A COMMON SCALE FOR CONTEXTUAL OVERRIDE: WHAT SURVIVES A CHANGE OF WORDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval systems are vulnerable to attacks embedded within the documents they retrieve. Presenting a claim as fiction, attributing it to a specific source, or repeating it across multiple documents are all known exploit vectors. However, since each finding originates from a different experimental setup, these attacks have never been systematically ranked against one another. Furthermore, they have never been measured against the baseline control required to make such a ranking interpretable: the same claim, embedded in the same document, but explicitly denied. In this work, we aim to construct this unified scale. A controlled document generator maintains the fact, counterfactual object, prompt question, sentence count, and mention count as strict invariants, changing only the framing of a single sentence. The resulting behavior is then scored as a signed margin against each model's measured parametric prior. Because each rung of this scale uses a distinct sentence, a surface-level ranking of framings risks reducing the analysis to a simple ranking of arbitrary wordings. To mitigate this, we rebuild both evaluation ladders using an independent set of claim templates and re-assess all 11 models. Accumulated corroboration emerges as the most effective attack we tested. All models are overridden on a majority of facts by at least one accumulation condition (achieving an 89% override rate across 185 facts at eight sources, compared to just 26% when the claim is denied). Instruction tuning halves the efficacy of single-mention attacks while simultaneously sharpening the impact of accumulation. This asymmetry is critical for real-world deployment, where a retrieval ranker dictates the exact number of corroborating passages presented to the model. When matched on their baseline parametric priors, models differ primarily in susceptibility (the magnitude by which a single mention shifts their confidence), not in memorization or polarity sensitivity. Ultimately, while the hierarchical ordering between framing tiers survives a change of wording, the assumption that denial constitutes the weakest possible framing does not.

1 INTRODUCTION

When a retrieved document contradicts what a language model learned during pretraining, the model must arbitrate. Determining which evidence prevails is a highly practical question, since a retrieval system does not filter what it surfaces: a query regarding a false claim may simultaneously return fact-checks debunking it, rumors repeating it, fiction entertaining it, and passages explicitly corroborating it. The prevailing view, which holds that parametric knowledge and contextual evidence each carry an intrinsic strength and the stronger wins, predicts that a document *denying* a claim should be safe; however, this assumption fails in practice. When prompted with the text "It is false that the capital of France is Sucre", OLMo-2-7B erroneously outputs Sucre on 31% of the facts it otherwise knows, and Qwen2.5-7B commits the same error on 84%. Recent works have examined these framing effects in isolation: presenting claims as fiction (Basmova et al., 2024), attributing them to specific sources (Schuster et al., 2026), repeating them (Jin et al., 2024a), and training on documents that flag a claim as *false* (Mayne et al., 2026) have all proven effective. Because each finding originates from a distinct methodological construction (such as a forced choice between two contexts, fine-tuning on synthetic corpora, or entity substitution), the effect magnitudes remain fundamentally incomparable. Consequently, these attacks have never been systematically ranked, leaving it unresolved whether a fact-check or a corroborated passage constitutes the more concerning artifact.

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.

Establishing a rigorous ranking requires two critical elements absent from existing literature. The first is a *common scale*: a unified document generator that maintains the fact, the counterfactual object, the prompt question, the sentence count, and the mention count as invariants, modifying only the framing of a single sentence to be scored against each model's measured parametric prior. The second is an *anchor* to make this scale interpretable: the same claim, embedded within the same document, but explicitly denied. Explicit denial represents the critical control condition that the field has thus far omitted. Its inclusion transforms the binary question of "does manipulation X work?" into a quantifiable measurement of "how far does X move a model relative to denying the claim?"

This unified scale has an inherent limitation that forms the central methodological focus of this paper. Since each rung utilizes a distinct sentence, any resulting hierarchy of framings is confounded with the specific wording used. This confound is directly measurable within our experimental design: evaluating the same logical condition using a different sentence pool shifts the margin by up to 2.07 log-odds and the override rate by 38 percentage points, whereas two independent evaluations of the *identical* sentence agree to within 0.14. To isolate the conceptual framings from arbitrary phrasing effects, we rebuild both ladders utilizing an independently authored set of claim templates and subsequently re-evaluate all 11 models. This rigorous control, alongside the robust findings that survive it, constitutes the main contribution of this work. Our evaluation reveals a clear hierarchical ordering between broad framing categories, while the ordering within categories is substantially less consistent. Accumulated corroboration emerges as the most effective attack we tested: presenting eight distinct sources overrides 89% of facts, and multi-source confirmation overrides 86%, compared to just 26% for the same claim when explicitly denied. Moreover, both a fictional framing and an unverified rumor outperform a direct denial, successfully replicating the findings of (Basmova et al., 2024). However, denial is *not* reliably the weakest framing: it ranks lowest on only 4 and 5 of the 11 models across the two respective template sets, and its relative ordering against a debunking framing inverts depending on the specific statistical metric applied. We report this methodological instability alongside our primary results. The main contributions of this work are as follows:

