PHANTOM MERGE: WHEN YOUR LARGE LANGUAGE MODEL AGENTS PICK ONE BUT TELL YOU ABOUT ANOTHER

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents increasingly operate over multiple entities, where reliable response generation requires not only retrieving relevant evidence but also preserving its association with the correct entity. We identify *Phantom Merge* (PM), an entity–evidence binding failure in which an agent attributes unsupported claims to its committed entity, potentially even when the task is completed correctly. We formalize PM at the claim level and characterize its prevalence and error subtypes across four agent domains and four backbone families. To address this problem, we propose a two-stage framework for claim-level risk estimation and correction. First, Anchor-Grounded Risk (AGR) combines a representational readout of the agent's pre-emission hidden state with Jacobian-based functional information, providing complementary slot-level and anchor-value risk scores. Second, Fixed-Anchor Correction (FAC) audits flagged claims and selectively revises them using observed anchor evidence or removes them when no supported replacement is available. On the held-out Shopping detection cohort, AGR-slot achieves an AUROC of 96.2% and an average precision of 97.3%, outperforming the evaluated detection baselines. In the end-to-end Shopping mitigation evaluation, FAC reduces trajectory-level PM from 51.4% to 6.2%, with an output-to-original claim ratio of 80.7%. These results demonstrate the prevalence of entity-specific grounding failures and the effectiveness of combining internal risk estimation with evidence-constrained claim correction. Our code is publicly available at `https://github.com/PhantomMerge00/PhantomMerge`

1 INTRODUCTION

LLM agents extend language models from single-turn generation to interactive systems that interleave reasoning, tool calls, and observations from external environments (Yao et al., 2023; Nakano et al., 2021). This capability allows agents to search, compare, and act over multiple external entities before producing a final response (Lee et al., 2025). In such settings, reliable response generation requires each final-answer claim to be supported by observations associated with the entity that the claim describes. We refer to the entity to which the agent commits in its final response as the *anchor*. However, during final-answer synthesis, an agent may attribute to the anchor a value observed for another entity, project a requested constraint onto the anchor as an observed attribute, or generate a value that conflicts with the anchor's observed evidence. We define this class of entity–evidence binding failures as *Phantom Merge* (PM), where an anchor-attributed claim lacks support from the anchor's own observed evidence. PM differs from generic hallucination because problematic content can already be supported elsewhere in the trajectory while remaining unsupported for the anchor. Accordingly, task-level success and trajectory-level factual support provide only partial information about whether the final response is correctly grounded in the committed entity.

However, existing approaches leave two limitations in establishing and restoring entity-specific grounding. *Gap 1. Entity-specific support verification.* Current agent evaluations (Liu et al., 2023) primarily assess task completion or entity selection, while factuality methods examine whether generated claims are supported by available evidence (Rashkin et al., 2023). These approaches do not explicitly verify whether the supporting evidence belongs to the entity referenced by each final-answer claim. Consequently, the first challenge is to assess claim-level grounding across evidence

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.

from multiple entities collected during successive agent interactions. *Gap 2. Entity-specific claim correction.* Existing post-hoc factuality methods revise or remove unsupported content (Dhuliawala et al., 2024), but generally do not explicitly constrain corrections to evidence associated with the entity already selected by the agent. Therefore, the second challenge is to correct incorrectly attributed claims by entity-specific evidence while preserving committed entity and correctly grounded content.

To address these challenges, we propose a two-stage framework for diagnosing and correcting Phantom Merge. For the first challenge, we introduce *Anchor-Grounded Risk* (AGR), which combines two complementary readouts of the agent's pre-emission hidden state. Specifically, a representational readout estimates claim-level binding risk, while a Jacobian-based functional readout evaluates downstream compatibility with anchor-derived targets at attribute or value granularity. We develop a fixed-additive slot-level score for primary detection and a calibrated anchor-value configuration for complementary risk estimation. For the second challenge, we introduce *Fixed-Anchor Correction*, which combines claim selection, functional auditing, and evidence-grounded revision to correct claims associated with the committed entity. On completed agent trajectories, high-risk claims are revised when a supported replacement is recoverable and removed otherwise, while unaffected claims and the committed entity remain unchanged without updating the underlying agent.

