# TRUST AT THE MARGIN: PROVENANCE-VISIBILITY EFFECTS AND MODEL-DEPENDENT RISK LOCALIZATION IN LLM KNOWLEDGE CONFLICTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

LLMs increasingly rely on external information that may conflict with their parametric knowledge. Recent work demonstrates that provenance can influence conflict resolution, however, three questions remain: how provenance visibility affects model decisions, where resulting risk is concentrated, and how this information can guide mitigation. In this paper, we address these questions in three steps. First, we introduce PROVMARGIN, a benchmark that varies provenance visibility while holding others fixed and measures model epistemic margins before intervention. Second, we characterize provenance effects and their risk localization. Across eight models, removing provenance cues indicating low credibility increases false-answer adoption by 9.95%, with a larger effect near decision boundary. Extensions across model families, response measures, and tasks show that overall effect generalizes more consistently than its boundary localization. Real-world evaluations further reveal heterogeneous effects for erroneous claims, while official provenance consistently supports correct updates. Finally, we translate these findings into PROVROUTE, which combines provenance, epistemic margins, and observable responses to prioritize verification. At a budget of $1/6$ of requests, PROVROUTE reduces residual false-adoption risk to 27.56%, improving over the strongest static baseline by 2.07 %. Together, these results connect causal measurement, model-dependent risk localization, and selective verification, while emphasizing the need to reduce errors without hindering useful knowledge updates.

## 1 INTRODUCTION

LLM systems increasingly combine parametric knowledge with external information through retrieval-augmented generation (Lewis et al., 2020), tool use (Schick et al., 2023), long-term memory (Packer et al., 2023), multi-agent workflows (Wu et al., 2023). Although external information can improve factual accuracy and provide up-to-date knowledge, it may also conflict with what a model already knows, requiring the model to decide which information to trust. External information often includes *provenance*: cues about its origin, institutional affiliation, or authority. These cues can influence how model interprets otherwise identical content (Xu et al., 2024; Jin et al., 2024; Peng et al., 2025; Li et al., 2025; Khan et al., 2026). Reliable conflict resolution therefore requires models to weigh the evidence, the relevance and credibility of its provenance, and their own uncertainty.

Existing research has examined knowledge conflicts in LLMs through the influence of provenance cues (Schuster et al., 2026; Yao et al., 2026; Li et al., 2026), robustness to imperfect retrieval (Wang et al., 2025), and provenance-aware system design (Louck, 2026; Ouyang & Hou, 2026; Alvarez et al., 2026) (see Appendix A for a broader review). Despite this progress, three key questions remain. **First, causal identification:** prior studies show that provenance cues and citations can influence model responses (Khan et al., 2026; Schuster et al., 2026; Khurana et al., 2026), but how does removing explicit provenance information affect a model's choice when external information conflicts with its parametric knowledge? **Second, risk localization:** which localization are most sensitive to provenance visibility? An average effect cannot establish whether this sensitivity is greater near the model's decision boundary or far from the decision boundary. **Third, practical mitigation:** how can these findings reduce errors in deployment? Identifying a risk does not tell us which responses to verify, how many errors a fixed verification budget can prevent, or how much verification is needed.



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

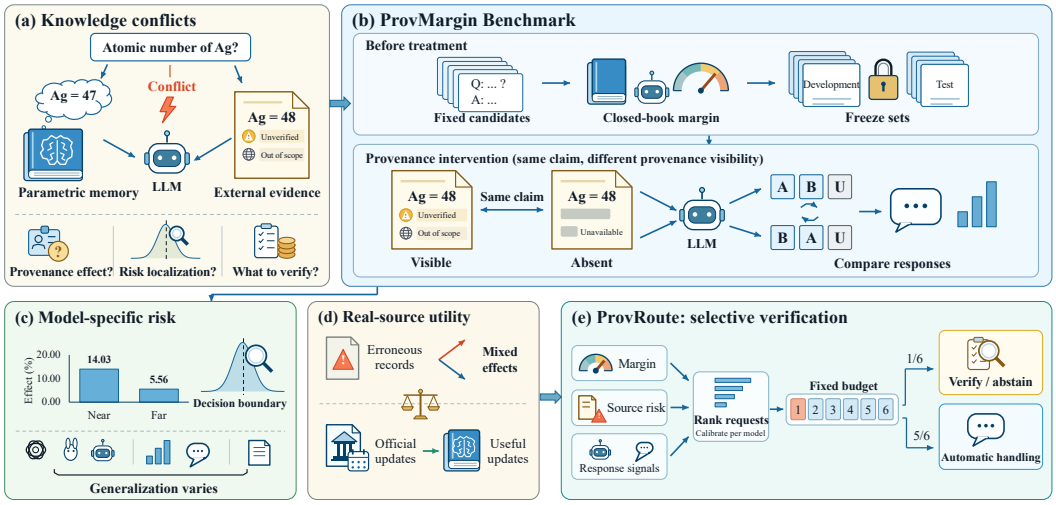


Figure 1: Overview of our framework. **(a)** Conflicts between parametric memory and external evidence. **(b)** PROVMARGIN freezes candidates, model-specific closed-book margins, and evaluation sets before comparing the same erroneous claim with weak-source provenance visible versus absent. **(c)** Provenance effects are different at decision boundary, while generalization varies across models, response metrics, and task requirements. **(d)** Real-source evaluations reveal heterogeneous effects for erroneous records and the utility of provenance for adopting correct official updates. **(e)** PROVROUTE combines margin, source risk, and observable response signals to prioritize verification or abstention under a fixed budget.

We address these questions through *Trust at the Margin*, a study of the provenance visibility effect and the model-dependent risk localization in parametric–contextual knowledge conflicts. Figure 1 summarizes our framework. Our study asks four research questions. **RQ1 (provenance visibility):** with external content held fixed, does provenance visibility causally affect the adoption of conflicting information? **RQ2 (boundary localization):** is this effect stronger near the model's decision boundary? **RQ3 (generalization and measurement):** which findings hold across models, response metrics, and task requirements? **RQ4 (safety and utility):** how can provenance-aware verification reduce the adoption of false information while preserving the benefits of reliable external evidence?

Specifically, the main contributions are summarized as follows:

1. **PROVMARGIN benchmark and measurement protocol.** To address RQ1, we introduce PROVMARGIN, which evaluates the effect of provenance visibility by varying provenance cues while holding all other fixed. It further measures each model's closed-book epistemic margin before the provenance intervention and evaluates responses using different metrics. This protocol avoids outcome-dependent instance selection and controls for content and answer-position confounds, and provides a reproducible estimate of how provenance visibility changes trust in conflicting external information.
2. **Boundary-localized provenance effects and their generalization limits.** To address RQ2, we use closed-book margins to group samples by their distance from each model's decision boundary. Across the four models (Qwen-3 4B and Qwen-3 8B (Yang et al., 2025), Granite-3.3 8B (IBM Granite Team, 2025), and Mistral-7B (Jiang et al., 2023)), provenance visibility has a consistently larger effect near the boundary than farther away. To address RQ3, we extend the evaluation to Phi (Abdin et al., 2024; Abouelenin et al., 2025), Gemma (Gemma Team et al., 2025), OLMo (Team OLMo et al., 2025), and DeepSeek (Guo et al., 2025), as well as alternative metrics and task instructions. The overall provenance effect generalizes more reliably than its boundary localization, whose strength and direction vary across settings. These findings support the epistemic margin as a model-specific indicator of risk, with limited transferability across models.
3. **Safety–utility tradeoffs on real-world information.** To address RQ4, we evaluate provenance visibility on erroneous real-world records and correct first-party updates. Effects on

erroneous records vary without a consistent aggregate direction, whereas complete official provenance consistently improves the adoption of correct updates. These findings show that provenance can support useful knowledge updates while also influencing the adoption of false information, motivating selective verification.