- **A Scale Whose Resolution Is Empirically Established** (§3, §4.2). We construct a length- and mention-matched dose ladder spanning 185 facts and 11 models, fully reconstructing it using two independently authored claim-template sets. The specific template choice shifts the entire ladder (ranging from -1.28 to $+0.25$ log-odds) but alters its internal spacing (resulting in a median cross-condition standard deviation of 0.27). This confirms that while the structural ordering survives a change in wording, the absolute magnitudes do not. We identify which comparisons this instrument can and cannot reliably support, while reporting an instance in which one of our headline orderings fails this robustness test.
- **Empirical Findings from the Validated Scale** (§4.1, §4.3). Accumulated corroboration defeats every model we tested. Instruction tuning (the intervention available to a practitioner) halves the efficacy of every single-mention attack across the five base-instruct pairs where the base model is vulnerable; however, it shifts the impact of accumulation in the *wrong* direction in 10 out of the 12 pair-by-axis comparisons. Matched on their baseline parametric prior, models are differentiated by their susceptibility, not their memorization or polarity sensitivity: the models that resist manipulation register a direct denial worse.
- **Four Mitigations That Fail** (§4.4). Plausibility, corpus exposure, source filtering, and prompt structure each appear to offer protection; none does, and prompt structure is actively harmful for one model.

2 Related Work

The Closest Work. Basmova et al. (2024) evaluate language models across varying knowledge conditions and linguistic environments, finding that hypothetical contexts are particularly vulnerable to knowledge conflict and that non-affirmative errors typically treat entailment-canceled statements as affirmative. While our fiction-versus-denial result aligns with their findings, our evaluation metric differs. They score abstention accuracy on questions rendered unanswerable, a measure that inherently conflates counterfactual adoption with reliance on parametric memory. By contrast, we compute a signed margin between two candidates against the model's measured prior. This isolates these behaviors and enables linguistic framings, source attribution, and evidence accumulation to be ranked on a single continuous axis (§3.1).

What We Build On. Mayne et al. (2026) demonstrate that fine-tuning models on documents that refute a claim causes them to assert it. Their in-context control corroborates §4.1, but we newly resolve the internal ordering of this regime, which they report only as a single aggregate number. In forced-choice contextual evaluations, Schuster et al. (2026) establish a credibility hierarchy and its reversal under repetition, while Jin et al. (2024a) demonstrate the majority-rule behavior that anchors the top of our scale. Other studies of knowledge conflict rely exclusively on assertive evidence (Longpre et al., 2021; Xie et al., 2024; Xu et al., 2024); our polarity control indicates this design choice obscures the underlying override mechanism. Finally, we confirm established findings that conflict resolution occurs in late network layers (Jin et al., 2024b; Ortu et al., 2024), while contributing an additional negative result (Appendix E).

The Gap. Prior research relates conflict override to prior strength using base-model log-probability (Cheng et al., 2026) or entity popularity (Mallen et al., 2023) as proxies. Since these studies measure prior strength and override on the same population, they cannot separate a model's capacity to resist evidence from its initial distance to the decision boundary; §3.3 formally supplies this separation. On the effect of instruction tuning, previous works report conflicting directional outcomes (Goyal et al., 2025; Han et al., 2025; Kim et al., 2026). §4.2 measures this regime difference without claiming priority. However, to the best of our knowledge, the distinct *asymmetry* between single contextual mentions and accumulated corroboration remains unexplored.

3 The Instrument

We propose a unified measurement framework: an experimental construction that renders four families of contextual attacks as directly comparable quantities, accompanied by controls that establish precisely what these metrics can and cannot distinguish.

Notation. A fact is defined as a triple (s, r, o) (subject, relation, object) associated with a prompt question $q_r(s)$ whose true answer is o; each fact is assigned a counterfactual object o' of the identical entity type. A document $d(f, \ell, \rho)$ is constructed from a given fact f, a ladder level ℓ governing the framing, and a seed ρ determining the filler text. T_a and T_b denote the two independently authored claim-template sets (§3.4). Finally, $M_{AB}(\tau)$ defines the subset of facts for which models A and B exhibit baseline parametric priors within τ log-odds of one another (§3.3).

3.1 What we measure, and why a signed margin

At answer position, we test the log-probability of the first token for o and o' and report the difference:

$$\Delta(f, \ell); =; \mathbb{E}_\rho\big[\log p(o' \mid d(f, \ell, \rho)); -; \log p(o \mid d(f, \ell, \rho))\big], \tag{1}$$

Hence, a positive margin ($\Delta > 0$) indicates the model outputs the falsehood. Three properties of Eq. 1 are critical. First, it is *continuous*: unlike binary flip rates that quickly saturate on our ladders, the margin captures shifts even after override rates plateau, exposing the instruction-tuning effect (§4.3). Second, as a *difference between two candidates*, it cancels position-specific scaling and separates two failure modes conflated by single-answer metrics: actively asserting a falsehood versus declining to answer. Standard accuracy penalizes both identically, a flaw noted by Basmova et al. (2024) that makes conflict-induced abstention appear as false robustness. Isolating these mechanistically distinct behaviors allows us to rank framing, attribution, and accumulation along a single unified axis. Finally, the metric is *signed at the decision boundary of zero*, yielding both a continuous effect size and a discrete override rate. Because these can diverge near the boundary (§4.2), we report both and specify which metric supports each structural claim.