Contributions. Our contributions are summarized as follows:

- **A novel agent reliability problem.** We formalize Phantom Merge as an entity–evidence binding failure and systematically characterize its prevalence, subtype composition, and relationship with task correctness across agent domains and backbones.
- **An anchor-grounded diagnosis method.** We introduce AGR, combining representational and Jacobian-based functional readouts for claim-level PM detection, and demonstrate strong discrimination on Shopping, including test claims with non-overlapping quote templates.
- **An evidence-grounded correction method.** We develop Fixed-Anchor Correction and evaluate its effectiveness in reducing PM and preserving response content through end-to-end and shared-gate experiments across multiple agent domains.

2 Problem Definition

We consider an LLM-agent trajectory $\tau = (q, x_{1:T}, y)$, comprising a user query q, interactions and observations $x_{1:T}$, and final response y. Let V_τ denote the observed entities and $\mathrm{Obs}_\tau(e)$ the evidence observed for entity e. The agent commits to an *anchor* $a_\tau \in V_\tau$, and its response contains anchor-attributed claims $C_\tau = \{c_1, \ldots, c_n\}$, each asserting a value $v(c)$ for attribute $\mathrm{slot}(c)$. We write $\mathrm{Obs}_\tau(e) \models c$ when the evidence supports c, and use $\mathrm{Src}_\tau(c) \subseteq V_\tau \cup \{q\}$ for the identifiable sources of its asserted value, including user constraints from q.

Definition 1 (Phantom Merge)**.** A claim $c \in C_\tau$ exhibits *Phantom Merge* (PM) when it lacks anchor-specific support and satisfies at least one of three conditions: CEM (Cross-Entity Misattribution), where a non-anchor entity supports c; CAP (Constraint-to-Anchor Projection), where a user constraint is asserted as an unsupported anchor attribute; or ASF (Anchor-Slot Fabrication), where neither preceding condition applies and the claim contradicts an observed anchor value. CEM and CAP may co-occur. Let $\mathcal{L}_\tau(c)$ denote the applicable labels and $\mathcal{M} = \{\mathrm{CEM}, \mathrm{CAP}, \mathrm{ASF}\}$. We define

$$\mathrm{PM}(c) = \mathbb{1}[\mathcal{L}_\tau(c) \cap \mathcal{M} \neq \varnothing]\,. \tag{1}$$

A trajectory exhibits PM if any $c \in C_\tau$ satisfies $\mathrm{PM}(c) = 1$.

Definition 2 (Pure Hallucination)**.** A claim is a *Pure Hallucination* (PH) when $\mathrm{Src}_\tau(c) = \varnothing$ and it does not contradict any observed anchor value for $\mathrm{slot}(c)$. PH is excluded from PM.

3 Methodology

3.1 Overview

We propose a two-stage framework for diagnosing and correcting Phantom Merge, as illustrated in Figure 1. Given an agent trajectory τ, its committed anchor a_τ, and anchor-attributed claims

