresponding inference, removing an explicit sentence stating C does not remove its warrant. Likewise, deleting an intermediate proof step need not make a theorem undetermined when that step remains derivable. An insufficiency variant must remove an indispensable evidential dependency without leaving an alternative route to either conclusion. This need not remove all relevant background, and it does not imply that empirical warrant is deductive.

C.2 Measured Priors and Label-Preserving Access

Interpretation of a measured prior. The evidence-free probe asks for a factual judgment on the same scoped claim, rather than its support by an empty context. The value $b_\theta(C) = 0$ denotes a stable judgment accepting $\neg C$, not failure to recall or endorse C; uncertain or unstable responses are ineligible and receive $\perp$. Contrary orientations are probed separately,

because a prior for one does not in general determine a prior for the other. Eliciting model-specific judgments before testing conflicting evidence has precedent in knowledge-conflict evaluation (Xie et al., 2024).

The prompts, repetition count, decoding settings, and stability rule are part of the probing protocol and must be reported. The map h_θ expresses the measured factual orientation in the benchmark's label space; it does not make that prior a source-relative warrant. On an unresolved context, an incorrect prediction agreeing with $h_\theta(C)$ is a *prior-consistent error*, not proof that the prior caused the error. Prior eligibility does not require the prior to be factually correct. Model-specific eligible counts must accompany prior-conditioned results; comparisons using different eligible subsets need not compare the same scientific problems.

Valid access comparisons. The condition $R_\Pi(\rho_v(X, C), C) = y$ is checked against the actual rendered premises. Extraction, highlighting, truncation, and added material do not automatically preserve the label. Matched views must preserve the claim's scope, the referents needed to interpret retained evidence, and the reference relation. They need not preserve reading difficulty or interpretability, which are not consequences of label preservation.

For U instances, there is no sufficient-evidence oracle for a supported or contradicted answer. An evidence-localized view can still be evaluated, but it must preserve the unresolved dependency without introducing a label-bearing explanation, such as an annotation stating that the decisive result is absent. An instance without a valid matched view is omitted from that particular access comparison, not from the general evaluation. Empty or otherwise diagnostic views must not turn insufficiency into a formatting shortcut. Length and layout manipulations are likewise validated for their semantic effect.

C.3 Structural Properties of the Six-Cell Core

Consider N valid, complete six-cell families, evaluated with equal weight per anchor and per cell. For a deterministic classifier $f : \mathcal{X} \times \mathcal{C} \to \{\mathsf{S}, \mathsf{C}, \mathsf{U}\}$, define

$$\begin{aligned} c_i^{rb} &= \mathbf{1}\{f(P_i^r, C_i^b) = Y_{rb}\}, \\ A_i &= \frac{1}{6}\sum_{r,b} c_i^{rb}, & J_i &= \prod_{r,b} c_i^{rb}, \\ A &= \frac{1}{N}\sum_{i=1}^{N} A_i, & J &= \frac{1}{N}\sum_{i=1}^{N} J_i. \end{aligned} \tag{5}$$

Here A is cell accuracy and J is exact six-cell family accuracy. The following properties concern the balanced core, not an arbitrarily reweighted collection of additional controls or prior-eligible cells.

Proposition 1 (Blind baselines cannot satisfy the core). On every valid six-cell family: (i) a premise-blind classifier $f(X, C) = g(C)$ has $A_i = 1/3$ and $J_i = 0$; (ii) a claim-blind classifier $f(X, C) = q(X)$ has $A_i \le 2/3$ and $J_i = 0$; and (iii) a classifier that always returns U has $A_i = 1/3$ and $J_i = 0$.

Proof. For either fixed claim orientation, its three reference labels are a permutation of $(\mathsf{S}, \mathsf{C}, \mathsf{U})$. A premise-blind classifier produces the same label in all three cells, so exactly one is correct. Summing over the two orientations gives two correct cells out of six and rules out exact family correctness.

For a claim-blind classifier, the two predictions in each row coincide. In each of the first two rows, the labels differ, so at most one of the two predictions is correct. In the third row, at most both are correct. Thus at most $1 + 1 + 2 = 4$ cells are correct, and neither of the first two rows can be jointly correct. Finally, always returning U is correct exactly in the two cells of the third row. □

Stochastic extension. If the output distribution for a fixed claim is independent of the premises, its expected cell accuracy is also $1/3$. Indeed, writing $p_b(y)$ for that distribution,

the expected number correct in column b is

$$\sum_{r\in\{+,-,0\}} p_b(Y_{rb}) = p_b(\mathsf{S}) + p_b(\mathsf{C}) + p_b(\mathsf{U}) = 1. \tag{6}$$

A claim-blind output distribution similarly has expected cell accuracy at most 2/3, by the same row-wise argument. No independence between calls is needed for these expectation bounds. The zero exact-family accuracy statements in Proposition 1 apply to deterministic classifiers; stochastic classifiers can occasionally pass a family by chance.

Proposition 2 (Pointwise accuracy does not determine family correctness). For the fixed six-cell core,

$$\max\{0, 6A - 5\} \le J \le A. \tag{7}$$

In particular, accuracy $A = 5/6$ is compatible with $J = 0$.

Proof. A perfect family has $J_i = A_i = 1$. An imperfect family has $J_i = 0$ and at most five correct cells, giving

$$J_i \le A_i \le \frac{5 + J_i}{6}. \tag{8}$$

Averaging over anchors yields $J \le A$ and $6A - 5 \le J$. Combining the latter with $J \ge 0$ proves the bound. To obtain $A = 5/6$ and $J = 0$, consider a prediction pattern with exactly one error in every family. For comparison, when N is divisible by six, a pattern that is entirely correct on five sixths of families and entirely wrong on the rest has the same $A = 5/6$, but $J = 5/6$. Thus the same cell accuracy can correspond to very different matched-family performance. □

Reporting extended controls. Exact correctness on $\mathcal{F}_i$ additionally tests the specified relation-preserving edits. Because the number of such edits may differ across anchors, exact six-cell accuracy and extended-control performance should be reported separately, or control counts should be standardized. Incorrect invariance and arbitrary label changes do not count as correct responses, even when they satisfy an informal "same" or "different" pattern. Stochastic results additionally require a specified sampling and aggregation protocol.

Remark (Finite agreement does not identify a mechanism). Even perfect performance on a finite evaluation set does not uniquely identify the internal reasoning process. To see this, let $\mathcal{E}$ be the finite set of evaluated input pairs (X, C). Construct a lookup table storing $R_\Pi(X, C)$ for each pair in $\mathcal{E}$, and define a classifier to return the stored label when its input is in the table and an arbitrary label otherwise. The classifier agrees with every reference label on $\mathcal{E}$, but its implementation only performs input identification and label lookup, not an evidential inference. Hence agreement on these inputs cannot uniquely identify the internal mechanism producing it.