4. **Provenance-aware selective verification under a fixed budget.** To further address RQ4, we propose PROVROUTE, which combines epistemic margins, provenance risk indicators, and observable response signals to prioritize requests for verification. Under a fixed $1/6$ verification budget, PROVROUTE reduces residual false-adoption risk to 27.56%, outperforming the strongest static margin–provenance baseline by 2.07% and also improving on random review.

## 2 PROBLEM SETUP AND EVALUATION METRICS

We study how provenance visibility affects a fixed model $f_m$ when an external claim conflicts with an audited factual answer. Each instance contains a query, a gold answer, and a plausible but false alternative supported by an external record. We evaluate each instance using fixed prompt variants and both answer orders. For example, for a question about the capital of France, Paris is the gold answer and Lyon is a distractor, their A/B labels are reversed in the mirrored presentation.

### 2.1 KNOWLEDGE-CONFLICT INSTANCES AND THE PROVENANCE INTERVENTION

Each controlled instance is represented as

$$\mathcal{I}_i = (q_i, g_i, e_i, c_i), \tag{1}$$

where $q_i$ is a factual query, $g_i$ is the audited gold answer, $e_i \neq g_i$ is a plausible distractor, and $c_i$ is an external record supporting $e_i$. We vary provenance visibility through two conditions, $z \in \{V, A\}$. Under $V$ (*visible*), explicit provenance cues indicate low credibility, limited authority, or expertise outside the query's domain. Under $A$ (*absent*), the corresponding identity, authority, scope, and origin-trace information is unavailable. The query, claim, answer candidates, presentation order, and all non-provenance text remain fixed, with neutral padding used to control prompt length. The paired comparison therefore isolates the effect of provenance visibility while preserving the external claim.

Let $Y_{imr}(z) \in \{0, 1\}$ indicate whether model $m$ adopts the external answer $e_i$ under provenance condition $z$ and repeated-measurement condition $r \in \mathcal{R}$. In the primary A/B/U protocol, A and B identify the two answer candidates, and U indicates uncertainty. After accounting for the mirrored answer order, we define

$$Y_{imr}(z) = \mathbf{1}\{f_m(q_i, c_i, z; r) = e_i\}\,. \tag{2}$$

Uncertain responses, refusals, and responses that cannot be parsed count as non-adoption and remain in the denominator. Two answer orders and three prompt variants yield $|\mathcal{R}| = 6$ measurements per provenance condition. The paired effect for instance $i$ and model $m$ is

$$\delta_{im} = \bar{Y}_{im}(A) - \bar{Y}_{im}(V), \qquad \bar{Y}_{im}(z) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} Y_{imr}(z). \tag{3}$$

Our primary estimand is the family-balanced provenance effect:

$$\Delta_{\text{prov}} = \sum_{m \in \mathcal{M}} w_m \, \mathbb{E}_i[\delta_{im}], \qquad \sum_m w_m = 1. \tag{4}$$

In the prespecified four-model evaluation, each Qwen model receives weight $1/6$, while Granite and Mistral each receive weight $1/3$, giving equal total weight to each model family. Because $e_i$ is false, $\Delta_{\text{prov}} > 0$ means that removing provenance cues indicating low credibility causally increases false-answer adoption. The primary metric is the false-answer selection rate in the mirrored A/B/U task. Following free-generation analyses of knowledge conflicts (Longpre et al., 2021), we also evaluate generated answers using the same semantic definition of adoption as a supplementary check.

Table 1: The main evaluation metrics.

| Metric | Contrast | Positive value means |
|---|---|---|
| $\Delta_{\text{prov}}$ | absent − visible false adoption | provenance removal increases errors |
| $\Delta_{\text{NF}}$ | near effect − far effect | the signed effect is larger near the boundary |
| $\text{SP}_{AB}$ | visible − absent external probability | visibility favors the external answer |
| $\Delta_{\text{err}}$ | stripped − full false adoption | provenance removal increases errors |
| $\Delta_{\text{upd}}$ | full − stripped correct adoption | provenance supports correct updates |
| $\Psi$ | PROVROUTE − lite residual risk | PROVROUTE leaves more residual risk |

### 2.2 PRE-TREATMENT EPISTEMIC MARGIN

Before either provenance intervention, we measure the model's preference between the gold answer and the distractor without external information. Let $C_o(y)$ denote the representation of answer $y$ under answer order $o \in \{g{\to}A, g{\to}B\}$, using either its option label or answer text. Let $s_m(C_o(y) \mid q_i)$ be its closed-book log-likelihood score under a fixed scoring protocol. The model-specific epistemic margin is

$$M_{im} = \frac{1}{2} \sum_{o} \left[ s_m(C_o(g_i) \mid q_i) - s_m(C_o(e_i) \mid q_i) \right]. \tag{5}$$

Positive $M_{im}$ indicates a preference for the gold answer, while negative $M_{im}$ indicates a preference for the distractor. We define boundary coordinate as $D_{im} = |M_{im}|$: smaller values indicate a weaker preference between two answers and thus greater proximity to the pairwise decision boundary.

Within each prespecified evaluation set, we rank instances separately for each model by $D_{im}$, breaking ties by instance index $i$, and divide them into equally sized near, middle, and far strata.

For stratum $s \in \{\text{near}, \text{middle}, \text{far}\}$, let $\Delta_{s,m} = \mathbb{E}_i[\delta_{im} \mid S_{im} = s]$, where $S_{im}$ is the stratum assigned to instance $i$ for model $m$. We measure boundary localization through the family-balanced near–far contrast:

$$\Delta_{\text{NF}} = \sum_{m} w_m \left( \Delta_{\text{near},m} - \Delta_{\text{far},m} \right). \tag{6}$$

A positive $\Delta_{\text{NF}}$ indicates a larger signed provenance-removal effect near the decision boundary; a negative value indicates a larger effect farther away. We follow a prespecified testing sequence: we first test the overall provenance effect in Equation 4, and test the near–far contrast in Equation 6 only if the first test meets the prespecified significance criterion.

These two estimands form the confirmatory analysis. Additional metrics assess deterministic label probabilities, task–provenance interactions, safety–utility tradeoffs on real-world records, and selective-verification performance. Appendix B provides their definitions and assumptions. Table 1 summarizes the sign conventions used throughout the paper.

## 3 THE PROVMARGIN BENCHMARK AND EXPERIMENTAL PROTOCOL

We construct PROVMARGIN before any provenance intervention: we fix factual candidates, measure model-specific closed-book margins, and select development and test sets without using treatment outcomes. We then evaluate each selected instance under paired provenance conditions with identical claim content (Figure 1b).

### 3.1 CANDIDATE CONSTRUCTION AND AUDIT

We use a frozen August 3, 2026 snapshot of truthy statements from the Wikidata Query Service (Vrandečić & Krötzsch, 2014). Of 17 queried relations, 13 yield sufficient usable records: asteroid discoverer, building architect, chemical-element atomic number, company founder, film director, mountain range, novel author, programming-language designer, river mouth, spacecraft manufacturer, television-series network, university founder, and video-game developer. We retain facts with a usable English item label and exactly one distinct value for the target relation, excluding semantic overlap with an earlier benchmark.