We score the first token instead of the full candidate string for structural reasons. The model commits to an answer at the token immediately following `Answer:`. Since both candidates are evaluated at this exact position under a shared context, the metric compares two continuations of a single prefix, allowing position-specific scale factors to cancel out. Full-string scoring provides neither guarantee; candidate strings span varying numbers of tokens, forcing length normalization to substitute for a direct comparison. This approach also fails empirically: on two of the evaluated models, length-normalized teacher forcing drives closed-book accuracy to chance and compresses the dynamic range below 2 log-odds (Appendix B). As a result, candidate pairs must differ at the evaluated position. Any facts whose true and counterfactual objects share a first token under a specific model's tokenizer are excluded before analysis, and this exclusion count is recorded per run.

3.2 DOCUMENT CONSTRUCTION AND ITS INVARIANTS

Every document contains exactly $D{=}8$ sentences, one of which carries the target claim. Along the framing axes, the counterfactual is mentioned exactly once. Both structural invariants are enforced in code and validated by regression tests designed to fail on earlier builds.

Sentence count: holding D fixed requires adjusting the amount of filler text based on the inserted framing's actual sentence count, rather than applying a constant offset. For instance, a framing that spans two sentences must displace exactly two filler sentences. An earlier pipeline iteration applied a fixed padding of $D-1$ regardless of the framing length, which inadvertently produced documents one sentence longer under three of the six evaluated framings. **Mention count:** on the polarity and credibility axes, the counterfactual appears exactly once at every level, ensuring that these ladders vary framing alone. The quantity axis is the exception by design: its dose is defined by the mention count, and its successive levels replace filler text with additional paraphrases of the same claim.

Word count cannot be equalized, as a refutation is longer than the assertion it refutes. We therefore report word counts instead of claiming they are matched (Appendix A). This residual variation in length actually works against our conclusion, because allocating more words to deny a claim provides the model with a stronger signal that a denial is occurring. **Counterfactuals** are assigned via derangement within each relation, ensuring that every incorrect answer remains type-appropriate and no fact retains its true object. For example, *Sucre* is the actual capital of another country in the inventory, not a fabricated string. Finally, filler text is drawn from a pool of 25 register-matched sentences and shuffled, allowing the position of the target claim to vary across documents.

The experimental design comprises five distinct axes, each isolating a single property: **polarity** (the framing of the single mention: asserted, negated, refuted, debunked, or fictional, plus a true-answer baseline); **credibility** (attribution ranging from an unverified online rumor to multi-source official confirmation, with the assertion count held at one); **quantity** (0 to 8 corroborating sources); **plausibility** (arbitrary versus near-miss counterfactuals); and **position**. Each axis functions as a dose ladder in which level 0 is always the identical filler-only document.

3.3 TWO CONTROLS THAT MAKE THE COMPARISON POSSIBLE

The Polarity Anchor. A ladder of manipulations requires a baseline to be interpretable. Our baseline is the identical claim presented within the same document, but explicitly denied. The graded account of knowledge conflict predicts this condition should be safe: if a model weighs evidence, a denial should not push it toward the denied claim. However, this is precisely the control that prior work has omitted in context. Anchoring the scale to this baseline shifts the experimental question from simply asking "does manipulation X work?" to measuring "how far does X move a model relative to denying the claim outright?". This comparative metric constitutes the exact question a practitioner deciding what to retrieve must answer.

The Matched-Prior Subset. Override rates are not directly comparable across models that begin at different distances from the decision boundary: our no-evidence margins range from -7.75 to -13.73. Recent work relates prior strength to override observationally (Cheng et al., 2026; Mallen et al., 2023), measuring both on the same population. Hence, "resist evidence" and "start further away" are never separated. We isolate these factors by restricting our evaluation to:

$$M_{AB}(\tau); =; , f : |\Delta_A(f, 0) - \Delta_B(f, 0)| < \tau,, \tag{2}$$

the subset of facts on which two models begin within τ log-odds of each other, and we compare the *paired* per-fact swing within this region. We set $\tau = 1.0$ and report a sweep over τ in §4.3, since an effect that grew as the matching window narrowed would indicate selection on noise. Because $M_{AB}(\tau)$ identifies the precise region where a strong-prior and a weak-prior model coincide rather than a random sample of facts, the resulting comparison remains strictly conditional on it.

3.4 SEPARATING THE ORDERING FROM THE WORDING

Since every ladder rung uses a distinct sentence, rung ordering is confounded with sentence wording, *i.e.*, the instrument's central weakness, which we explicitly measure. Although polarity ASSERT and quantity one-source represent the identical logical condition (a single unattributed assertion), they draw from distinct sentence pools and differ by up to 2.07 log-odds and 38 override points.