This existence argument does not assert that an evaluated model has memorized the benchmark. It limits the interpretation of the score: Proposition 1 rules out complete premise blindness for perfect *deterministic* family behavior, but context dependence alone does not establish dependence on the intended scientific evidence rather than textual artifacts or memorized input–label mappings. Held-out anchors, independent prompts, and construction-artifact checks address practical validity concerns; none turns a finite behavioral test into a mechanistic certificate. Likewise, differences between label-preserving access views or prior-conflict strata characterize behavioral differences without uniquely establishing retrieval failure or causal memory override.

D Scope of the Behavioral Conclusions

D.1 Claim Requirements

An anchor claim is a sentence from a citing paper that asserts something its cited works can establish, and that a reader can interpret with nothing else from the citing paper in front of them. Two screens enforce this, in that order, and neither sees the cited full text.

Claim type. A citation-bearing sentence is kept only if the citation is offered as *evidence* for what the sentence claims, and the whole claim is something works with those titles and years could themselves state or report. We call these *background* claims, following the citation function taxonomy of SciCite (Cohan et al., 2019), and they are the only type that reaches the benchmark.

> *Recurrent networks have been firmly established as state of the art approaches* [CITATION]
> *The base BERT model* [CITATION] *has 12 layers and a hidden size of 768*

Sentences whose citation plays any other role are dropped. A *method* citation names something the authors used or built on, and the sentence is about the authors, not the cited work.

> *We use the Adam optimizer* [CITATION]

A *comparison* reports the authors' own results against the cited work, so the evidence lives in the citing paper. An *entity* citation identifies something named in passing while the sentence asserts something the cited work does not report, such as a later trend. A sentence whose main claim, or a second claim of equal weight, goes beyond what the cited works could support is dropped as *partial*. The test in every case is whether a reader could verify the whole claim by reading only the cited works.

Context. A kept sentence must stand on its own. It is rejected if it contains a pronoun, demonstrative or definite phrase pointing at earlier text, a symbol or acronym it does not itself define and the field does not universally know, a continuation opening such as "where" or "Doing so", or a scope that the surrounding text sets and without which it reads as a broader claim. A sentence that reports what the citing paper itself did or found is rejected as *self* even when it is otherwise self-contained, since a first-person result sentence would pass every other test. No context is ever added to a claim. Sentences that need it are dropped, so every claim in the release is a single sentence and no row carries a passage from the citing paper.

Both screens are prompted LLM classifiers whose full instructions are in the released `data/README.md`. The direct-extraction path applies the same two tests by hand, under a written brief, and every claim on either path was then read by an author against its located evidence before augmentation.

D.2 Negation Requirements

A contrasting claim C^- is built from C^+ by changing one proposition and keeping every other word. It must be *contrary*, so that a reader who accepts the documents must reject it, and it must be *subtle*, so that a reader cannot tell which of the pair is original without the documents. The planner is given these rules.

What to change. Prefer, in order, a proposition the documents imply but never state, so that a premise can be changed and the consequence left for the reader to redo; a proposition a knowledgeable reader is confident about, so that C^- looks wrong from outside and only the documents say otherwise; and the smallest set of passages that must change. The change is recorded as one of six types, a *value*, an *entity*, a *polarity*, a *comparative*, a *quantifier*, or a *modal*.

What not to do. Never a bare "not". Do not target a value inside an "up to", "at least" or "more than" frame, since a different number there is not contrary. Avoid quantifier flips to universal negatives such as "no work has", which a single surviving mention refutes. Keep the original's hedging, scope and style. Prefer not to leave a sentence that simply announces C^-; if the claim's only proposition is stated outright, target it anyway, since a stated item is better than no item, and record the core as *stated* rather than *derived*.

How to edit. Touch as little as possible. A passage changes only if leaving it would make C^- false or leave the documents visibly contradicting themselves. One passage in, one passage out, with the same number of sentences, similar length, and the same register, wording and markup. Never add a sentence or a justification; the documents must state the new proposition no more plainly than they stated the old one. When a number changes, change the numbers that follow from it. Each rewritten passage must read as something that paper could have written. A rewrite that adds a sentence or grows by more than thirty percent is rejected in code before it is applied.

The same standard is what the human audit checks in Appendix E, where a reviewer who has just judged a C^- row is shown C^+ beside it and asked whether the pair reads as subtle, too easy, or broken.

E Details of Human Verifying

Two human reviews exist at different points of construction. The first reads every anchor claim before any augmentation. The second blind samples the finished release.

E.1 Stage one: every anchor claim

An author reads each collected claim against its cited works and the passages `sol` located in them, and answers a fixed checklist.

1. Is this a claim at all, rather than signposting, a definition or a table fragment?
2. Is it a *background* claim in the sense of Appendix D.1, something the cited works can establish, and not a description of the citing paper's own work?
3. Is it decontextualised, interpretable with nothing else from the citing paper?
4. Do the located passages establish it, so that the natural cell carries its label?
5. Can a contrary version be built, so that it will support a P^- row?

A claim failing any item is removed and does not enter augmentation. This is the only exhaustive human pass, and it is the reason the release contains no claim that is trivially true, mangled, or about the citing authors themselves. It is a review of claims, not of labels. No reference label is assigned by hand at any point; the cell fixes it.

E.2 Stage two: blinded sampled audit

After construction, rows are sampled by a hash of their identifier, so two auditors given the same seed see the same rows, and each row is walked through five steps in a local page. The steps are ordered so that the auditor commits to a judgment before seeing anything the pipeline decided.

1. **Answer blind.** The auditor sees the premise documents and the claim, and nothing else. Not the cell, not the reference label, not the sibling claim. They answer supported, contradicted or undetermined. An optional toggle highlights the located evidence, and whether it was used is recorded.
2. **Reveal.** The reference label is shown and the auditor's answer is recorded as right or wrong. A row humans get wrong is not thereby defective; the audit measures the construction, not the auditor.
3. **The claim.** Is this a real claim worth testing, something the cited work settles that takes a step of reasoning rather than a lookup? The reasons offered for *bad* are the stage-one checklist in reverse, not a claim, trivial, not background, needs context, vague, mangled. A bad claim retires all cells of the core.
4. **The negation**, on C^- rows. C^+ and C^- are shown side by side. The question is not whether C^- is false but whether a reader could tell which is original without the documents, the standard of Appendix D.2. The auditor chooses *subtle but critical*,

it reads naturally and only the documents settle it; *too easy*, a flip visible without reading the premises; or *broken*, not actually contrary or nonsense. Easy or broken retires the three C^- cells and leaves the C^+ cells.