For each fact, we select three false alternatives from other items with the same relation and value type. We prioritize similar Wikidata sitelink counts for entity distractors and numerical proximity for integer

Table 2: Examples from PROVMARGIN, shown with the gold answer assigned to A. Claim content is identical across provenance conditions.

| Domain | Question | A: gold | B: external | Shared claim |
|---|---|---|---|---|
| Astronomy | Who discovered asteroid 102 Miriam? | C. H. F. Peters | G. Schiaparelli | 102 Miriam was discovered by G. Schiaparelli. |
| Chemistry | What is the atomic number of silver? | 47 | 48 | The atomic number of silver is 48. |

Table 3: Provenance-visibility effects (%). Brackets denote 95% CI for locked-test estimates.

| Model | Dev. $\Delta_{\text{prov}}$ | Dev. $\Delta_{\text{NF}}$ | Locked $\Delta_{\text{prov}}$ | Locked $\Delta_{\text{NF}}$ |
|---|---|---|---|---|
| Qwen-3 4B | $+9.35$ | $+14.72$ | $+9.17\ [+4.72, +13.61]$ | $+11.67\ [+0.68, +22.65]$ |
| Qwen-3 8B | $+12.31$ | $+11.11$ | $+14.44\ [+9.72, +19.17]$ | $+9.17\ [-0.49, +18.82]$ |
| Granite-3.3 8B | $+6.48$ | $+11.39$ | $+8.89\ [+5.13, +12.65]$ | $+7.50\ [-1.98, +16.98]$ |
| Mistral-7B | $+5.93$ | $+4.44$ | $+9.17\ [+5.76, +12.57]$ | $+7.50\ [-0.38, +15.38]$ |
| Family-balanced | $\mathbf{+7.75}$ | $\mathbf{+9.58}$ | $\mathbf{+9.95}\ [\mathbf{+7.55}, \mathbf{+12.35}]$ | $\mathbf{+8.47}\ [\mathbf{+3.70}, \mathbf{+13.24}]$ |

distractors, breaking ties with deterministic hashes. This yields 306 facts and 918 fact–distractor candidates. Each distractor differs from the gold answer while remaining compatible with the relation. At most one candidate per fact is included across the final development and test sets.

### 3.2 DEVELOPMENT AND TEST SPLITS

Each model scores all 918 candidates using a fixed closed-book pairwise prompt. We average the gold-minus-distractor likelihood margins across both answer orders and retain their range to measure order sensitivity. We divide candidates into signed-margin tertiles for each model and use mixed-integer linear programming to select a shared 240-candidate set. The optimization requires distinct facts, 80 candidates per signed-margin tertile for every model, and 17–20 candidates per domain, while favoring lower order sensitivity and higher-ranked distractors. The development split contains 180 candidates, with 60 per signed tertile for each model and 12–16 per domain. The remaining 60 form the locked test set, with 20 per signed tertile. The splits share no facts, and selection uses no provenance-conditioned responses.

After fixing both splits, we separately rank instances within each split and model by $D_{im} = |M_{im}|$ to define near, middle, and far tertiles. Each stratum contains 60 development instances and 20 test instances per model. These strata are model-specific: the same instance may be near the decision boundary for one model and far from it for another.

### 3.3 CONTENT-MATCHED PROVENANCE INTERVENTION

Each instance is evaluated under both provenance conditions. Visible provenance identifies an unverified community discussion post, indicates that the topic falls outside its declared expertise, preserves its origin trace, and signals low topical authority. With provenance absent, the same claim appears as an imported memory with its origin, scope, trace, and authority unspecified. The controlled intervention uses no real publisher or institution names. Across conditions, we hold fixed the query, answer candidates, claim, record layout, and claim position. Table 2 illustrates two development instances. Both answer orders are evaluated.

Appendix C.1, C.2, C.3 detail response protocols, model coverage, and statistical analysis.

## 4 VALIDATING PROVENANCE-VISIBILITY EFFECTS

We test whether removing provenance cues indicating low credibility increases false-answer adoption. Positive $\Delta_{\text{prov}}$ indicates increased false adoption, positive $\Delta_{\text{NF}}$ indicates a larger effect near the model-specific closed-book decision boundary.

Table 3 reports the overall and near–far effects under mirrored A/B/U evaluation on the 180-instance development set and the 60-instance test set. The family-balanced overall and near–far effects are positive on both sets. On the test set, removing provenance increases false adoption by 9.95% (95% CI $[7.55, 12.35]$, $p = 1.74 \times 10^{-11}$), with a near–far contrast of 8.47% (95% CI $[3.70, 13.24]$, $p = 7.58 \times 10^{-4}$). All four models have positive overall and near–far point estimates on both sets, although three model-specific locked-test near–far intervals include zero. These results support a provenance-visibility effect and greater near-boundary risk in the pooled four-model evaluation.

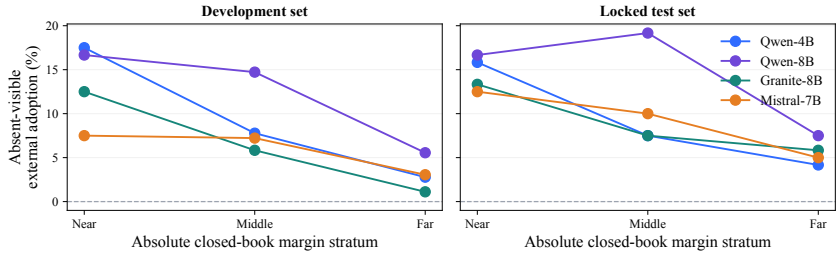


Figure 2: Provenance-visibility effects across model-specific closed-book margin.

Table 4: Cross-model and scale extensions on two fixed 60-instance sets (%).

| Model | Locked $\Delta_{\text{prov}}$ | Independent $\Delta_{\text{prov}}$ | Locked $\Delta_{\text{NF}}$ | Independent $\Delta_{\text{NF}}$ |
|---|---|---|---|---|
| Phi-4-mini, 3.8B | $+6.94$ $[+4.17, +9.72]$ | $+4.72$ $[+1.39, +8.33]$ | $+2.50$ $[-4.17, +9.17]$ | $-0.83$ $[-8.33, +6.67]$ |
| Phi-4, 14B | $+5.28$ $[+2.22, +8.89]$ | $+9.72$ $[+5.28, +15.00]$ | $+13.33$ $[+5.83, +22.50]$ | $+20.83$ $[+10.83, +31.67]$ |
| Gemma-3, 27B | $+13.61$ $[+9.44, +17.78]$ | $+15.56$ $[+11.94, +19.17]$ | $+13.33$ $[+3.33, +22.50]$ | $+5.83$ $[-2.50, +14.17]$ |
| OLMo-2, 32B | $+6.94$ $[+4.44, +9.44]$ | $+8.06$ $[+5.28, +11.11]$ | $-8.33$ $[-13.33, -4.17]$ | $-12.50$ $[-19.17, -6.67]$ |

### 4.1 Risk localization along the closed-book margin

Figure 2 shows provenance-visibility effects across model-specific near, middle, and far margin tertiles. All four models have larger near- than far-stratum effects on both sets. On the test set, the family-balanced effect is 14.03 % near the boundary and 5.56 % in the far stratum. Thus, provenance removal has a greater effect when competing answers receive similar closed-book support, but remains harmful in aggregate even when prior preferences are more decisive.

### 4.2 Cross-model and scale extensions

We extend the evaluation to Phi-4-mini, Phi-4, Gemma-3 27B, and OLMo-2 32B on the original locked test set and a separate, non-overlapping 60-instance set. The results are shown in Table 4.

All eight overall-effect intervals lie above zero, extending the evidence that provenance removal increases false adoption across model families and scales. Boundary localization, however, varies by model. Phi-4 has positive near–far contrasts with intervals excluding zero on both sets. Gemma-3 has positive estimates on both sets, but only the locked-set interval excludes zero. Phi-4-mini shows no consistent localization: its estimates change sign and both intervals include zero. OLMo-2 exhibits the reverse pattern, with both near–far intervals entirely below zero. Its near-stratum effect is 0.83 % on both sets, compared with far-stratum effects of 9.17 and 13.33 %. Provenance removal therefore increases false adoption more strongly farther from its decision boundary.