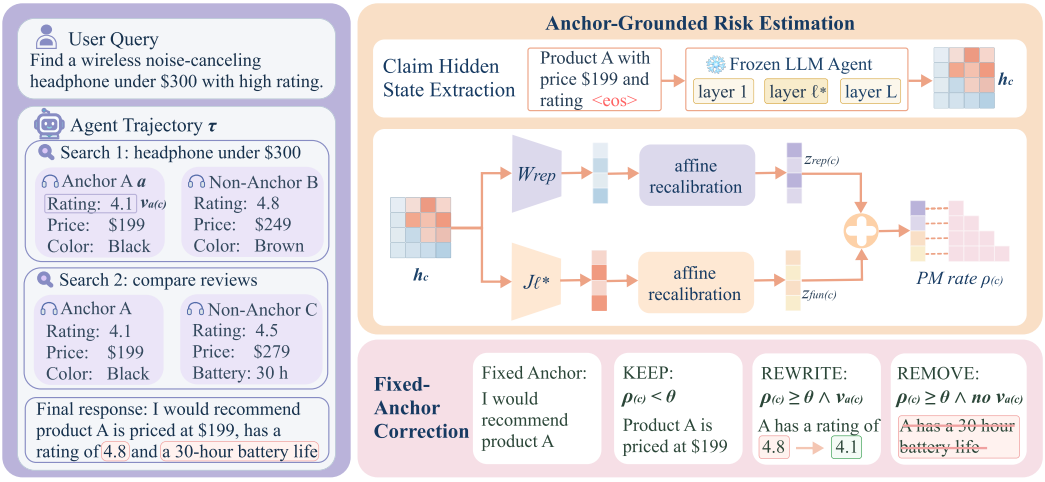


Figure 1: **Overview of the proposed framework.** Anchor-Grounded Risk (AGR) combines representational and Jacobian-based functional information for claim-level PM detection. Fixed-Anchor Correction (FAC) integrates risk-based selection, functional auditing, and anchor-specific evidence to retain, revise, or remove individual claims.

C_τ, the first stage introduces *Anchor-Grounded Risk* (AGR), which integrates representational and Jacobian-based functional information from the agent's pre-emission state. We develop two complementary scoring configurations using attribute-level and anchor-value functional targets. The second stage introduces *Fixed-Anchor Correction* (FAC), which combines claim-level risk assessment with functional auditing and observed anchor evidence to determine appropriate corrections.

3.2 Anchor-Grounded Risk Estimation

We estimate claim-level PM risk from the agent's internal state during final-answer generation. Let t_c denote the position immediately preceding the first token of $v(c)$, and let $h_c \in \mathbb{R}^d$ be the corresponding hidden state at layer ℓ^*, selected on validation data. We extract two complementary signals from this frozen representation: a supervised readout of PM-related information and a functional readout of the model's downstream computation.

Representational Readout. Given the claim-level label $m_c = \mathrm{PM}(c) \in \{0, 1\}$, we train a linear readout $z_{\mathrm{rep}}(c) = w_{\mathrm{rep}}^\top h_c + b_{\mathrm{rep}}$, with $p_{\mathrm{rep}}(c) = \sigma(z_{\mathrm{rep}}(c))$, using binary cross-entropy on labeled anchor–claim instances. The restricted linear formulation measures the accessibility of PM-related information in the frozen representation (Hewitt & Liang, 2019; Belinkov, 2022). The resulting representational signal provides the supervised component of AGR. We denote the output of the frozen representational probe used in the slot-level configuration by $p_L(c)$.

Jacobian-Based Functional Readout. The functional readout estimates the information accessible through the model's downstream computation. We construct two functional targets: an attribute lexicon $T_{\mathrm{slot}}(c)$ associated with $\mathrm{slot}(c)$ and an anchor-value target $T_{\mathrm{val}}(c)$ constructed from $\mathrm{Obs}_\tau(a_\tau)$. The former captures slot-level information, while the latter evaluates compatibility with the observed value of the committed anchor.

Let $F_{\ell^* \to L}$ denote the frozen computation from layer ℓ^* to the final residual representation at position t_c. We abbreviate this mapping as F and approximate its output using an averaged Jacobian estimated from unlabeled generation states. With reference state $\bar{h}$, the resulting approximation is

$$\begin{aligned} \widehat{h}_c^{\,L} &= F(\bar{h}) + J_{\ell^*}(h_c - \bar{h}), \\ J_{\ell^*} &= \mathbb{E}_h[\nabla_h F(h)]\,. \end{aligned} \tag{2}$$

Following the Jacobian-lens approach (Gurnee et al., 2026), we obtain the approximated output logits $\widehat{g}_c = W_U \widehat{h}_c^{\,L}$ using the unembedding matrix W_U. Let $\pi_J(T \mid h_c)$ denote the resulting probability mass assigned to token set T. This provides a shared functional representation from which we construct the two AGR scoring configurations.