5. **Coverage**, on P^- and P^0 rows. Both claims and the reference label are shown, with the located evidence highlighted, and the auditor scans the documents for anything that fights the label. On P^- that is a passage still supporting C^+; on P^0 it is anything left that settles either claim. Any such passage is selected and stored verbatim. *Competing* retires the two cells that share that premise; *clean* and *borderline* do not.

F Details of Augmented Process

Augmentation follows natural-instance review and export. Each eligible anchor (P_i^+, C_i^+) provides the starting point for a matched family satisfying equation 2. Every generated variant retains its connection to the original scientific problem. The intended relations guide construction, but the transformation type alone does not establish a reference label: acceptance requires validation against the complete resulting premises under Π.

F.1 Models and the two-pass construction.

Construction uses two models, neither of which sees the benchmark prompt. `GPT-5.6 Sol` (the reader) is the only model that reads whole papers. `GPT-6 Astra` (the planner) sees only the passages the reader returns. Keeping these roles apart means the expensive full-text reads go to the reader, while the planner's judgement is spent on a short list of candidates. Cores whose cited documents together exceed about 350k tokens are not built.

Pass 1: read and plan. The reader is called once per cited document with the claim C^+ and the full text. It quotes every passage that states the claim, the settings, values or results the claim rests on, restatements of it, and measurements whose numbers would have to differ if the claim were false. It is told to include a passage when unsure, because a passage it leaves out can never be edited. Each quote is located verbatim in the source, and a document that could not be read is recorded.

The planner receives the claim and the numbered passages, each with its source document, and returns one plan: a role for every passage (*assertion*, *premise* or *unrelated*), the target proposition and its edit type, C^-, a rewrite for each passage that must change, the route (*derived* or *stated*), and two self-checks: whether the passages entail C^+ and whether the edited passages rule it out. Code then checks whether either claim is stated outright by a single sentence in the edited or original documents. If one is, the planner is asked once to retarget a proposition the documents imply but never state. If the retry is still stated outright, the original plan is kept and the core is recorded as *stated*. Every rewrite must add no sentence and grow by at most 30% (plus 30 characters); rewrites that break this are rejected in code and logged. A core fails if the planner declares no coherent contrary possible, if C^- equals C^+, or if no passage changes.

Pass 2: check and repair. The reader re-reads each edited document in full, given both reports at most five passages that would strike a reader as *flatly* inconsistent with C^-. These are passages that still state the original value, direction or property, or numbers that contradict the edited ones. Indirect tension and background are deliberate and a document the reader cannot read counts as flagged. The planner rewrites the flagged passages in one round, shown the edits and held to the same in-place bound. Flags that survive the repair are stored with the core, and its two P^- cells are marked u This pass is intentionally light. It catches blatant misses cheaply and leaves arguable coverage to the human audit.

Evidence removal and its control. P^0 is built without a model. Every passage labelled assertion or premise is cut from the original documents, including every copy of it. P^0 has no second pass, so its standard is removal of the located evidence, not the absence of every trace of the topi anything left still settles either claim is checked by the audit. $P^=$ is also

deterministic span cut from P^0, it removes a comparable volume of text from the same document at a similar relative position, touching only located passage overlaps. The two variants therefore differ in *what* was removed, not in how much or where. $P^=$ i the same assumption as P^0, namely that the located evidence is complete.

F.2 Claim contrast and evidence reversal.

The claim transformation constructs C_i^- by changing a decisive assertion while retaining the remaining meaning, scope, and evaluation conditions as far as possible. The premise transformation constructs P_i^- by changing the evidence governing that assertion. This produces a coherent counterfactual premise context, not a negation of the entire document or of every assertion in the cited works.

Crossing the two claim orientations with the original and opposing premises produces four candidate cells. The accepted relations must satisfy

$$R_\Pi(P_i^+, C_i^+) = \mathsf{S}, \quad R_\Pi(P_i^+, C_i^-) = \mathsf{C},$$
$$R_\Pi(P_i^-, C_i^+) = \mathsf{C}, \quad R_\Pi(P_i^-, C_i^-) = \mathsf{S}.$$

Each relation is checked independently rather than assigned from the presence of a negation, changed number, or reversed comparison. In particular, contrasting claims need not be exact logical complements, so validating one orientation does not automatically validate the other.

F.3 Decisive-evidence removal.

An insufficient context P_i^0 is constructed by removing information necessary to settle the target claim, such as an assumption, theoretical connection, or decisive experimental comparison. Acceptance requires

$$R_\Pi(P_i^0, C_i^+) = R_\Pi(P_i^0, C_i^-) = \mathsf{U}.$$

This requirement is stronger than merely removing the originally localized evidence. Review considers all remaining documents and the contextual premise G, including repeated results, summaries, and alternative evidential routes. If the relation remains established by another passage or by reasoning permitted under Π, the deletion does not justify an undetermined label.

Neither missing support nor failure to establish one claim orientation constitutes evidence for the other. Conversely, lack of support for both orientations is not sufficient when the remaining premises contradict them. For example, a reported tie under the same evaluation setting may contradict both "A outperforms B" and "B outperforms A," rather than leave both undetermined. Each orientation therefore requires its own three-way assessment.

Deletion must also preserve claim interpretability, including the referents and conditions needed to understand the assertion. Removing this information makes a candidate unsuitable rather than a valid insufficiency example. Extensive relevant background may remain; the intended change is the loss of a necessary evidential dependency, not simply a reduction in document length.

F.4 Horizon Ablations

The natural row is unaffected. Table 7 repeats the comparison of the main text on P^+. No system outside the Gemini family moves by more than 2.3 points under either rendering, in either direction, except `gemini-3.8-flash` under the oracle, which is its hedging and not the horizon. Where the model already knows the answer, removing the search removes nothing it was doing.

The counterfactual row by cell. Table 5 pools the two claims. Table 8 separates them, and shows that the gains of the main text sit almost entirely on P^-C^+, where the reference label contradicts the claim the model remembers.

Table 7: The natural row under three renderings, pooled over both claims and paired on shared rows, in the layout of Table 5.

Model	Prior	Accuracy on P^+			Gain over native	
	says S	native	highlighted	oracle	highlighted	oracle
GPT-4.1	91.7	75.0	75.5	75.9	+0.5	+0.9
GPT-5.6 Luna	86.1	81.0	80.1	80.1	−0.9	−0.9
GPT-5.6 Terra	86.1	81.9	83.3	79.6	+1.4	−2.3
Claude Opus 5	83.3	91.7	91.7	91.7	+0.0	+0.0
Claude Fable 5.1	81.5	90.3	92.1	92.6	+1.9	+2.3
GPT-6 Astra	77.8	86.1	87.0	85.2	+0.9	−0.9
Claude Sonnet 5	75.0	83.8	85.6	84.3	+1.9	+0.5
DeepSeek Flash	59.3	85.2	85.6	83.3	+0.5	−1.9
Gemini 2.5 Flash	53.7	79.2	81.5	70.4	+2.3	**−8.8**
Gemini 3.8 Flash	50.0	69.9	73.6	53.2	**+3.7**	**−16.7**

Table 8: The counterfactual row under three renderings, one block per claim, paired on shared rows over 108 anchors. Within each block, native accuracy, the gain of each intervention over native, and oracle accuracy. *Prior* as in Table 5. Bold where an anchor-level bootstrap interval excludes zero.