Overall, the provenance-visibility effect generalizes more consistently than its boundary localization. The epistemic margin can help rank risk within a model, but its relationship to provenance sensitivity requires validation for each target model.

## 5 Factors Influencing Provenance-Visibility Effects

Following the results in Section 4, we examine how provenance-visibility effects vary across response measures and task requirements.

**Sensitivity to response measures.** Figure 3 a compares boundary localization across response measures. The near–far contrast ranges from 4.44 to 14.72 % under mirrored A/B/U choice, but from 0.00 to 10.56 % under free generation. The corresponding contrasts in continuous gold-margin scores are 9.21, 3.03, 0.72, and $-0.05$ for Qwen-4B, Qwen-8B, Granite, and Mistral, respectively, in model-specific units. Thus, both the magnitude and localization of provenance-visibility effects depend on the response measure, patterns observed in constrained choice do not necessarily carry over to free generation.

**Parser-free label probability.** To assess whether the effect persists without response parsing, we remove the abstention option and directly compare the probabilities assigned to the gold and erroneous answer labels after A/B renormalization. Visible provenance indicating low credibility reduces support for the erroneous answer in all four original models (Figure 3b). The visible-minus-

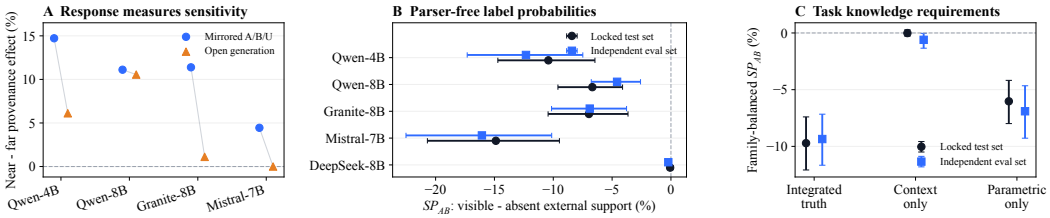


Figure 3: Factors influence provenance-visibility effects.

Table 5: Real-world provenance effects erroneous records and correct updates.

| Record group and metric | Estimate | 95% CI |
|---|---|---|
| Erroneous (18), A/B/U $\Delta_{\text{err}}$ | $-4.48$ | $[-10.03, +1.39]$ |
| Erroneous (18), continuous | $+0.025$ | $[-0.136, +0.222]$ |
| Erroneous (18), free generation | $+6.48$ | $[-3.09, +17.59]$ |
| Erroneous (18), A/B/U near–far | $+7.41$ | $[-6.02, +20.37]$ |
| Official updates (12), A/B/U $\Delta_{\text{upd}}$ | $\mathbf{+24.77}$ | $[\mathbf{+18.75}, \mathbf{+30.79}]$ |
| Official updates (12), continuous | $\mathbf{+0.284}$ | $[\mathbf{+0.216}, \mathbf{+0.353}]$ |
| Official updates (12), free generation | $\mathbf{+10.19}$ | $[\mathbf{+3.24}, \mathbf{+18.52}]$ |

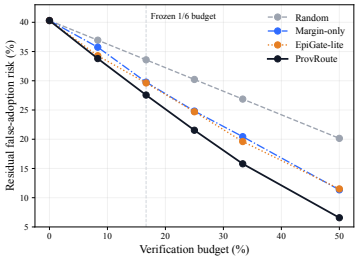


Figure 4: Sensitivity to budgets.

absent contrast, $\text{SP}_{AB}$, ranges from $-14.88$ to $-6.68$ % on the test set and from $-16.07$ to $-4.58$ % on the independent set, all 95% confidence intervals exclude zero. The overall effect is therefore also present in label probabilities and cannot be explained solely by generation or parsing behavior.

**Task knowledge requirements.** We compare three task instructions (Figure 3c): `integrated_truth` asks models to combine parametric knowledge with context to identify the factually correct answer; `context_only` asks them to reproduce the context's claim; and `parametric_only` asks them to ignore the context.

Under `integrated_truth`, visible provenance indicating low credibility reduces the erroneous answer's probability by approximately 9–10 %, the largest effect among the three tasks. The effect is close to zero under `context_only`, consistent with its focus on reproducing the supplied claim. A shift of approximately 6–7 % persists under `parametric_only`, suggesting that contextual information continues to influence predictions despite instructions to ignore it.

Together, these results show that provenance sensitivity depends on how responses are measured and what the task requires. Direct probability measurements support the overall effect, while its magnitude and boundary localization remain specific to the model and evaluation setting.

## 6 Real-World Provenance: Safety and Update Utility

We assess whether the controlled findings extend to authentic provenance using 30 fixed records: 18 erroneous claims and 12 correct first-party updates that may postdate the models' parametric knowledge. Each claim is evaluated with full or stripped provenance. Full provenance includes the page title, publisher role, timestamp or revision date, permanent URL, and topical authority. The stripped condition replaces these fields with metadata-unavailable placeholders and tokenizer-calibrated neutral padding, preserving the claim content.

For erroneous records, $\Delta_{\text{err}}$ measures the increase in false adoption after provenance removal (Equation 12). For correct updates, $\Delta_{\text{upd}}$ measures the increase in correct adoption with full provenance (Equation 13). Thus, positive values indicate harm from removal and utility from visibility, respectively. We evaluate both groups using mirrored A/B/U choice, free generation, and continuous likelihood margins.

### 6.1 Erroneous Records Show No Consistent Aggregate Effect

For erroneous records, the family-balanced A/B/U effect is $-4.48$ % (95% CI $[-10.03, 1.39]$). The continuous and free-generation estimates also include zero, as does the near–far contrast (Table 5).

This panel therefore does not establish a consistent aggregate direction or reliable boundary localization. The effect observed with controlled low-credibility provenance does not transfer uniformly to authentic provenance.

### 6.2 Official provenance supports correct updates

For correct first-party updates, full provenance increases A/B/U correct adoption by 24.77%, with positive effects for all 12 records. Continuous margins and free generation support the same conclusion, with positive means in every confirmatory model family.

These gains show that provenance can help models accept useful external information, including updates that may conflict with outdated parametric knowledge. We also distinguish factual accuracy from faithfulness to the external claim: adopting an erroneous claim reduces accuracy, whereas adopting a correct update improves it. The pooled decomposition shows that full provenance reduces abstention and increases both accuracy and claim faithfulness, but these aggregate changes must be interpreted by record correctness (Appendix D).

### 6.3 Implications for provenance-aware verification

The two groups reveal a practical safety–utility tradeoff. Removing provenance can discard information that helps models recognize valid updates, while preserving provenance alone does not reliably prevent false adoption. Accordingly, provenance visibility should not serve as a universal acceptance or rejection rule.

A useful design principle is to assess what provenance establishes for the specific claim. Institutional identity indicates who issued a record, topical authority indicates relevant expertise, and timestamps help determine whether the record may supersede older knowledge. These cues can inform trust, but none independently establishes factual correctness. Systems should therefore preserve provenance and use it alongside claim content and model uncertainty when deciding which responses require verification.

Evaluation should likewise measure both errors prevented and correct updates retained. A policy that lowers false adoption by broadly rejecting external information may also reduce the ability to incorporate new knowledge. This motivates the selective-verification approach in the next section: use provenance and model-specific signals to allocate a limited verification budget.

## 7 Selective Verification with ProvRoute

The preceding results motivate selective verification: provenance removal can increase false adoption, but risk varies across models, and official provenance can support correct updates. A useful routing policy should therefore consider both provenance and the model's response to the external claim.

We propose ProvRoute, which ranks requests before release and routes the highest-priority fraction to `VERIFY_OR_ABSTAIN` under a fixed budget. It selects responses for review without changing their answers. Appendices E.1 and E.2 define the deployment objective and equal-budget baselines.