Our primary configuration, *AGR-slot*, combines the frozen representational probe with the Jacobian-based slot-level signal. Therefore, we define

$$B(c) = \mathrm{logit}(p_L(c)) + \log \pi_J(T_{\mathrm{slot}}(c) \mid h_c). \tag{3}$$

The score multiplies probe-derived PM odds by slot-level probability mass in log space, downweighting claims with low slot compatibility. This fixed fusion requires no additional weights and ranks claim-level PM risk. Slot mass modulates the probe evidence rather than indicating PM on its own.

Our complementary configuration, *AGR-value*, uses the anchor-value target to estimate functional compatibility with the committed entity's observed evidence. We transform the corresponding probability mass into a risk-oriented score:

$$z^{\mathrm{val}}_{\mathrm{fun}}(c) = \log \frac{1 - \pi_J(T_{\mathrm{val}}(c) \mid h_c)}{\pi_J(T_{\mathrm{val}}(c) \mid h_c)}. \tag{4}$$

This formulation assigns higher functional risk to lower estimated compatibility with the observed anchor value. The value-level signal is computed for claims with recoverable anchor-value targets and subsequently integrated with the representational readout.

Calibrated Risk Fusion. AGR-value integrates the representational and anchor-value functional signals through calibrated fusion. We first apply affine calibration $\widetilde{z}_i = a_i z_i + b_i$ to each available signal on a held-out split $\mathcal{D}_c$, where $i \in \{\mathrm{rep}, \mathrm{fun}\}$. For claims with both signals available, let $s(c) = [\widetilde{z}_{\mathrm{rep}}(c), \widetilde{z}^{\mathrm{val}}_{\mathrm{fun}}(c)]^\top$. Assuming class-conditional Gaussian scores $s \mid m_c = k \sim \mathcal{N}(\mu_k, \Sigma)$ with shared covariance Σ, the corresponding Bayes log-posterior odds are

$$\log \frac{P(m_c = 1 \mid s)}{P(m_c = 0 \mid s)} = (\mu_1 - \mu_0)^\top \Sigma^{-1} s + b. \tag{5}$$

This formulation motivates an affine combination of the two calibrated signals that incorporates their class separation and covariance. Under diagonal covariance, each signal's coefficient is proportional to its class-conditional mean difference divided by its variance. We define the AGR-value score as

$$\rho(c) = \begin{cases} \sigma\big(w^\top s(c) + b\big), & T_{\mathrm{val}}(c) \text{ resolves}, \\ \sigma(\widetilde{z}_{\mathrm{rep}}(c)), & \text{otherwise}. \end{cases} \tag{6}$$

The fusion parameters (w, b) are fitted on a held-out split $\mathcal{D}_f$. Probe fitting, calibration, and fusion use disjoint data partitions, while the Jacobian operator is estimated independently from unlabeled states. The value-level configuration uses the representational signal when an anchor-value target is unavailable. Together, AGR-slot and AGR-value provide two implementations of representational–functional risk estimation, combining their respective targets through fixed additive scoring and calibrated fusion for complementary claim-level PM detection.

3.3 Fixed-Anchor Claim Correction

We introduce *Fixed-Anchor Correction* (FAC) to revise risky claims using evidence associated with the committed entity. FAC consists of two components: a claim-selection gate and an internal functional audit. The gate identifies claims requiring intervention, while the audit provides additional information for selecting an appropriate correction. For each claim c, we retrieve the value of $\mathrm{slot}(c)$ from $\mathrm{Obs}_\tau(a_\tau)$ and denote the recovered anchor-supported value by $v_a(c)$ when available.

Anchor-Constrained Routing. Let $g(c)$ denote the score supplied by the claim-selection gate, and let u_c denote the functional audit state. The audit combines representational risk, slot-level functional information, and attribute-alignment features to distinguish different intervention conditions. Given threshold θ, the resulting claim-level operator is

$$\mathcal{A}_\theta(c) = \begin{cases} c, & g(c) \le \theta, \\ \mathrm{Route}(c, u_c, \mathrm{Obs}_\tau(a_\tau)), & g(c) > \theta. \end{cases} \tag{7}$$

The routing procedure uses the functional audit to assess flagged claims and determine whether to retain, revise, or remove them. Claims eligible for revision enter an evidence-grounded rewriting procedure, which uses a recovered anchor-supported value to correct the asserted attribute while preserving the remaining claim content. Claims requiring removal are excluded from the corrected re-

sponse. Applying the operator to the final-answer claims produces y' while preserving the committed anchor and the completed interaction trajectory.