Model	Prior	P^-C^+, reference C				P^-C^-, reference S			
	says S	native	Δhighl.	Δoracle	oracle	native	Δhighl.	Δoracle	oracle
GPT-4.1	91.7	26.9	**+12.0**	**+13.9**	40.7	75.0	+0.0	+2.8	77.8
GPT-5.6 Luna	86.1	45.4	**+10.2**	+5.6	50.9	76.9	+3.7	**+8.3**	85.2
GPT-5.6 Terra	86.1	60.2	**+10.2**	**+7.4**	67.6	76.9	−3.7	−0.9	75.9
Claude Opus 5	83.3	72.2	**+5.6**	+4.6	76.9	83.3	+3.7	**+5.6**	88.9
Claude Fable 5.1	81.5	73.1	+2.8	+4.6	77.8	88.9	−2.8	+0.9	89.8
Claude Sonnet 5	75.0	60.2	+0.9	+4.6	64.8	77.6	−4.7	−3.7	73.8
Deepseek Flash	59.3	68.5	−1.9	+0.9	69.4	75.9	+5.6	+1.9	77.8
Gemini 2.5 Flash	53.7	65.7	+2.8	+4.6	70.4	75.9	−0.9	**−13.0**	63.0
Gemini 3.8 Flash	50.0	72.2	+0.9	**−10.2**	62.0	46.3	+3.7	**−19.4**	26.9

The cost does not scale. For each system and each of the oracle and highlighted views we take the paired gain over native on every anchor, averaged over its determinate cells, and correlate it with three measures of horizon, the length of the premise in characters, the number of located evidence statements, and the share of the document the intervention changed. Figure 3 bins the same gains.

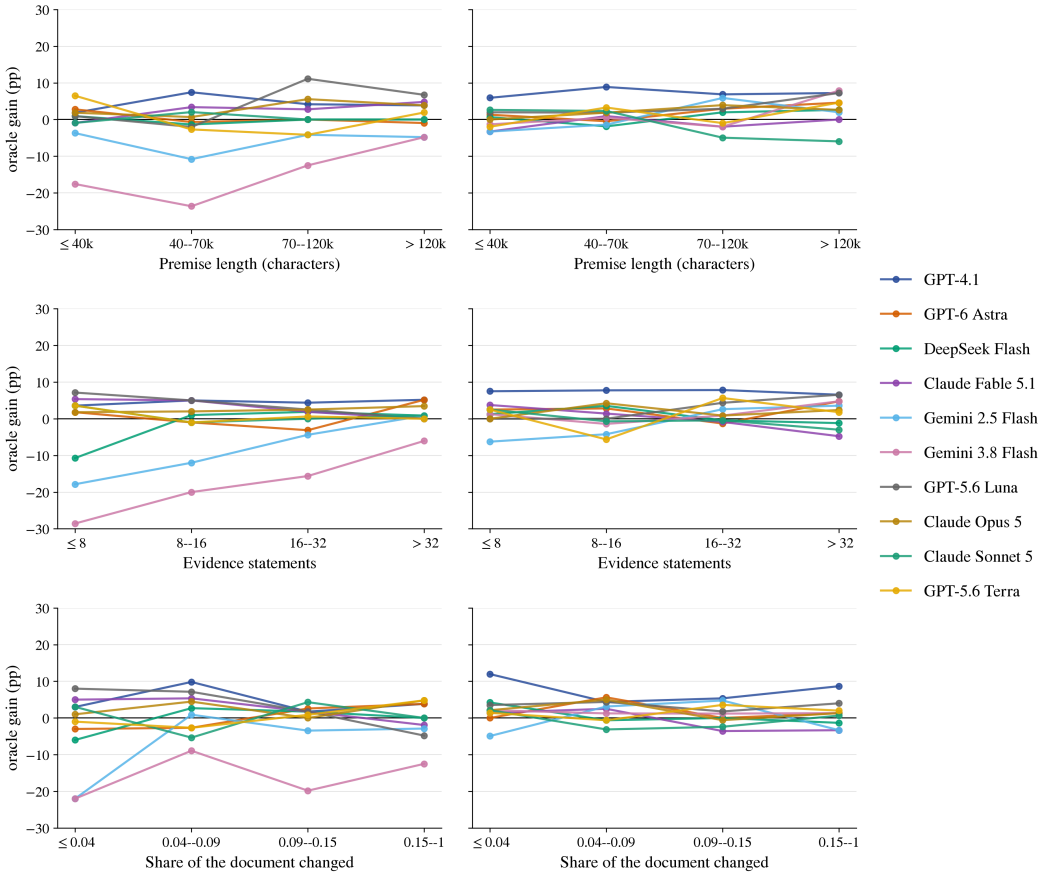


Figure 3: Per-system gain over native under the oracle (left) and highlighted (right) views, against three horizon measures, on the determinate cells. The black line is the mean.

Highlighting is a level effect. `gpt-4.1` gains six to nine points in every bin of every measure, `opus-5` and `luna` two to three, `sonnet-5` and `fable` nothing, and none of those levels changes with the horizon. The one system whose gain grows is `gemini-3.8-flash`, with length and with statement count, and it is the system that reads the natural cell worst. The oracle shows one trend, and it is the two `gemini` systems, whose oracle penalty is largest when there are fewest statements and vanishes above thirty. Handed three sentences and no paper they decline to commit. Handed forty they behave like everyone else. That is a disposition, and it is the same one Appendix F.5 measures under matched removal.

The regression separates what the bins conflate (Table 9). Document length has no effect on P^{+} or P^{0} and a small one on P^{-}. The number of evidence statements is the only property significant on every row, and it is strongest on P^{0}, where the more evidence a claim had, the less a system notices when it is gone. That effect survives in the oracle and is untouched by highlighting, so it is a cost of integrating evidence rather than of finding it. Whether the evidence had been in the abstract cuts both ways. It helps on P^{-}, in the oracle as much as natively, since abstract statements are cleaner evidence. It hurts on P^{0}, where a paper whose abstract has been cut still frames itself around the claim, and that framing is read as support.