### 7.1 The ProvRoute routing score

ProvRoute combines pre-treatment features with observable responses to the external claim (Figure 1e). Let $B_u$ denote model-specific boundary closeness, $A_u$ indicate provenance absence, and $H_u$ identify missing or low-authority provenance. Two mirrored A/B/U probes provide the external-answer adoption rate $P_u$ and uncertainty rate $U_u$, while reducing answer-position bias. The fixed routing score is

$$s_{\mathrm{PR}}(u) = 10P_uH_u + B_u + 0.25U_u + 0.10A_u. \tag{7}$$

The adoption term prioritizes requests in which the model follows external content with risky provenance. Boundary closeness, uncertainty, and provenance absence refine the ranking. The mask $H_u$ ensures that adopting a well-formed authoritative update does not itself trigger the adoption penalty. Requests are ranked within each model, and the top fraction $b$ is routed for review. Routing uses neither audited truth labels nor held-out evaluation responses.

Table 6: Selective verification at budget $b = 1/6$. Results are family-balanced. Residual risk is measured over all requests, selective risk is measured over the 83.33% handled automatically.

| Policy | Residual risk | Selective risk | Risk reduction | Relative reduction | Residual prov. effect |
|---|---|---|---|---|---|
| Random | 33.58 | 40.30 | 6.72 | 16.67 | +8.91 |
| Margin-only | 29.77 | 35.72 | 10.53 | 26.13 | +7.69 |
| EpiGate-lite | 29.63 | 35.56 | 10.67 | 26.48 | −1.20 |
| PROVROUTE | **27.56** | **33.07** | **12.74** | **31.62** | **−2.15** |

### 7.2 MAIN MITIGATION RESULTS

We evaluate the fixed policy at $b = 1/6$ on 60 new facts whose outcomes were unseen during method development. Cross-fitted probes separate routing responses from evaluation responses (Appendix E.3). To isolate routing quality, we assume that verification successfully prevents false adoption in reviewed requests.

At the same verification budget, PROVROUTE achieves the lowest residual risk among the evaluated policies, reducing false adoption from 40.30% to 27.56% (Table 6). This is a 2.07 % improvement over the strongest static baseline, EpiGate-lite, and prevents 31.62% of baseline errors, compared with 16.67% for random review. PROVROUTE also has lower residual-risk point estimates than all baselines at every tested nonzero budget from $1/12$ to $1/2$ (Figure 4).

### 7.3 PRACTICAL IMPLICATIONS AND SCOPE

At the observed error rate, reviewing about 167 of every 1,000 requests would prevent approximately 127 false adoptions with PROVROUTE, compared with 67 under random review, assuming successful verification.

This gain does not make the remaining responses uniformly reliable: false-adoption risk among automatically handled requests is still 33.07%. The verification budget must therefore reflect the application's tolerance for residual errors. Likewise, the ideal-verifier assumption means that these results measure routing quality; realized benefits also depend on verifier accuracy and the cost of probing and review.

Deployment further requires model-specific validation. The incremental gains and the value of a second mirrored probe vary across models (Appendix E.4). A practical policy should assess these gains against inference cost and check whether additional review preserves the adoption of correct updates (Appendix E.5). Thus, PROVROUTE provides a model-specific approach to prioritizing verification.

## 8 DISCUSSION

Our findings highlight provenance-visibility effect, model-dependent risk localization, and the need to balance error reduction with useful knowledge updates. Appendix F discusses these implications, measurement limitations, and the verification costs and assumptions that constrain deployment.

## 9 CONCLUSION

In this paper, to explore how provenance visibility shapes LLM knowledge conflicts and how to mitigate the resulting risks, we propose PROVMARGIN, a benchmark that isolates provenance effects while measuring models' pre-treatment epistemic margins. Our evaluation indicates that removing provenance cues signaling low credibility increases false adoption, while the localization of this effect near the decision boundary depends on the model and evaluation setting. Real-world results further show that official provenance supports correct updates, highlighting the need to balance safety and utility. To translate these findings into mitigation, we propose PROVROUTE, which combines provenance, epistemic margins, and observable responses to improve verification allocation under a fixed budget. Together, these contributions connect causal evaluation of provenance-visibility effects with model-dependent risk localization and selective verification for LLM knowledge conflicts.

### AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for generating synthetic data sets and assisting with translation. We have not used generative AI tools for designing or providing feedback on research methodology or experiments, formulating mathematical claims, providing critical ingredients for proving mathematical claim and helping develop theoretical models or conceptual frameworks are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

### REPRODUCIBILITY STATEMENT

More detailed information can ne found in Appendix.

## REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Ander Alvarez, Santhiya Rajan, Samuel Mugel, and Román Orús. Provenanceguard: Source-aware factuality verification for mcp-based llm agents. *arXiv preprint arXiv:2606.18037*, 2026.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models. In *International Conference on Learning Representations*, volume 2026, pp. 106786–106808, 2026.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26969–26983, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

IBM Granite Team. Granite-3.3-8B-Instruct, 2025. URL `https://huggingface.co/ibm-granite/granite-3.3-8b-instruct`.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Zhengda Jin, Bingbing Wang, Jing Li, Ruifeng Xu, and Min Zhang. Mitigating provenance-role collapse in long-term agents via typed memory representation. *arXiv preprint arXiv:2605.25869*, 2026.

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 *Proceedings of LREC-COLING*, 2024.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5684–5696, 2020.

Mohammad Aflah Khan, Mahsa Amani, Soumi Das, Bishwamittra Ghosh, Qinyuan Wu, Krishna Gummadi, Manish Gupta, and Abhilasha Ravichander. In agents we trust, but who do agents trust? latent source preferences steer llm generations. In *International Conference on Learning Representations*, volume 2026, pp. 115154–115207, 2026.

Aryan Khurana, Aravind Ramana RN, and Dhruv Kumar. Authority, truth, and citation bias: A large-scale multi-domain benchmark for studying epistemic susceptibility in large language models. *arXiv preprint arXiv:2606.13104*, 2026.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Shuowei Li, Haoxin Li, Wenda Chu, and Yi Fang. How large language models balance internal knowledge with user and document assertions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25323–25346, 2026.

Yuxuan Li, Xinwei Guo, Jiashi Gao, Guanhua Chen, Xiangyu Zhao, Jiaxin Zhang, Quanying Liu, Haiyan Wu, Xin Yao, and Xuetao Wei. Llms trust humans more, that'sa problem! unveiling and mitigating the authority bias in retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28844–28858, 2025.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of EMNLP*, 2021.

Yedidel Louck. Securing LLM-agent long-term memory against poisoning: Non-malleable, origin-bound authority with machine-checked guarantees. *arXiv preprint arXiv:2606.24322*, 2026.

Ciyan Ouyang and Rui Hou. MemLineage: Lineage-guided enforcement for LLM agent memory. *arXiv preprint arXiv:2605.14421*, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Miao Peng, Nuo Chen, Jianheng Tang, and Jia Li. How does misinformation affect large language model behaviors and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13711–13748, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? llm source preferences under knowledge conflicts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29430–29459, 2026.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Jonas Wallat, Wolfgang Nejdl, and Sandipan Sikdar. When facts change: Temporal knowledge conflict resolution in llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2154–2184, 2026.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of EMNLP*, 2024.

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

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

Min-Hsuan Yeh, Sharon Li, and Tanwi Mallick. Lumina: Detecting hallucinations in rag system with context–knowledge signals. In *International Conference on Learning Representations*, volume 2026, pp. 151430–151448, 2026.