Conversely, credibility PLAIN uses the *same* sentence as polarity ASSERT in a separate run and agrees within 0.14 log-odds, confirming the gap stems from carrier wording, not run-to-run noise. To isolate this effect, we author an independent template set, T_b, re-evaluating all 11 models (separated by a strict aggregation key). Precise agreement at level 0 (which contains no claim sentence, differing by only 0.009 log-odds) serves as a manipulation check. We decompose this cross-template comparison into a *level shift* (overall ladder weakening) and a *spacing change* (shifts in between-rung gaps). A large level shift alongside a small spacing change indicates that wording dictates absolute attack strength while preserving relative ordering. §4.2 reports both metrics, using the spacing change as a resolution bound to flag any claims falling below this threshold.

3.5 Does the instrument measure what we think?

Tokenizer Failure. Answer-position scoring can produce plausible but meaningless values under certain tokenizers, requiring all runs to be screened before analysis. Specifically, a document asserting the *true* answer must not leave the model preferring the falsehood, and the agreement-to-assertion span must exceed 4 log-odds. All 11 evaluated runs successfully pass this gate (Appendix B).

Forced Choice. Eq. 1 evaluates which of two candidate answers a model prefers. Consequently, it cannot observe a third potential behavior: declining to answer, a response Basmova et al. (2024) found to be common under knowledge conflict. To investigate this, we generate and classify continuations for 60 facts per model across 7 conditions using both greedy and sampled decoding. Abstention proved rare (0.9-2.5%). The sign of the margin predicts the emitted answer on 98.3% of items at the median, and the rate of emitted counterfactuals tracks the computed override rate across all 77 model-by-condition cells ($r = 0.989$). We thus write that a model gives the false answer.

Prompting Regime and Sensitivity. All preceding evaluations rely on raw completion. §4.2 details how these outcomes shift under a chat template, establishing a critical scope condition for every quantitative result reported here. Additionally, a positive control utilizing 28 transformers trained from scratch demonstrates that graded resistance is recoverable where it exists (Appendix F). Because this control operates through the tier-1 pipeline rather than the answer-position scorer, it validates our broader analytical machinery, not the scoring mechanism itself.

3.6 Facts and inference

We use 185 facts across four relations (capitals, currencies, authorship, element symbols), selected for temporal stability, unique answers, and broad corpus frequency distributions. This deliberately small inventory permits exhaustive testing (690,000 scored documents) and paired per-fact comparisons. The binding constraint is the number of *relations*, not facts: with only four relations, relation type is confounded with corpus frequency (§4.4), making a fifth relation far more informative than a thousand additional capitals. All axes use this full inventory except plausibility, which requires type-matched near misses (8187 tokenizer-dependent facts). We report point estimates with 95% bootstrap intervals (5000 fact resamples). Because marginal intervals can falsely overlap even when effects are reliable, within-model comparisons directly bootstrap the *paired* per-fact differences.

4 Does the Ordering Survive Its Own Wording?

We report the empirical measurements provided by the scale (§4.1), test whether it orders the framings, not our specific sentences (§4.2), and apply the findings that survive this control (§4.3, §4.4).

4.1 The scale: one mention, seven framings

Closed-book accuracy (0.978 to 1.000) and no-evidence margins (-7.75 to -13.73) confirm these models know the evaluated facts and decisively prefer the true answer. Retaining the few facts that 5 models miss closed-book represents a conservative choice, inflating override rates by at most 0.8 percentage points. Changing a single sentence within an eight-sentence document leads to the behavioral spread in Fig. 1. A plain assertion is the strongest single-mention framing, overriding 70% of facts on average (8.1%-84.3% across models). Explicit denial is substantially weaker, averaging a 26% override rate and remaining below 50% on 10 of the 11 evaluated models. For instance,

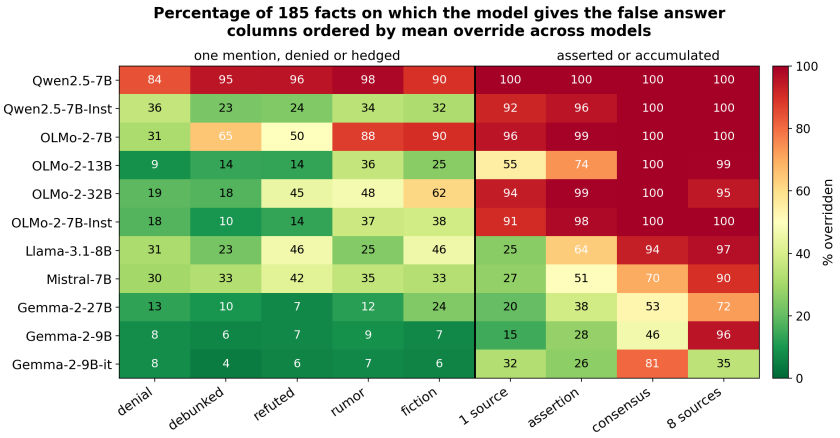

Figure 1: Percentage of 185 facts on which each model gives the false answer, detailed per condition. Columns are ordered by the mean override rate. Conditions to the left of the vertical rule mention the counterfactual exactly once. The overarching block structure (a single denied or framed mention, a plain assertion, and accumulated sources) represents our primary claim. The specific ordering within the single-mention block, however, is finer than the instrument can reliably resolve (§4.2).