Table 9: Logistic regression of correctness on premise properties, with model and cell fixed effects, one fit per row of the core and per view. Coefficients are log-odds per standard deviation; the abstract indicator is per unit. Anchors repeat across systems, so the standard errors are optimistic and the stars are an ordering rather than a test.

Premise property	native P+	native P-	native P0	oracle P-	highlighted P-
Document length ($\log_{10}$ chars)	$+0.00$	-0.21^{**}	$+0.01$	-0.07	-0.17^{*}
Evidence statements ($\log_2$)	-0.22^{*}	-0.17^{*}	-0.49^{***}	-0.15	-0.22^{*}
Fraction removed in P^0	$+0.16$	-0.20^{*}	-0.20^{*}	-0.17^{*}	-0.24^{**}
Evidence in the abstract	$+0.16$	$+0.42^{**}$	-0.31^{*}	$+0.38^{**}$	$+0.51^{***}$

F.5 Default Behaviour Study

Ten systems span twenty-four points of Acc_6, and the aggregate says almost nothing about how any of them gets there. We decompose the same six-cell core three ways, by what a system reaches for, by how it trades unwarranted commitment against unwarranted hedging, and by whether it can tell a missing dependency from a missing paragraph. Each decomposition recovers the same dispositions, and one system separates from the other nine under every one of them.

Measures. All measures use the six-cell core. Alongside per-cell accuracy and the three-class confusion matrix we report the model's own answer distribution over the same rows, the same distribution conditional on the answer being wrong, and two directional error rates,

$$\begin{aligned} \mathrm{FD} &= \Pr\big(\widehat{y} \in \{\mathsf{S}, \mathsf{C}\} \mid y = \mathsf{U}\big), \\ \mathrm{FU} &= \Pr\big(\widehat{y} = \mathsf{U} \mid y \in \{\mathsf{S}, \mathsf{C}\}\big). \end{aligned} \tag{9}$$

Probabilities use the anchor-balanced weighting defined above. FD measures unwarranted determination and FU measures failure to decide despite sufficient evidence. Invalid outputs are a separate error category, so FD is not assumed to equal one minus U recall. Error destinations are reported conditional on being wrong, which isolates the direction of failure from its frequency.

What each system reaches for. The core is balanced, so a system reading only the evidence would answer each label a third of the time. Nine of ten answer S more often than that, and the lean is strongest in the weakest systems (Table 10). Failures inherit the disposition and sharpen it. Every system that leans toward a label leans further toward it when wrong, so the answer distribution conditional on error is the unconditional one pushed outward, and being wrong never pulls a system toward the middle. Eight fail toward S. `GPT-6 Astra`, the flattest system in the table, fails toward C by a hair. `Gemini 3.8 Flash` is the mirror of the rest rather than an exception to the pattern. It answers U on nearly half of rows, and on nearly six in ten of the rows it gets wrong.

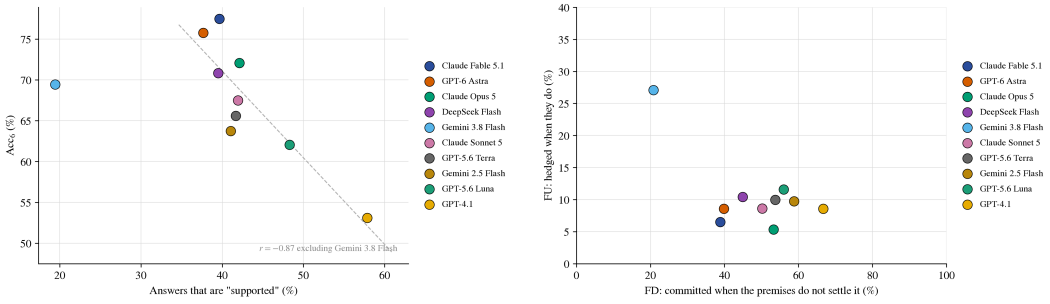
Commitment against hedging. The lean shows up as a trade (Figure 4). A system that endorses readily buys the supported cells and pays under FD, and a system that hedges readily makes the opposite bargain. Nine cluster at high FD and low FU, and within that cluster accuracy falls as the rate of answering S rises, with $r = -0.87$. One system sits alone in the opposite corner, committing least often where the premises settle nothing and hedging most often where they settle something. It is the same system that inverts the answer distribution, and it lies a point and a half from `DeepSeek Flash` on Acc_6 while running the opposite policy.

F.6 Whole-premise coherence and validity.

Validation covers the complete resulting premise bundle rather than only the edited passage. Relevant summaries, comparisons, and repeated occurrences must remain consistent with the intended change, including any information supplied through G. For example, reversing an experimental result while leaving a dependent summary asserting the original comparison does

Table 10: Acc_6 beside the distribution of answers, overall and conditional on the answer being incorrect. Bold marks a label a system reaches for more often in failure than overall. *Errors* names where a system's failures most often land.

Model	Acc_6	Answers			Answers when wrong			Errors
		S	C	U	S	C	U	→
Claude Fable 5.1	77.5	39.7	35.6	24.7	**42.5**	**38.4**	19.2	S 42
GPT-6 Astra	75.8	37.7	36.6	25.8	37.6	**38.9**	23.6	C 39
Claude Opus 5	72.1	42.1	38.7	19.1	**45.9**	**41.4**	12.7	S 46
DeepSeek Flash	70.8	39.5	35.2	25.3	**42.9**	33.3	23.8	S 43
Gemini 3.8 Flash	69.4	19.4	36.1	44.4	9.1	31.8	**59.1**	U 59
Claude Sonnet 5	67.5	42.0	35.8	22.3	**45.7**	**36.7**	17.6	S 46
GPT-5.6 Terra	65.6	41.7	36.3	22.1	**43.9**	**36.8**	19.3	S 44
Gemini 2.5 Flash	63.7	41.0	38.7	20.2	**41.3**	**40.9**	17.9	S 41
GPT-5.6 Luna	62.0	48.3	29.3	22.4	**54.5**	25.2	20.3	S 54
GPT-4.1	53.1	57.9	25.3	16.8	**64.5**	23.4	12.2	S 64



(a) Acc_6 against the rate of answering S. The fit excludes Gemini 3.8 Flash.

(b) The determination plane. Nine systems cluster at high FD and low FU.

Figure 4: FD is the rate of committing to S or C where the premises settle neither, and FU the rate of answering U where they settle one. A system that endorses readily buys accuracy on the supported cells and pays for it under FD. A system that hedges readily makes the opposite trade.

Table 11: Design cells. The label is determined by construction from the premise condition and conclusion polarity. **P=** exists for the 82 anchors that contained removable redundant text.