FAC supports alternative risk estimators through its claim-selection interface. In the Shopping shared-gate comparison, all correction methods receive the same claims selected by the frozen representational probe, $p_{\mathrm{pm}}(c) > 0.05$. FAC then applies functional auditing and anchor-constrained correction to these claims. We separately examine alternative AGR detection heads under fixed downstream routing to evaluate how risk estimation influences intervention outcomes.

4 Experiment Settings

Benchmark. We evaluate Phantom Merge on four multi-entity agent domains: ShoppingBench (Wang et al., 2025) and the airline, telecom, and retail domains of τ^3-Bench (Yao et al., 2024; Barres et al., 2025). PM characterization uses 2,000 PM-enriched Shopping trajectories and 1,000 trajectories from each τ^3-Bench domain. Detection and mitigation additionally use an independently annotated Shopping cohort, with a held-out test set of 243 trajectories and 274 claims (156 PM and 118 correctly bound). The three τ^3-Bench domains also serve as cross-domain evaluation sets.

Baselines. For detection, we compare AGR with TF-IDF, two ModernBERT classifiers (Warner et al., 2025), SelfCheckGPT (Manakul et al., 2023), FacTool (Chern et al., 2023), and FActScore (Min et al., 2023), alongside individual readouts and AGR-value. Additional anchor-aware baselines include DeBERTa-NLI (He et al., 2021), LLM verifiers, and cite-or-drop. For mitigation, we compare Fixed-Anchor Correction with Fixed-Anchor Deletion, RARR (Gao et al., 2023a), CoVe (Dhuliawala et al., 2024), and a FEVER-style baseline (Thorne et al., 2018), under both complete intervention and shared-gate settings. Implementation details are provided in Appendix E.

Metrics. For PM characterization, we report trajectory-level prevalence, task correctness, conditional statistics, and claim-level subtype composition. Detection metrics include AUROC, average precision (AP), precision, recall, F1, and recall-matched precision. Mitigation metrics include trajectory-level PM rate, correctly bound claim retention, and the output-to-original claim ratio to quantify mitigation and content preservation. Post-intervention outputs are re-evaluated using the same AGR procedure, with manual spot checks of the resulting assessments.

Backbones. Qwen3-32B (Yang et al., 2025) is our primary agent backbone for characterization, detection, and mitigation. We additionally evaluate Gemma-4-31B (Team et al., 2026), Llama-3.3-70B (Grattafiori et al., 2024), and Mixtral-8x7B (Jiang et al., 2024) for PM characterization. Qwen3-8B performs deterministic claim rewriting. All agent backbones remain frozen.

5 Experimental Evaluations

We evaluate Phantom Merge and the proposed framework through three research questions. **RQ1:** How does Phantom Merge manifest across agent domains and model backbones? **RQ2:** How effectively can AGR detect Phantom Merge, and what contributes to its performance? **RQ3:** How effectively can Fixed-Anchor Correction reduce Phantom Merge while preserving response content?

5.1 Prevalence and Characteristics of Phantom Merge

We examine Phantom Merge across four agent domains and four model backbones, measuring its trajectory-level prevalence, occurrence among task-correct trajectories, and claim-level subtype composition. Table 1 summarizes the results.