# A Related Work

**Knowledge conflicts and provenance cues.** Knowledge-conflict research examines how LLMs reconcile parametric memory with contradictory context (Longpre et al., 2021; Xu et al., 2024). Prior evaluations identify preferences driven by evidence frequency, agreement with memory, and misinformation style (Jin et al., 2024). Real-document and temporal-update studies also show why retaining a memorized answer need not be correct (Wallat et al., 2026). Provenance identity adds a distinct influence: authority-bias and three-source studies compare reliance on user assertions, documents, and internal knowledge (Li et al., 2025; 2026), while controlled experiments reveal latent publisher preferences and repetition-sensitive preferences among synthetic sources (Khan et al., 2026; Schuster et al., 2026). AuthorityBench evaluates perceived domain and entity authority (Yao et al., 2026), and citation-veracity experiments examine attribution-induced errors (Khurana et al., 2026). These findings establish that provenance can influence responses. ProvMargin asks a more specific question: how does removing explicit weak-source provenance change adoption of the *same* conflicting claim? Its paired comparison holds content fixed and uses outcome-independent instance selection.

**Uncertainty and risk localization.** Semantic uncertainty measures disagreement among semantically distinct answers (Kuhn et al., 2023). In RAG, LUMINA detects hallucinations using signals of contextual and internal knowledge utilization (Yeh et al., 2026), while CK-PLUG measures context-induced entropy changes and adjusts decoding to control knowledge reliance (Bi et al., 2026). These methods motivate using model signals to assess unreliable decisions, but predicting an error and locating sensitivity to a provenance intervention are different objectives. We measure the closed-book margin *before* either provenance condition and use it to stratify paired effects within each model. This separates the overall visibility effect from its near–far interaction and tests whether boundary localization transfers across models, rather than assuming a universal confidence scale.

**Conflict mitigation and selective verification.** ASTUTE RAG consolidates internal and retrieved knowledge with source awareness (Wang et al., 2025); CARE learns a context assessor through grounded and adversarial soft prompting (Choi et al., 2025). Complementary agent-memory approaches preserve evidence structure: MemLineage tracks derivation lineage, origin-bound authority limits authority laundering, and MemIR separates evidence, retrieval cues, and supported claims (Ouyang & Hou, 2026; Louck, 2026; Jin et al., 2026). ProvenanceGuard verifies atomic claims against source-specific tool evidence and checks their attribution (Alvarez et al., 2026). Our behavioral intervention examines what happens when provenance is visible to the model, it does not authenticate sources or establish memory-integrity guarantees. For mitigation, ProvRoute builds on the risk–coverage perspective of selective classification and selective question answering (Kamath et al., 2020). It combines margins, source-risk indicators, and observable response signals to prioritize verification after a decision and before release.

# B Additional Evaluation Metrics and Definitions

This appendix gives the complete definitions of the supportive evaluation metrics introduced briefly in Section 2.

## B.1 Deterministic A/B/U label-logit metrics

To isolate option preferences from free generation and parsing, we read the next-token logits for A, B, and U at the start of the assistant response. For $\mathcal{K} \in \{\{A, B\}, \{A, B, U\}\}$,

$$p^{\mathcal{K}}_{imor}(k \mid z) = \frac{\exp \ell_{imor}(k \mid z)}{\sum_{j \in \mathcal{K}} \exp \ell_{imor}(j \mid z)}, \tag{8}$$

where $o$ is the mirrored order. We map option probabilities to the semantic gold and external answers before averaging, and define

$$\mathrm{SP}_{AB} = \mathbb{E}_{w,i,o}\left[p^{AB}(e_i \mid V) - p^{AB}(e_i \mid A)\right], \tag{9}$$

with $\mathrm{SP}_{ABU}$ defined analogously while retaining U in the normalization. Because Equation 9 uses visible-minus-absent, a negative $\mathrm{SP}_{AB}$ is directionally consistent with a positive harmful effect in

Equation 4. We additionally report $\Delta P(U) = \mathbb{E}_{w,i,o}[p^{ABU}(U \mid V) - p^{ABU}(U \mid A)]$, the directional KL divergence $D_{\mathrm{KL}}(p_V^{ABU} \| p_A^{ABU})$, the symmetric Jensen–Shannon divergence, and target-wise negative-log-likelihood shifts

$$\Delta\mathrm{NLL}_k = \mathbb{E}_{w,i,o}\left[-\log_2 p^{ABU}(k \mid V) + \log_2 p^{ABU}(k \mid A)\right], \quad k \in \{g_i, e_i\}. \tag{10}$$

In particular, $\Delta\mathrm{NLL}_g < 0$ means visible provenance increases probability on the gold answer. Reporting A/B- and A/B/U-normalized quantities separates discrimination between the factual candidates from willingness to answer.

### B.2 Task–provenance interaction

We evaluate the same record under three instructions. $T = \mathrm{int}$ asks the model to integrate parametric and contextual evidence for factual accuracy; $T = \mathrm{ctx}$ asks what the record states; and $T = \mathrm{par}$ asks the model to ignore the record and answer from parametric knowledge. The primary interaction is

$$\Gamma_{\mathrm{ctx-int}} = \mathrm{SP}_{AB}^{\mathrm{ctx}} - \mathrm{SP}_{AB}^{\mathrm{int}}. \tag{11}$$

A positive value means provenance suppresses the false external answer more strongly when the task requires evidence integration than when it requires record reproduction. The parametric-only contrast and corresponding NLL and JS interactions are secondary, supportive measurements.

### B.3 Real-source safety–utility metrics

For a real-world record, the *full* condition displays its authentic page title, publisher role, timestamp, permanent URL, and topical scope; the *stripped* condition preserves the claim but replaces source-identity fields with metadata-unavailable placeholders and calibrated neutral padding. For records containing erroneous claims, we retain the controlled risk direction:

$$\Delta_{\mathrm{err}} = \mathbb{E}[Y_{\mathrm{false}}(\mathrm{stripped}) - Y_{\mathrm{false}}(\mathrm{full})]. \tag{12}$$

For correct first-party updates, utility is

$$\Delta_{\mathrm{upd}} = \mathbb{E}[Y_{\mathrm{correct}}(\mathrm{full}) - Y_{\mathrm{correct}}(\mathrm{stripped})]. \tag{13}$$

Positive values denote harm in Equation 12 but benefit in Equation 13. We also report analogues of parametric answer retention (PAR+) and source-driven revision (SDR+). Unlike prior bare-question selection protocols, our runs use pairwise closed-book likelihood margins. Full-set values are therefore the main analogous measures; restrictions to $M_{im} > 0$ for PAR+ and $M_{im} < 0$ for SDR+ are proxies only. Each response is separately coded for correctness against the audited answer (factuality) and agreement with the supplied record (source faithfulness), with U and parser failures retained as distinct outcomes.

### B.4 Deployment metrics

Let $u = (i, m, z, r)$ index a candidate–model request and $Y_u$ denote its false adoption outcome. A verification budget $b$ allows a fraction $b$ of requests to be reviewed. Let $V_u^\pi(b) \in [0, 1]$ be the probability that policy $\pi$ routes request $u$ for verification. This value is normally binary; a fractional value represents uniform randomization among exact score ties at the cutoff. Under an ideal-verifier accounting assumption, every routed false response is detected and corrected. The unconditional residual risk is

$$\mathcal{R}_\pi(b) = \mathbb{E}_{w,u}\left[Y_u\left(1 - V_u^\pi(b)\right)\right]. \tag{14}$$

Let $\mathcal{R}_0 = \mathbb{E}_{w,u}[Y_u]$ be risk without verification and $G_\pi(b) = \mathcal{R}_0 - \mathcal{R}_\pi(b)$ the absolute fraction of false adoptions prevented. At fixed $b$, we report automatic coverage $1-b$; selective risk $\mathcal{R}_\pi(b)/(1-b)$ among automatically handled requests; errors prevented per 1,000 queries $1000G_\pi(b)$; review yield $G_\pi(b)/b$; lift over random review $[G_\pi(b)/b]/\mathcal{R}_0$; and number needed to verify $\mathrm{NNV} = b/G_\pi(b)$. The prespecified policy comparison is

$$\Psi(b) = \mathcal{R}_{\textsc{ProvRoute}}(b) - \mathcal{R}_{\textsc{EpiGate-lite}}(b), \qquad b = 1/6, \tag{15}$$

where $\Psi(b) < 0$ favors ProvRoute. These metrics isolate routing quality under idealized successful verification; they do not estimate verifier accuracy, latency, or monetary cost.