Premise condition	Perturbation	**C+**	**C–**
P+ Evidence intact	none	SUPPORTED (108)	CONTRADICTED (108)
P– Evidence reversed	counterfactual edit	CONTRADICTED (108)	SUPPORTED (108)
P0 Evidence removed	evidence deletion	UNDETERMINED (108)	UNDETERMINED (108)
P= Evidence preserved	redundancy deletion	SUPPORTED (82)	CONTRADICTED (82)
Total		812 instances	

not constitute a valid reversal merely because the local edit succeeded. Such inconsistencies must be resolved before the candidate can enter the matched core.

This requirement also applies to insufficiency and relation-preserving controls. A deletion cannot establish insufficiency while another evidential route remains, and a rewrite cannot be treated as relation-preserving if it changes a necessary qualification. All decisions concern the complete evaluated assertion under the same annotation policy Π.

F.7 Core selection and family-level evaluation.

A six-cell core is eligible for core evaluation only when all six reference relations hold. Incomplete or invalid cores are not scored as though they were complete. Additional valid removals, rewrites, and other constructions remain family extensions rather than increasing the number of core cells.

The core-version selection rule is fixed before model evaluation and does not depend on model predictions. Selection retains the matched structure: both claim orientations are evaluated against the same selected premise for each row. The selected core, rather than the total number of generated variants, determines an anchor's contribution to six-cell evaluation. This prevents anchors with more augmentation candidates from receiving greater weight.

G Dataset

G.1 Construction and Design

Each instance pairs a natural-language *conclusion* drawn from a scientific paper with the *full text* of the works it cites as evidence, and asks the model to decide whether the cited material SUPPORTS the conclusion, CONTRADICTS it, or leaves it UNDETERMINED. Instances are organised into *anchors*: a single claim–evidence item from which every design cell is generated by controlled perturbation. Because all cells of an anchor share a provenance, comparisons across cells are within-item and hold topic, writing style and document length approximately fixed.

We cross four premise conditions with two conclusion polarities, giving eight cells (Table 11). The premise conditions are: **P+**, the cited papers unmodified; **P–**, a minimal counterfactual edit that reverses the load-bearing evidence; **P0**, deletion of the load-bearing evidence, leaving the claim underdetermined; and **P=**, a meaning-preserving control in which a comparable span of *redundant* text is deleted while the evidence is left verbatim. The conclusion is either the original claim (**C+**) or its negation (**C–**). The gold label is fixed by construction from the cell, so no annotation step is required to assign it.

Table 12: Corpus scale and input length. Token counts use the `cl100k_base` tokenizer.

Instances (premise–conclusion pairs)	812
Anchors (claim–evidence items)	108
Source papers	27
Distinct cited works	121
Distinct premise document texts	465
Premise documents per instance (mean / max)	1.3 / 4
Premise tokens, median	16,078
Premise tokens, mean	24,640
Premise tokens, 95th pct. / max	75,880 / 120,097
Instances >16K / >32K / >64K tokens	51.2% / 24.6% / 8.4%
Total premise tokens	20.0M
Conclusion tokens, mean / max	38.7 / 137
Evidence spans (`ref_statements`)	15,252
Spans with exactly resolving offsets	100.0%

Table 13: Perturbation magnitude, measured per anchor against the unperturbed **P+** premise. The **P0** / **P=** comparison is over the 82 anchors carrying both conditions.

Condition	Perturbation	Median change	Label
P–	words rewritten	34 words	flips
C–	words rewritten	1 word	flips
P0	premise text deleted	7.0% of characters	→ UNDETERMINED
P=	premise text deleted	7.1% of characters	unchanged

G.2 Scale and Length

The corpus comprises 812 instances built from 108 anchors, spanning 27 source papers and 121 distinct cited works (465 distinct premise document texts once perturbed variants are counted). Premises are entire papers rather than retrieved snippets, which places the benchmark squarely in the long-context regime: the median instance carries 16.1K tokens of premise and the largest carries 120K, with 51.2% of instances exceeding 16K tokens and 24.6% exceeding 32K (Table 12). Conclusions, by contrast, are single sentences of 39 tokens on average, so essentially all of the input length is evidence the model must actually traverse.

G.3 Perturbations are Minimal and Length-Matched

Two properties of the perturbations guard against shortcut solutions (Table 13).

First, the label-flipping edits are *small*. Reversing the evidence (**P–**) changes a median of 34 words of the premise, and negating the conclusion (**C–**) changes a median of a single word. A model cannot locate the decisive evidence by looking for the part of the document that is unusual, distinctive, or stylistically out of place; it has to find the span that bears on the claim.

Second, and more importantly, the UNDETERMINED class is *length matched*. On the 82 anchors carrying both conditions, **P0** deletes a median of 7.0% of the premise and **P=** deletes a median of 7.1% – statistically indistinguishable – yet **P0** is labelled UNDETERMINED while **P=** retains the label of the unperturbed item. The two conditions differ only in *which* text was removed. A model that infers "something was deleted, therefore the evidence is missing" will succeed on **P0** and fail on **P=** at equal rates, so abstention cannot be produced by a length or perturbation heuristic. We verify that **P=** is genuinely meaning-preserving: all 1,870 annotated evidence spans in **P=** appear verbatim and at resolving offsets in the shortened document, whereas in **P–** every anchor's spans differ from **P+**, tracking the counterfactual edit.

G.4 Evidence Annotations

Instances in the evidence-bearing conditions carry span-level annotations (`ref_statements`) marking the passages that bear on the conclusion: 15,252 spans in total, a median of 21 per annotated instance, with a median quote length of 35 words. Every span resolves exactly at its recorded character offsets in the corresponding premise document, and offsets are re-aligned after perturbation, so the annotations remain valid within each cell. Spans are tagged by role (`premise`, 9,384; `assertion`, 5,084).

H Prompt Ablation

Every result in the paper uses one system prompt. Since the benchmark asks whether a model reasons from the documents or from what it already knows, the instruction could in principle be doing that work for it. We therefore re-ran four systems on the full core under two further prompts, one stricter and one looser than the default, and compare all three on the rows every arm judged.

Prompts. **Default** is the instruction used throughout.

```
You are judging how a CONCLUSION follows, given premise documents
Answer with exactly one of:
  Supported    - the premises back the conclusion.  They need
not word every detail the same way:  if they establish what the
conclusion asserts, it is supported.
  Contradicted - the premises say otherwise:  they state the
opposite, or give a different value for something the conclusion
states.
  Uncertain    - the premises neither establish the conclusion nor
contradict it.  Use this when they simply do not settle it, not when
you merely feel unsure.
Respond with the answer and nothing else.
```

Strict is the default with one line appended.

```
Use the premise documents STRICTLY. Judge only from what they say:
do not use world knowledge, prior familiarity with the work, or what
you believe to be true outside these documents.
```

Loose replaces the instruction entirely. It gives no definitions and does not name the premises.

```
Here is some information, followed by a conclusion.  Is the
conclusion supported, contradicted, or uncertain?  Answer with one
word:  Supported, Contradicted, or Uncertain.
```

The user message is identical under all three, and the three labels are parsed the same way.