Prevalence and task correctness. PM occurs in all 16 evaluated backbone-domain combinations, spanning dense and mixture-of-experts models and both recommendation and customer-service tasks. Its prevalence ranges from 0.4% for Gemma on Retail to 54.1% for Qwen on Telecom, indicating substantial variation across configurations. More importantly, PM is not confined to unsuccessful trajectories. In the PM-enriched Shopping cohort, 29.6–55.6% of task-correct trajectories still exhibit PM across the four backbones. This discrepancy also appears outside Shopping: PM affects 32.0% of task-correct trajectories for Mixtral on Airline and 17.1% for Qwen on Telecom, although it is

Table 1: Phantom Merge across four agent domains and backbone families. TC denotes task correctness, PM denotes trajectory-level Phantom Merge, and $X \mid Y$ denotes $P(X \mid Y)$. CEM, CAP, and ASF report claim-level PM subtype composition. All values are percentages.

Domain	Backbone	TC	PM	PM|TC	TC|PM	CEM	CAP	ASF
Retail	Mixtral	8.7	16.5	18.4	9.7	4.2	1.6	94.2
	Gemma	21.8	0.4	0.5	25.0	0.0	0.0	100.0
	Llama	9.2	2.5	0.0	0.0	3.6	7.1	89.3
	Qwen	15.9	3.1	1.3	6.5	69.0	0.0	31.0
Airline	Mixtral	35.6	34.0	32.0	33.5	0.0	0.3	99.7
	Gemma	29.3	4.5	4.1	26.7	0.0	75.0	25.0
	Llama	16.8	5.1	2.4	7.8	0.0	7.7	92.3
	Qwen	13.7	18.6	5.1	3.8	26.4	30.6	43.1
Telecom	Mixtral	4.0	11.7	7.7	2.7	0.0	0.0	100.0
	Gemma	11.6	0.6	0.9	16.7	62.5	0.0	37.5
	Llama	3.6	27.4	5.6	0.7	100.0	0.0	0.0
	Qwen	4.1	54.1	17.1	1.3	100.0	0.0	0.0
Shopping	Mixtral	41.0	33.0	37.1	46.0	10.9	85.2	3.9
	Gemma	27.7	41.7	45.5	30.3	2.6	95.9	1.5
	Llama	35.1	26.8	29.6	38.8	1.6	96.4	2.0
	Qwen	20.7	46.8	55.6	24.6	29.5	63.3	7.2

absent among task-correct trajectories for Llama on Retail. These findings show that successful task completion can coexist with incorrectly attributed claims. Conventional task correctness therefore cannot fully capture the PM of final responses, motivating explicit claim-level assessment.

Error composition. PM subtype composition varies markedly across both domains and backbones. Shopping is dominated by Constraint-to-Anchor Projection (63.3–96.4%), whereas Telecom exhibits particularly pronounced backbone-dependent differences: all observed PM claims are classified as Cross-Entity Misattribution for Qwen and Llama, but as Anchor-Slot Fabrication for Mixtral. These extremes occur at substantially different PM prevalence rates, including 54.1% for Qwen and 11.7% for Mixtral. Airline also illustrates this heterogeneity: Anchor-Slot Fabrication accounts for 99.7% of Mixtral's PM claims, while Qwen exhibits a more balanced distribution across all three subtypes. Importantly, extreme subtype proportions must be interpreted alongside PM prevalence and the number of observed PM claims. For instance, Gemma's 100% Anchor-Slot Fabrication on Retail accompanies a PM prevalence of only 0.4%, making its subtype estimate particularly sensitive to the limited number of PM instances. These results establish that PM is neither restricted to a single error subtype nor characterized by a uniform error distribution across agent configurations.

5.2 Detection Effectiveness and Readout Analysis

We answer **RQ2** by comparing AGR with lexical, embedding-based, and adapted factual verification methods on the same held-out Shopping cohort. We evaluate overall detection performance and examine how the representational and functional readouts contribute to AGR.