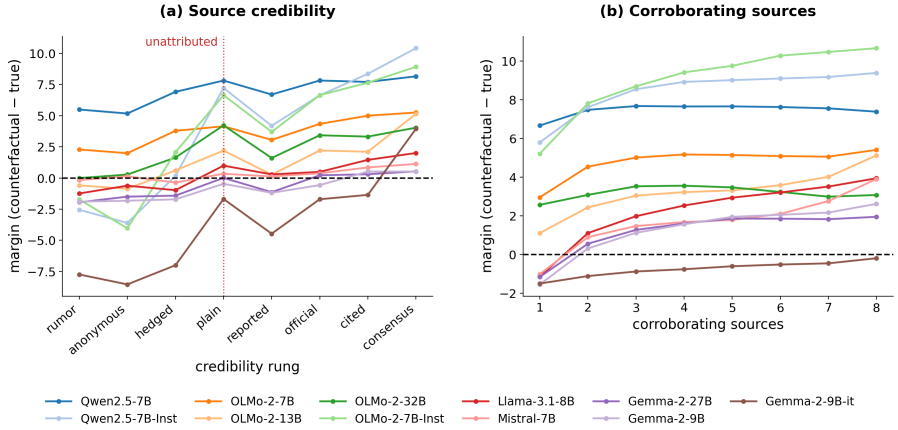

Figure 2: The evidence ladders evaluated across all 11 models. (a) Source credibility, ranging from an unverified rumor to multi-source official confirmation; the dotted line marks the *plain* framing, an unattributed assertion that represents the ladder's only inversion. (b) One to eight corroborating sources, with the dashed line indicating the decision boundary.

OLMo-2-7B's margin is restricted to -0.66 under "It is false that …", compared to $+4.20$ for a bare assertion (Appendix J). Contrary to cloze-based negation probing (Kassner & Schütze, 2020; Truong et al., 2023), these models successfully distinguish explicit denials from assertions.

Three single-mention framings surpass explicit denial. A fictional framing ("Imagine a fictional world in which …") results in a higher margin than denial on 10 models, reversing only on Mistral-7B. Source attribution establishes a graded credibility ladder, monotonic ($\rho \geq 0.85$, range $+0.69$-$+0.98$) across 7 models; notably, even its *weakest* rung (an unverified rumor) beats denial on 9 models. Finally, corroboration pushes 10 of 11 models above the decision boundary. Gemma-2-9B-it remains the sole exception, reaching a margin of only -0.19 at eight sources (Fig. 2 and Appendix K). None of these interventions increases credibility: a fictional world is explicitly counterfactual, the weakest attribution is an unverified rumor, and eight paraphrases of one claim yield no independent evidence. One hypothesis is that these framings shift the counterfactual from the sentence's *target* to its *content*, leaving denial as the only framing that actively opposes the claim. However, because we did not manipulate syntactic role independently, we cannot rule out a simpler alternative: that these framings differ in how much of the sentence qualifies rather than asserts.

Table 1: Cross-model means computed across all 11 models for the two independently authored claim-template sets. Although template set T_b weakens the effect of every condition, it preserves the structural ordering. Notably, the denial and debunking framings exchange rank depending on whether they are tested by override rate or margin, a reversal that occurs across both template sets.

	override rate (%)		margin	
Condition	set T_a	set T_b	set T_a	set T_b
Denial	26.1	22.2	-2.16	-2.88
Debunked	27.3	22.9	-2.35	-2.99
Refuted	31.9	26.1	-1.77	-2.51
Fiction	41.2	35.4	-0.77	-1.42

4.2 THE TEST: REBUILDING THE SCALE FROM INDEPENDENT WORDING

The Template Shifts the Ladder Without Reordering. Template set T_b weakens every condition by a mean of -1.28 to $+0.25$ log-odds per model, yet this shift remains nearly uniform internally. Its median standard deviation is 0.27 log-odds (0.16-0.73) across the five polarity conditions and 0.28 across the eight credibility rungs. The carrier sentence thus dictates the ladder's strength while preserving its relative spacing. Consequently, although rungs separated by almost a third of a log-odd remain comparable, absolute override rates cannot be cited independently of their source templates.