What changes. On Acc_6 the instruction moves Gemini 3.8 Flash and Claude Sonnet 5 by under two points in either direction, within noise (Figure 5). It moves the two GPT-5.6 systems more. Terra gains four points under strict and loses seven under loose, and Luna loses three under loose, all significant. Every one of those changes sits on the P^0 row (Table 14). Told not to use world knowledge, Terra answers U where the evidence is gone on eleven more points of anchors; told nothing, it answers from memory and loses eighteen. The determinate rows are nearly still under every prompt, and the counterfactual row in particular moves by under three points for every system but Terra under loose.

The instruction is therefore a dial on hedging and not on evidence use. It changes how readily a system says U, and it changes that most for the systems with the strongest priors, but it does not change whether a system finds and accepts an edit that contradicts what it remembers. The default prompt sits between the two, and for Terra its P^0 accuracy is closer to loose than to strict, so the headline numbers do not overstate the benchmark's difficulty on that row.

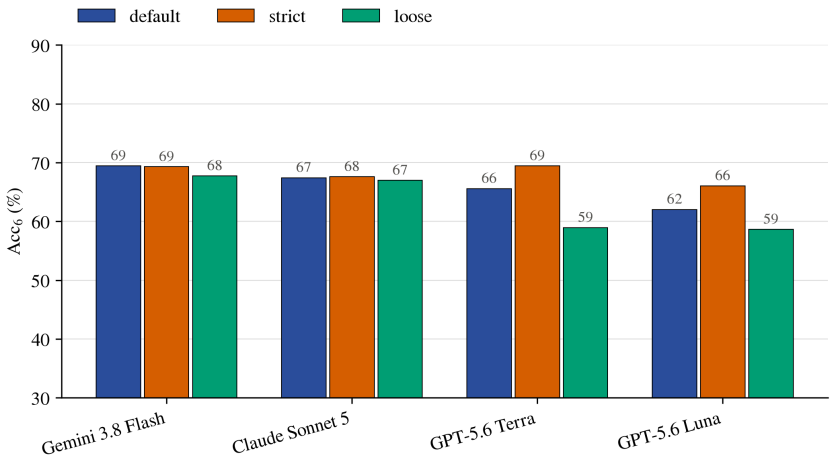


Figure 5: Acc_6 under each prompt.

Table 14: Accuracy under each prompt on the six-cell core and by premise row. Parentheses give the change from default, bold where an anchor-level bootstrap interval excludes zero.

Model	Prompt	Acc_6	P^+ row	P^- row	P^0 row
Gemini 3.8 Flash	default	69.4	69.9	59.3	79.2
	strict	69.3 (−0.2)	65.7 (−**4.2**)	57.9 (−1.4)	84.3 (+**5.1**)
	loose	67.7 (−1.7)	66.7 (−**3.2**)	57.4 (−1.9)	79.2 (+0.0)
Claude Sonnet 5	default	67.4	83.8	68.7	49.8
	strict	67.6 (+0.2)	83.8 (+0.0)	68.2 (−0.5)	50.7 (+0.9)
	loose	67.0 (−0.5)	85.2 (+1.4)	71.5 (+2.8)	44.2 (−**5.6**)
GPT-5.6 Terra	default	65.6	81.9	68.5	46.3
	strict	69.4 (+**3.9**)	82.9 (+0.9)	68.5 (+0.0)	56.9 (+**10.6**)
	loose	59.0 (−**6.6**)	85.2 (+3.2)	63.0 (−**5.6**)	28.7 (−**17.6**)
GPT-5.6 Luna	default	62.0	81.0	61.1	44.0
	strict	66.0 (+**4.0**)	79.6 (−1.4)	61.1 (+0.0)	57.4 (+**13.4**)
	loose	58.6 (−**3.4**)	78.7 (−2.3)	63.4 (+2.3)	33.8 (−**10.2**)

I Additional Analysis of Prior–Evidence Interactions

I.1 Alignment magnitude and statistical uncertainty.

The alignment coefficient is positive for every model, with $0.53 \leq \hat{\beta}_1 \leq 2.54$, corresponding to a 1.7–12.6-fold increase in the odds of correctness under alignment. Because the two-group logistic model is saturated, its fitted probabilities equal the observed aligned and conflicting accuracies; their difference is precisely the alignment gap reported in Table 15. The coefficient and probability-scale gap therefore express the same empirical contrast rather than independent findings. Nine models show an alignment effect under both the two-sided Wald test with anchor-cluster-robust standard errors and paired anchor-level resampling. Astra is the exception: its Wald test gives $p = 0.050$, while its probability-scale interval, $[-0.01, 0.13]$, includes zero.

I.2 Error direction under conflict.

The alignment gap measures how frequently a model is correct, whereas ErrDir_θ characterizes what it does when incorrect. For nine models, conflict errors more often reproduce the measured prior than return undetermined. The two Gemini models illustrate the distinction: Gemini 2.5 Flash follows its prior on 74% of conflict errors, whereas Gemini 3.8 Flash does so on 41%, abstaining more frequently. The latter's interval, $[0.26, 0.55]$, includes the direction-uniform reference 1/2; Fable's lower endpoint of 0.52 lies only narrowly above it. Thus, failure under conflicting evidence need not take the same form across models: some retain the prior-consistent answer, while others withhold judgment despite sufficient evidence.

I.3 Prior-directed commitment under insufficiency.

All ten models favor their measured prior when they give determinate answers to unresolved instances. Importantly, $\mathrm{DirComp}_\theta$ conditions on commitment: it measures the direction of unwarranted answers, not how frequently a model gives them. Fable's 97% and Astra's 93% therefore do not imply that either model answers determinately on nearly all insufficient-evidence instances. Rather, they show that even models with comparatively small alignment gaps produce strongly prior-consistent answers when they fail to withhold judgment. Performance against opposing evidence and appropriate withholding under missing evidence therefore capture distinct aspects of evidence dependence.

I.4 Eligibility and interpretation.

Stable evidence-free priors are available for 65–96% of the 216 claims, with the two Gemini models having the lowest coverage. Prior-conditioned comparisons therefore involve partially different claim sets and complement, rather than replace, the common-set RQ1 evaluation. The conditional error analyses also have smaller denominators: ErrDir_θ uses 31–127 conflict errors per model, and $\mathrm{DirComp}_\theta$ uses 34–142 determinate commitments on unresolved instances. These differences should be considered when comparing estimates and uncertainty across models. Throughout, prior-consistent errors describe an association with independently measured evidence-free judgments; they do not uniquely identify the internal process producing those errors.