Baseline comparison. Table 2 compares AGR with lexical, embedding-based, and adapted factual verification methods on the same held-out Shopping cohort. AGR-slot achieves the highest detection performance, reaching 96.2% AUROC and 97.3% AUPRC. Compared with DeBERTa-NLI, the strongest external baseline in overall ranking performance, AGR-slot improves AUROC by 13.3 percentage points and AUPRC by 7.9 points. Its advantage extends across the lexical, embedding-based, and LLM verification baselines, including methods supplied with anchor-specific evidence. The representational readout alone also exceeds all evaluated external baselines in

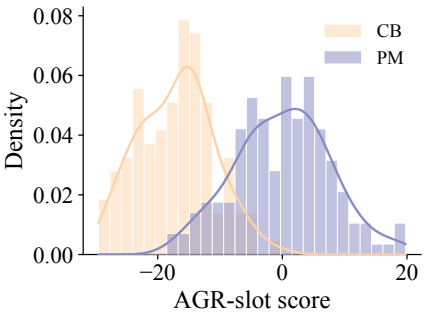


Figure 2: AGR-slot score distribution on the Shopping test cohort.

Table 2: Phantom Merge detection on the Shopping test cohort. AUROC and AUPRC are reported as percentages with 95% bootstrap confidence intervals. Precision, recall, and F1 are evaluated at $\tau = 0.05$ for applicable probability-like scores. "–" denotes scores evaluated separately using development-selected thresholds (Appendix D).

Method	AUROC	AUPRC	Precision	Recall	F1
RARR	49.9 [48.7, 50.8]	56.9 [50.2, 63.0]	50.0	0.6	1.3
FActScore	58.2 [51.6, 64.6]	65.5 [57.5, 73.0]	59.4	62.8	61.1
FacTool	67.9 [61.9, 74.2]	67.9 [60.8, 74.5]	77.6	62.2	69.0
Anchor-only LLM	71.9 [66.4, 77.4]	72.7 [65.9, 78.6]	83.6	59.0	69.2
Source-agnostic LLM	75.9 [70.6, 81.2]	76.6 [69.8, 82.6]	88.3	62.8	73.4
ModernBERT	76.2 [69.1, 82.6]	79.9 [71.6, 86.9]	61.2	96.2	74.8
TF-IDF	77.9 [71.7, 83.3]	78.3 [69.9, 85.9]	56.9	100.0	72.6
Cite-or-drop LLM	78.8 [73.5, 83.8]	77.6 [70.9, 83.6]	83.1	78.8	80.9
SelfCheck-NLI	81.5 [75.2, 87.0]	85.1 [79.2, 90.3]	70.4	88.5	78.4
DeBERTa-NLI	82.9 [77.4, 87.9]	89.4 [85.3, 92.9]	73.6	84.0	78.4
Representational	88.8 [84.6, 92.4]	90.8 [86.0, 94.5]	76.5	91.7	83.4
Jacobian	89.4 [85.1, 93.4]	90.9 [86.9, 94.3]	–	–	–
AGR-value	89.8 [85.7, 93.3]	91.8 [87.2, 95.2]	–	–	–
AGR-slot	**96.2 [94.0, 97.9]**	**97.3 [95.5, 98.6]**	–	–	–

AUROC, indicating that the agent's pre-emission representations contain useful information for identifying PM. Incorporating functional information further improves discrimination, suggesting that downstream computation provides predictive information about the model's compatibility with attribute-slot targets beyond what the linear representational probe captures on its own.

Readout contributions. Table 2 further examines the contributions of the representational and Jacobian-based functional readouts. Individually, the two signals achieve comparable detection performance, with AUROC of 88.8% and 89.4%, respectively, demonstrating that both capture information relevant to PM prediction. Combining them yields additional improvements, most prominently in AGR-slot, which integrates the representational probe with slot-level functional information. Its AUROC increases to 96.2%, accompanied by an AUPRC of 97.3%. AGR-value, which incorporates anchor-value compatibility through calibrated fusion, also achieves competitive detection performance. The gain over both individual readouts suggests that similar standalone accuracy can conceal complementary predictive information.

Claim-level discrimination. Figure 2 provides a complementary view of AGR-slot's detection performance through the score distributions of PM and correctly bound claims. The two groups exhibit distinct concentrations, with PM claims generally receiving higher scores, consistent with the strong ranking performance reported in Table 2. To further examine discrimination at comparable detection coverage, we evaluate the internal readouts at matched recall. At approximately 91.7% recall, AGR-slot achieves 90.5% precision, compared with 77.3% for the representational readout and 77.7% for AGR-value. This improvement corresponds to substantially fewer incorrectly flagged claims at the same PM detection coverage. On the 93 test claims with non-overlapping quote templates, AGR-slot achieves 95.0% AUROC, compared with 88.2% for the representational readout and 75.6% for TF-IDF (Appendix C.1).