What Replicates. Fiction surpasses denial on 10 models under T_a and 10 under T_b. This directional agreement holds across all 11 models, persisting even when the gap is under one log-odd; furthermore, Mistral-7B's initial reversal reliably reproduces rather than vanishing as noise. Similarly, an unverified rumor beats denial on 9 models under both T_a and T_b, with Llama-3.1-8B as the sole consistent exception. The credibility ladder is monotonic ($\rho \geq 0.85$) on 7 and 8 models, respectively. Its single inversion (the unattributed PLAIN framing) persists on 11 of 11 models in both sets, placing PLAIN strictly outside the credibility scale.

What Does Not. Denial is the weakest single-mention framing on 4 models under T_a and 5 under T_b, yet the *strongest* on 3 and 1. Table 1 illustrates this tier's broader instability: denial and debunking swap ranks depending on whether override rate or margin is evaluated, a reversal consistent across both template sets. Because metric-dependent differences do not constitute valid orderings, we restrict our claims to the robust block-level ordering: a single denied or framed mention, then a plain assertion, and finally accumulated sources.

A Second Perturbation: The Prompting Regime. While preceding results rely on raw completion, most deployments use chat templates. Re-evaluating the three instruction-tuned models using native chat templates and explicit instructions shifts their baseline priors $+6.06$ to $+11.89$ log-odds closer to the decision boundary *before any evidence is introduced*. Consequently, override rates rise. Gemma-2-9B-it (previously the sole resister of eight-source corroboration) jumps from 5.9% to 98.9% under fiction. Paradoxically, its assertion swing actually *falls*: the model is overridden more frequently not because the mention moves it further, but because its starting point shifted. Despite this perturbation, the structural ordering survives: accumulation beats the strongest single mention in 9 of 9 model-by-regime cells, maintaining a $+0.71$-$+0.94$ rank correlation against raw completion. Since absolute magnitudes shift, our reported raw-completion rates serve as lower bounds for chat deployments. Finally, this baseline displacement offers a mechanism for why our instruction-tuning conclusions diverge from (Kim et al., 2026), though strictly as a hypothesis rather than a formal demonstration, since base models lack native chat templates for direct comparison (Appendix I).

4.3 WHAT THE VALIDATED SCALE REVEALS

Assertion, Not Polarity, Separates the Families. Except for Qwen2.5-7B, all evaluated models resist explicit denial; what distinguishes them is their response to plain assertion (Table 2). Polarity handling does not separate the OLMo-2 and Gemma families, as their denial override rates overlap (closest approach -4.3 points), with OLMo-2-13B outperforming both Gemma scales (-2.94 versus -2.58 and -2.05). Scale is equally non-predictive: OLMo-2's rate spans 22.7 points non-monotonically, and Gemma-2-27B is marginally weaker than Gemma-2-9B. Instead, strict sepa-

Table 2: Polarity control across models (first-token scoring, one document build). "Refute" is the plain-negation framing; "Agree Flip" is the share of facts on which a document asserting the *true* answer causes the model to prefer the falsehood. All 11 runs pass the validity gate (Appendix B). Intervals are 95% bootstrap over facts; text differences use full precision and may differ in last digit.

Model	No Ev.	Assert	Refute	% Overridden [95% CI]	Assert Truth	Agree Flip
Qwen2.5-7B	−8.06	+7.87	+3.28	84.3 [78.9, 89.2]	−12.17	0.0
Qwen2.5-7B-Instruct	−11.78	+7.36	−2.71	35.7 [29.2, 42.7]	−15.95	0.0
OLMo-2-7B	−8.91	+4.20	−0.66	31.4 [24.3, 37.8]	−11.15	0.0
OLMo-2-13B	−8.51	+2.28	−2.94	8.6 [4.9, 13.0]	−11.08	0.0
OLMo-2-32B	−8.78	+4.27	−1.22	18.9 [13.5, 24.9]	−11.26	0.0
OLMo-2-7B-Instruct	−13.73	+6.65	−5.79	17.8 [12.4, 23.3]	−15.53	0.0
Gemma-2-9B	−11.05	−0.47	−2.58	8.1 [4.3, 12.4]	−12.59	0.0
Gemma-2-9B-it	−12.49	−1.67	−7.55	8.1 [4.3, 12.4]	−13.42	0.0
Gemma-2-27B	−9.28	+0.04	−2.05	13.0 [8.6, 17.8]	−10.73	0.0
Llama-3.1-8B	−8.84	+0.94	−1.12	31.4 [24.9, 38.4]	−9.77	0.0
Mistral-7B	−7.75	+0.32	−0.40	29.9 [23.4, 36.4]	−9.39	0.0

Table 3: Matched-prior comparison, restricted to facts where the two models' no-evidence margins differ by less than 1.0 log-odds. Reported intervals represent 95% bootstrap confidence over the paired per-fact differences.

Comparison	Matched Facts	Δ swing	Δ override
Gemma-2-9B vs OLMo-2-7B	59	−3.95 [−4.31, −3.56]	−16.95 [−27.12, −6.78]
Gemma-2-27B vs OLMo-2-7B	51	−3.80 [−4.35, −3.23]	−1.96 [−13.73, +9.80]

ration emerges under plain assertion, which drives every OLMo model well above the decision boundary while Gemma models remain at or barely above it.