I.5 Additional Analysis of Evidence Access

Logistic fit. Because anchors reuse surface forms, we test whether within-anchor dependence drives the alignment effect by regressing correctness on alignment over prior-eligible determinate instances:

$$\operatorname{logit} \Pr(\widehat{y} = y) = \beta_0 + \beta_1 \alpha, \tag{10}$$

where β_1 is the inferential counterpart of Gap_θ and β_0 gives the log-odds of correctness under conflict. Because the two-group model is saturated, each model is summarized by

these two coefficients; $\hat{\beta}_1$ is positive for all ten models and clears the two-sided Wald test (anchor-cluster-robust CR1 errors) for nine, matching the resampling intervals of Table 16.

Table 15: Per-model fit of equation 10: intercept $\hat{\beta}_0$, alignment coefficient $\hat{\beta}_1$, and the p-value of its two-sided Wald test (anchor-cluster-robust CR1 errors). The fit is saturated, so $\sigma(\hat{\beta}_0 + \hat{\beta}_1)$ and $\sigma(\hat{\beta}_0)$ equal the observed aligned and conflicting accuracies whose anchor-balanced difference is Gap_θ (Table 4).

Model	n	$\hat{\beta}_0$	$\hat{\beta}_1$	p
GPT-4.1	416	−0.45	2.54	<.001
GPT-5.6 Luna	402	0.27	1.73	<.001
GPT-5.6 terra	408	0.58	1.30	<.001
Claude Sonnet 5	363	0.78	1.06	<.001
Gemini 3.8 Flash	282	0.39	0.68	.001
Deepseek Flash	300	1.05	1.23	<.001
Claude Opus 5	386	1.37	1.10	<.001
Gemini 2.5 Flash	282	0.79	0.79	.009
Claude Fable 5.1	366	1.59	0.83	.018
GPT-6 Astra	364	1.51	0.53	.050

Intervals for the prior-conditioned measures. Table 16 reports the paired anchor-level resampling intervals ($B = 2000$) for the three prior-conditioned measures whose point estimates appear in Table 4, together with their reference values and per-measure denominators.

Table 16: Per-model prior-conditioned measures of equation 4: point estimates with paired anchor-level resampling intervals ($B = 2000$) in parentheses, estimated with anchor-balanced weighting. Models are ordered by Gap_θ; the reference values are 0 for Gap_θ (no alignment effect) and 1/2 for ErrDir_θ and DirComp_θ (no directional tendency). Denominators differ by model: ErrDir_θ is taken over 31–127 conflict errors and DirComp_θ over 34–142 committed responses to undetermined rows.

Model	Gap_θ	ErrDir_θ	DirComp_θ
GPT-4.1	0.51 (0.42–0.58)	0.84 (0.78–0.91)	0.84 (0.78–0.90)
GPT-5.6 Luna	0.30 (0.22–0.39)	0.70 (0.58–0.80)	0.79 (0.70–0.86)
GPT-5.6 Terra	0.23 (0.15–0.31)	0.70 (0.59–0.80)	0.85 (0.77–0.92)
Claude Sonnet 5	0.18 (0.10–0.26)	0.84 (0.74–0.93)	0.80 (0.71–0.88)
Gemini 3.8 Flash	0.17 (0.07–0.26)	0.41 (0.26–0.55)	0.91 (0.80–0.98)
Deepseek Flash	0.16 (0.07–0.26)	0.73 (0.58–0.85)	0.78 (0.68–0.89)
Claude Opus 5	0.14 (0.06–0.21)	0.87 (0.76–0.97)	0.94 (0.88–0.98)
Gemini 2.5 Flash	0.14 (0.04–0.24)	0.74 (0.58–0.88)	0.71 (0.61–0.81)
Claude Fable 5.1	0.09 (0.02–0.16)	0.72 (0.52–0.89)	0.97 (0.93–1.00)
GPT-6 Astra	0.06 (−0.01–0.13)	0.70 (0.56–0.85)	0.93 (0.88–0.97)

Intervals for the stratified gaps. Table 17 reports the paired anchor-level resampling intervals ($B = 2000$) for the belief-sign-stratified gaps and their memory and disposition components whose point estimates appear in Table 3, recomputed per model and belief-sign group with the same anchor resamples so that the components are paired contrasts of the group gaps.

Table 17: Point estimates of Table 3 with paired anchor-level resampling intervals ($B = 2000$) in parentheses.

Model	believes-true	believes-false	memory	disposition
GPT-4.1	0.53 (0.43–0.62)	0.30 (0.09–0.50)	0.41 (0.31–0.52)	0.11 (−0.02–0.24)
GPT-5.6 Luna	0.41 (0.30–0.53)	0.14 (0.02–0.26)	0.28 (0.19–0.36)	0.14 (0.06–0.22)
GPT-5.6 Terra	0.27 (0.16–0.38)	0.15 (0.05–0.25)	0.21 (0.14–0.28)	0.06 (−0.02–0.14)
Claude Sonnet 5	0.23 (0.13–0.35)	0.12 (0.00–0.23)	0.18 (0.10–0.25)	0.06 (−0.02–0.14)
Gemini 3.8 Flash	−0.14 (−0.28–0.00)	0.39 (0.27–0.50)	0.12 (0.04–0.21)	−0.26 (−0.36–−0.16)
Deepseek Flash	0.13 (0.01–0.25)	0.19 (0.08–0.31)	0.16 (0.08–0.25)	−0.03 (−0.11–0.05)
Claude Opus 5	0.15 (0.05–0.25)	0.09 (0.01–0.18)	0.12 (0.05–0.18)	0.03 (−0.04–0.09)
Gemini 2.5 Flash	0.13 (0.01–0.25)	0.13 (−0.05–0.30)	0.13 (0.02–0.23)	0.00 (−0.10–0.11)
Claude Fable 5.1	0.14 (0.05–0.24)	0.03 (−0.04–0.10)	0.09 (0.02–0.16)	0.06 (0.01–0.10)
GPT-6 Astra	0.11 (−0.01–0.22)	0.01 (−0.06–0.09)	0.06 (0.00–0.12)	0.05 (−0.03–0.13)

Scope of the conclusions. Counterfactual premise status is not equivalent to conflict with a model's independently measured prior. Accordingly, the P^+/P^- comparisons do not replace the prior-conditioned analysis in RQ2. Highlighting changes evidence salience, while oracle extraction also changes the surrounding context; neither intervention uniquely separates localization from interpretation or prior-consistent responding. The supported conclusion is that making evidence explicit can improve long-context judgments, sometimes without shortening the documents, but does not consistently eliminate errors.