Table 7: Model and protocol coverage. Supportive models do not change the confirmatory estimand. LL denotes deterministic label-logit measurement.

| Model | Family / scale | Protocol coverage | Evidence role |
|---|---|---|---|
| Qwen-3-4B-Instruct-2507 | Qwen, 4B | Dev; locked; LL; task; real; gate | Confirmatory core |
| Qwen-3-8B | Qwen, 8B | Dev; locked; LL; task; real; gate | Confirmatory core |
| Granite-3.3-8B-Instruct | Granite, 8B | Dev; locked; LL; task; real; gate | Confirmatory core |
| Mistral-7B-Instruct-v0.3 | Mistral, 7B | Dev; locked; LL; task; real; gate | Confirmatory core |
| Phi-4-mini-instruct | Phi, 3.8B | Two fixed panels; gate; probe cost | Supportive |
| Phi-4 | Phi, 14B | Two fixed panels; gate; probe cost | Supportive scale check |
| Gemma-3-27B-IT | Gemma, 27B | Two fixed panels; gate | Supportive |
| OLMo-2-0325-32B-Instruct | OLMo, 32B | Two fixed panels; gate | Supportive |
| DeepSeek-R1-Distill-Llama-8B | DeepSeek/Llama, 8B | LL on two fixed panels | Supportive metric only |

# C Benchmark Evaluation Protocol Details

## C.1 Response Protocols

The primary evaluation metric is external-answer vote rate under generated, mirrored A/B/U arbitration. Each prompt offers the frozen gold answer, the context-supported false answer, and "U: Cannot determine." For every candidate, model, and provenance condition, two mirrored answer orders are crossed with three prompt variants, yielding six responses. Outputs are mapped to gold, external, or U before aggregation. U, refusals, and parser-invalid responses remain in the denominator. This evaluation metric controls position bias without conditioning on a valid binary choice.

## C.2 Models and Evidence Coverage

The confirmatory population contains four open-weight models in three families, five additional open-weight models provide supportive extensions. Local inference uses bfloat16 weights and each model's native chat template. Qwen-3 8B runs with thinking disabled. Table 7 records the evidence role and protocol coverage.

The confirmatory weights are $1/6$, $1/6$, $1/3$, and $1/3$ for Qwen-3 4B, Qwen-3 8B, Granite, and Mistral, respectively, giving equal family weight to Qwen, Granite, and Mistral.

## C.3 Statistical Protocol

The fact–distractor candidate is the independent unit. Provenance conditions, prompt variants, answer orders, and models are paired repeated measurements. Development analyses report all model–metric combinations, the overall paired effect, signed low-minus-high and absolute near-minus-far contrasts, and 10,000-replicate candidate-bootstrap intervals. They are exploratory and are not filtered by uncorrected significance.

The locked analysis follows a prespecified two-stage sequence. First, one family-balanced external-vote effect is formed for each of the 60 candidates and its mean is tested with a two-sided one-sample Student $t$ test at $\alpha = .05$. Only if this gate passes is the family-balanced near-minus-far contrast tested with a two-sided candidate-cluster sandwich $t$ test under the frozen 20/20/20 strata. The resulting $t$ intervals are the primary 95% confidence intervals; a 10,000-replicate candidate-cluster bootstrap with seed 2711 is a sensitivity interval. The fixed sequence controls familywise Type-I error at .05. The frozen headline rule also requires both primary point estimates to be at least five percentage points and positive model-specific near-minus-far estimates in both Qwen sizes and at least one non-Qwen family.

Parser validity must exceed 95% for every confirmatory model. All fixed responses remain in the denominator; there is no outcome-dependent case or model exclusion, adaptive sample-size increase, or behavioral rerun after a usable locked response. Supportive fixed-panel extensions do not modify the original hypothesis family.

Table 8: Factuality by source-faithfulness composition over all mirrored real-source A/B/U responses.

| Provenance | Factual/faithful | Factual/unfaithful | Nonfactual/faithful | Nonfactual/unfaithful | U | Invalid |
|---|---|---|---|---|---|---|
| Full / visible | 29.26% | 25.65% | 22.13% | 1.11% | 21.67% | 0.19% |
| Stripped / absent | 19.35% | 20.74% | 19.44% | 0.37% | 39.72% | 0.37% |

# D Factuality and Source Faithfulness

To determine whether provenance visibility improves factual accuracy or primarily encourages adherence to the supplied source, we classify each substantive response along two orthogonal dimensions. Factuality indicates whether the response agrees with the independently audited correct answer, whereas source faithfulness indicates whether it follows the answer supported by the supplied record. $U$ responses and unclassified or parser-failure outputs are retained as separate outcomes. Distinguishing these dimensions is essential because their relationship differs across the two lanes: in Lane A, resisting an erroneous record is factual but source-unfaithful, whereas in Lane B, adopting a correct official update is both factual and source-faithful.

As shown in Table 8 , full/visible provenance increases factual-and-faithful responses from 19.35% to 29.26% ($+9.91\%$) and factual-but-unfaithful responses from 20.74% to 25.65% ($+4.91\%$), while reducing $U$ responses from 39.72% to 21.67% ($-18.05\%$). Aggregating across the two factuality cells, the proportion of factually correct responses rises from 40.09% to 54.91%, aggregating across the two faithfulness cells, source-faithful responses increase from 38.79% to 51.39%. These results indicate that the primary behavioral effect of visible provenance is to reduce abstention and elicit more substantive responses, accompanied by increases in both factual accuracy and source faithfulness in the pooled data. The safety implications, however, depend on the lane. In Lane A, following the supplied source can produce factually incorrect answers; in Lane B, the same source-following behavior facilitates the adoption of correct updates. Factuality and source faithfulness must therefore be reported separately and interpreted in light of the underlying correctness of the source.

# E Additional ProvRoute Evaluation Details

## E.1 Deployment Setting

Let $u = (i, m, z, r)$ index a request and $Y_u$ indicate false adoption of the context-supported answer in the controlled stable-false track. The objective is to minimize $\mathcal{R}_\pi(b)$ in Equation 14 at fixed $b$ using only signals available before release. A policy may use which candidate the context supports, whether provenance is missing or explicitly weak, the model-specific pre-treatment margin, and probe responses. It may not use the audited truth label or the held-out evaluation response. The primary budget is $b = 1/6$: 16.67% verification and 83.33% automatic coverage. We use the ideal-verifier accounting assumption to isolate routing quality from downstream verifier accuracy.

## E.2 Equal-Budget Baselines

**Random** review assigns every request the same routing probability. **Margin-only** ranks by within-model boundary closeness, placing smaller absolute margins first. **EpiGate-lite** combines normalized boundary closeness $B_u \in [0, 1]$, provenance absence $A_u \in \{0, 1\}$, and their interaction:

$$r_{\text{lite}}(u) = \sigma(-3.057 + 3.381B_u + 0.293A_u + 0.271B_uA_u)\,. \tag{16}$$

These policies successively add no instance information, model-specific boundary proximity, and static margin–provenance features. We exclude a global raw-margin threshold because Sections 4 and 5 show that margin scales and even near–far orderings are model specific. The central comparison asks whether a response-conditioned cross-check improves on the strongest pre-decision baseline at the same budget.