Resistance Is Not Simply a Stronger Prior. An alternative hypothesis is that resistant models exhibit stronger memorization, starting further from the boundary. While the no-evidence margin correlates with override (r =+0.52) and assertion swing (r =+0.67), Eq. 2 removes this starting-distance confound by enforcing matched priors. Table 3 demonstrates that the gap persists even at equal parametric conviction: a single mention moves both Gemma scales several log-odds less than OLMo-2-7B. Because these matched subsets are small, we swept the threshold τ. From -3.80 at τ=0.25 (n=16) to -3.80 at τ=4.0 (n=133), every interval excludes zero. The magnitude remains flat rather than artificially inflating as the window narrows, ruling out selection noise. Consequently, models fundamentally differ in *susceptibility*—how far a mention moves them—not memorization or polarity sensitivity. Notably, resistant models process explicit denial *worse*. Measured as the share of assertion swing removed by a denial, the Gemma scales achieve only 20% and 22%, compared to OLMo-2-7B's 37% (matched paired differences of -0.14 [-0.18, -0.10] and -0.16 [-0.20, -0.12]). Conditional on these matched regions, Gemma resists overall manipulation more effectively despite registering the denial less.

Instruction Tuning Closes the Single-mention Attacks, Not the Others. As practitioners' primary intervention, establishing exactly what instruction tuning fixes and leaves vulnerable is our most consequential result. Across six base/instruct pairs, tuning draws a boundary precisely along the block structure of Fig. 1 (also see Fig. 3). It mitigates **single-mention** attacks, acting most dramatically on Qwen2.5-7B. As the sole model failing the polarity control outright, its override drops from 84% to 36% under denial, 90% to 32% under fiction, and 98% to 34% under rumor. OLMo-2-7B exhibits a smaller but identical pattern (denial falling 31% to 18%), as do the Llama, Mistral, and Qwen-14B pairs. The Gemma pair, already at the floor, barely moves. Against **accumulation**, override rates prove uninformative: pinned at 100% for Qwen and OLMo both before and after tuning, this ceiling masks underlying shifts. This requires the margin metric (§3.1), which isolates contextual effects from parametric priors by measuring the swing from each model's no-evidence baseline. Evaluated on this margin, instruction tuning actually renders models *more* vulnerable to accumula-

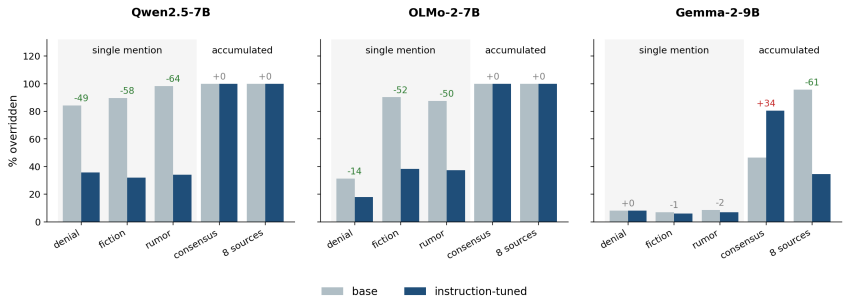

Figure 3: Three of the six base/instruct pairs. Bar labels show percentage-point changes (green = improvement); shading marks single-mention conditions. On the two vulnerable pairs, instruction tuning roughly halves every single-mention attack, while Gemma remains at the floor. Under accumulation, override rates prove uninformative: Qwen and OLMo hit a 100% ceiling before and after tuning, while Gemma moves by 34 (consensus) and -61 (eight sources) points.

tion in 10 of 12 pair-by-axis comparisons. For instance, OLMo's consensus swing increases from $+14.19$ to $+22.63$, with Qwen behaving similarly (the only two exceptions occur in eight-source cells). Relying solely on override rates artifactually suggests that accumulation remains unaffected. Consistent with (Itzhak et al., 2024), tuning teaches models to discount unsupported claims but defer heavily in corroborated ones. While sensible on the open web, this heuristic becomes a liability in retrieval: "instruction-tuned" does not mean a model is "more robust to contextual falsehoods."

4.4 What does not protect

Four properties that should make a fact harder to override do not. Replacing an arbitrary falsehood with a plausible near-miss shifts the margin just 0.87 log-odds against a 12.82 assertion swing: models respond to presentation, not intrinsic credibility. Corpus frequency (measurable only for OLMo) correlates with resistance at -0.49 pooled, but drops to -0.27 when controlling for relation, even reversing sign in our largest relation. Placing the claim after the question helps 7 of 11 models but reverses on OLMo-2-32B by 23.8 points. Finally, attention reduction to a refuted counterfactual remains small at all depths and fails to order models by resistance. Attention mass is blunt: early representational shifts can be consumed by later MLPs without altering macroscopic attention patterns. Moreover, Gemma's unpatched margin never turns positive under refutation, leaving no restoration threshold for causal patching. Our probes establish where behavioral divergence *finalizes* (the network's last third) but cannot rule out early suppression (Appendices C-E).

5 Discussion

Surfacing a debunking document inadvertently injects the target claim, and standard mitigations fail. Retrieving more evidence is ineffective; a single mention suffices. Source-authority filtering also fails: the weakest credibility rung overrides models failing the polarity control, while authoritative framing defeats even the resistant Gemma scales. The sole prompt-level defense (§4.4) is impractical behind APIs. Remaining interventions target the model rather than the context, leaving the critical vulnerability exposed: because a ranker, not reality, dictates passage count, retrieval systems artificially manufacture the exact accumulation attack that defeats every evaluated model. Since these findings rely on simulated inputs, not real-world measurements, future work must test these ladders inside deployed retrieval pipelines where the ranker actively sets the dose.

Limitations. All values are raw-completion first-token margins. While these track full-string generation (§3.5), our rates represent lower bounds for chat deployments (§4.2). Absolute override rates depend as much on the template set as the model. Confidence intervals omit document-randomization variance (at most 1.1 points) and multiple-comparison corrections. Finally, knowledge relation is confounded with corpus frequency (measurable only for OLMo); documents use templated filler rather than authentic fact-checking pages; model scale caps at 32B parameters; and mechanistic probes remain correlational, except for causal activation patching.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing and to retrieve and discover related work. We used generative AI tools to support research execution: auditing the code and analysis pipeline, writing the verification, and cross-checking the reported numbers against the experiment output files. We also used generative AI tools to help construct the templated materials from which our synthetic evaluation documents are generated. We have not used generative AI tools for proving mathematical claims, which does not apply to this work. We have not used generative AI tools for research ideation. We have reviewed all AI-assisted work. The research idea, the original manuscript, and the experimental design are the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to make this work reproducible. The measurement is specified in §3: the margin and its scoring position (§3.1, Eq. 1), the document construction and its invariants (§3.2), the matched-prior subset (Eq. 2), the two claim-template sets (§3.4), and the validity checks applied before any analysis (§3.5). Appendix A lists the complete fact inventory and document construction, Appendix B gives the validity gate and the scoring failure it caught, and Appendices C-I describe every additional experiment, including the controlled-corpus positive control (Appendix F) and the prompting-regime study (Appendix I); Appendix L records the measurement errors we found in our own pipeline and the automated guards added to prevent them. The supplementary material contains the implementation code. All evaluated models are publicly available open-weight checkpoints.

REFERENCES

Victoria Basmova, Yoav Goldberg, and Reut Tsarfaty. Llms' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. *CoRR*, abs/2404.06283, 2024. doi: 10.48550/ARXIV.2404.06283. URL https://doi.org/10.48550/arXiv.2404.06283.

Shuaizhi Cheng, Xiang Shi, and Mingwei Li. The override gap: A magnitude account of knowledge conflict failure in hypernetwork-based instant LLM adaptation. *CoRR*, abs/2604.23750, 2026.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. In *ICLR*. OpenReview.net, 2025.

Kyubeen Han, Junseo Jang, Hongjin Kim, Geunyeong Jeong, and Harksoo Kim. Exploring the impact of instruction-tuning on llm's susceptibility to misinformation. In *ACL (1)*, pp. 26711–26731. Association for Computational Linguistics, 2025.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785, 2024.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *LREC/COLING*, pp. 16867–16878. ELRA and ICCL, 2024a.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *ACL (Findings)*, volume ACL 2024 of *Findings of ACL*, pp. 1193–1215. Association for Computational Linguistics, 2024b.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *ACL*, pp. 7811–7818. Association for Computational Linguistics, 2020.

Minsung Kim, Dong-Kyum Kim, Jea Kwon, Nakyeong Yang, Kyomin Jung, and Meeyoung Cha. How training data shapes the use of parametric and in-context knowledge in language models. In *ACL (1)*, pp. 23242–23257. Association for Computational Linguistics, 2026.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *EMNLP (1)*, pp. 7052–7063. Association for Computational Linguistics, 2021.

Xuhua Ma, Richong Zhang, and Zhijie Nie. Corect: Context-aware logit contrast for hidden state rectification to resolve knowledge conflicts. *CoRR*, abs/2602.08221, 2026.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL (1)*, pp. 9802–9822. Association for Computational Linguistics, 2023.

Harry Mayne, Lev McKinney, Jan Dubinski, Adam Karvonen, James Chua, and Owain Evans. Negation neglect: When models fail to learn negations in training. *CoRR*, abs/2605.13829, 2026.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *ACL (1)*, pp. 8420–8436. Association for Computational Linguistics, 2024.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In *ACL (1)*, pp. 29430–29459. Association for Computational Linguistics, 2026.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In **SEM@ACL*, pp. 101–114. Association for Computational Linguistics, 2023.

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 *ICLR*. OpenReview.net, 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *EMNLP*, pp. 8541–8565. Association for Computational Linguistics, 2024.