## E.3 Outcome-Unseen Evaluation Protocol

Before any outcome on the final panel was generated, we froze Equation 7, the two probes, the $1/6$ budget, the primary comparison $\Psi(1/6)$, a two-point substantive threshold, bootstrap procedure,

Table 9: Supportive model and probe-cost heterogeneity at $b = 1/6$. Differences are residual-risk percentage points with bootstrap 95% intervals.

| Model | PROVROUTE − lite | PROVROUTE − margin-only | Two probes − one-probe average |
|---|---|---|---|
| Phi-4-mini, 3.8B | $+0.14$ $[-0.07, +0.42]$ | $-0.14$ $[-0.56, +0.21]$ | $0.00$ $[0.00, +0.07]$ |
| Phi-4, 14B | $\mathbf{-2.15}$ $[\mathbf{-5.35}, \mathbf{-0.21}]$ | $\mathbf{-2.57}$ $[\mathbf{-5.76}, \mathbf{-0.62}]$ | $-0.38$ $[-1.25, 0.00]$ |
| Gemma-3, 27B | $+0.07$ $[-0.42, +0.42]$ | $-0.07$ $[-1.46, +0.14]$ | $-0.03$ $[-0.24, 0.00]$ |
| OLMo-2, 32B | $+0.07$ $[-0.28, +0.42]$ | $-0.35$ $[-0.63, +0.28]$ | $-0.03$ $[-0.14, 0.00]$ |

Table 10: Deployment accounting per 1,000 requests at $b = 1/6$.

| Policy | Reviews | Errors prevented | Residual errors | Yield | Lift | NNV |
|---|---|---|---|---|---|---|
| Random | 166.7 | 67.2 | 335.8 | 0.403 | $1.00\times$ | 2.48 |
| Margin-only | 166.7 | 105.3 | 297.7 | 0.632 | $1.57\times$ | 1.58 |
| EpiGate-lite | 166.7 | 106.7 | 296.3 | 0.640 | $1.59\times$ | 1.56 |
| PROVROUTE | **166.7** | **127.4** | **275.6** | **0.765** | $\mathbf{1.90\times}$ | **1.31** |

parser, and cross-family direction gate. We then constructed 60 new facts using the PROVMARGIN protocol. This outcome-unseen panel spans 12 domains, is balanced across the four original models' closed-book strata, and has no fact overlap with earlier behavioral or routing data.

Three-fold prompt cross-fitting prevents a response from evaluating the route it created. In each fold, the two mirrored responses from one prompt variant produce $P_u$ and $U_u$; only the four responses from the other two variants measure residual false adoption. Rotation creates three disjoint probe–evaluation pairings. Policies rank separately within model and fold; exact cutoff ties use uniform expected routing, meeting the budget exactly without candidate identity or outcome as a tie-breaker.

### E.4 MODEL AND PROBE-COST HETEROGENEITY

We next test transfer to four supportive models and whether the second mirrored probe justifies its extra inference cost. Table 9 uses the same fixed panel and budget; negative values favor PROVROUTE or the full two-probe rule.

Phi-4 14B is the only supportive model whose PROVROUTE-minus-lite interval lies fully below zero. Phi-4-mini, Gemma, and OLMo show no reliable incremental gain. The adoption-conditioned component is useful for some models, not a universal add-on, consistent with model-specific localization in Section 4.

The probe-cost result is clearer. A one-probe variant requires one additional decision call, whereas the mirrored rule requires two, but their risk differences are close to zero on all four supportive models. The second probe is useful for controlling answer-order effects in evaluation, yet the current data do not show a general deployment benefit sufficient to justify doubling probe calls. One probe is therefore a reasonable low-cost candidate unless target model validation finds a material mirrored-probe gain.

### E.5 DEPLOYMENT ACCOUNTING AND UTILITY GUARDRAILS

Table 10 translates the confirmed reduction into operational quantities at the observed base rate and frozen budget. All values retain the ideal-verifier assumption.

PROVROUTE sends 166.7 of every 1,000 requests for verification and is expected to prevent 127.4 false adoptions, for a review yield of 0.765, a $1.90\times$ lift over random review, and NNV 1.31. The 33.07% selective risk among automatically handled requests remains substantial: the gate prioritizes risk but does not make the remainder safe. These quantities depend on the 40.30% base rate, the $1/6$ budget, the controlled false-adoption definition, and successful verification. They are not estimates of verifier accuracy, latency, or monetary cost in a deployed RAG or agent system.

The real-source study adds a development-only utility guardrail. On 18 erroneous records, PROVROUTE reduces residual false-adoption risk from 27.08% under lite to 20.41%, a descrip-

tive 6.67-point improvement. On 12 official updates, automatic correct-update coverage is 50.93% for PROVROUTE versus 51.04% for lite, and the perfect-source-check resolved-correct rate is 67.59% versus 67.71%; both differences are about $-0.12$ points. PROVROUTE routes 0% of visible-official requests, compared with 8.33% for lite, showing that the source-risk mask avoids specifically targeting visible first-party updates, but not establishing improved update utility. These small development panels cannot substitute for the controlled outcome-unseen confirmation or an end-to-end verifier evaluation.

# F DISCUSSION

## F.1 PROVENANCE AS EVIDENCE FOR TRUST ALLOCATION

Our results show that identical claims can elicit different responses when their provenance changes. This suggests that evaluating knowledge conflicts requires attention to both factual content and the information accompanying it. In practical systems, preserving a claim while dropping its provenance may change how the model weighs that claim. PROVMARGIN isolates this behavioral effect for a specific package of credibility, authority, scope, and lineage cues. Because these fields change together, the results do not identify their individual contributions or the model's internal mechanism. Separating these cues and tracing their loss through retrieval, summarization, and memory reuse would clarify which information systems should preserve.

## F.2 RISK LOCALIZATION REQUIRES MODEL-SPECIFIC VALIDATION

The stronger near-boundary effect in the confirmatory models makes the closed-book margin useful for studying provenance sensitivity. However, the reversed pattern in OLMo shows that this relationship cannot be assumed across models. The margin measures relative support for two candidates; a small value can also arise when both receive little support. It therefore needs validation for each model, candidate representation, and task. Moreover, sensitivity to a provenance intervention and baseline error risk are distinct: the instances whose answers change most need not be those most likely to be wrong. This distinction motivates evaluating risk localization and verification performance separately, using response measures appropriate to the intended application.

## F.3 PRESERVING PROVENANCE WHILE BALANCING SAFETY AND UTILITY

The benefit of official provenance for correct updates highlights a cost of indiscriminately rejecting external information: useful knowledge updates may also be lost. Conversely, authentic provenance does not consistently prevent adoption of erroneous claims in our real-world panel. Systems should therefore retain provenance as evidence to assess, while checking whether its authority, scope, and recency are relevant to the particular claim. Evaluation should measure factual accuracy, claim faithfulness, abstention, and correct-update coverage separately. Lower false adoption alone can conceal an increase in unanswered questions. The current 30-record panel provides an initial utility check; broader panels with independent human review are needed to characterize this tradeoff under realistic mixtures of correct and erroneous information.

## F.4 FROM VERIFICATION PRIORITY TO DEPLOYMENT BENEFIT

PROVROUTE shows that observable responses can improve verification allocation beyond static margin and provenance features in the primary evaluation. Its gains remain model dependent, and the ideal-verifier assumption isolates routing quality from verification success. The remaining automatic responses still have a false-adoption rate of 33.07%, so deployment must jointly consider review budget, residual risk, and verifier accuracy. Additional model calls also incur cost; the small gains from a second mirrored call in the supportive models motivate testing whether one call is sufficient. Open-ended use further requires candidate construction, extraction of the context-supported answer, and reliable provenance assessment. These upstream steps, including resilience to forged provenance, require end-to-end evaluation.