Cross-domain analysis. Figure 3 extends the readout comparison to Telecom, Airline, and Retail. The representational readout achieves AUROC values of 0.94, 0.83, and 0.87 across these domains, while AGR-value achieves 0.94, 0.82, and 0.87, respectively. AGR-slot achieves its strongest performance on Shopping. Across the three additional domains, the representational readout outperforms the standalone functional readout, while AGR-value closely matches the representational readout. This comparison identifies the representational signal as the more consistently discriminative component across these evaluated environments.

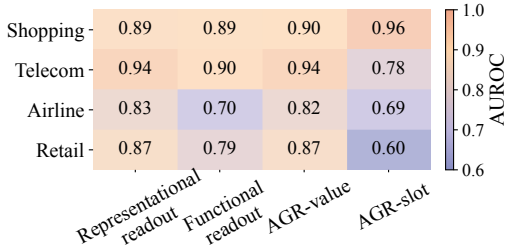


Figure 3: Cross-domain diagnostic AUROC of the evaluated readouts under in-domain and zero-shot configurations.

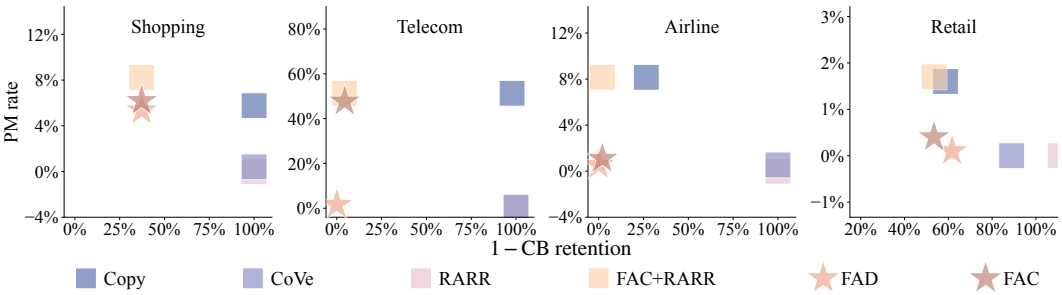


Figure 4: End-to-end mitigation across four agent domains. Lower-left is better: FAC achieves a favorable trade-off between PM reduction and correctly bound claim retention.

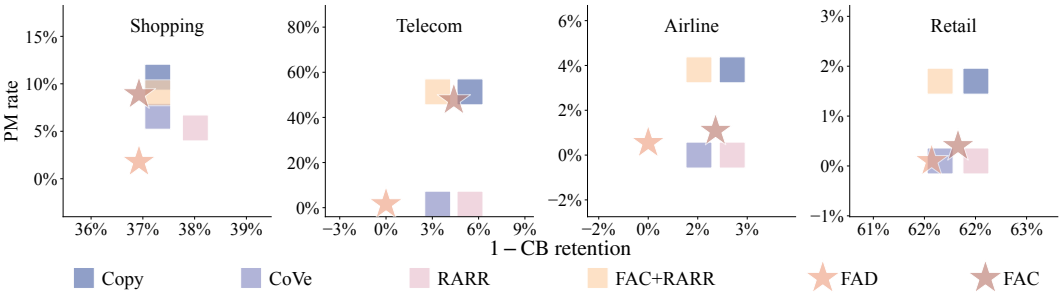


Figure 5: Shared-gate mitigation across four agent domains. All methods receive identical claims flagged by a frozen representational probe, isolating differences in downstream routing and correction. Marker positions are jittered to reduce overlap. Exact residual PM rates are reported in Table G.1

5.3 Mitigation Effectiveness and Claim Preservation

We answer **RQ3** through end-to-end mitigation, shared-gate correction, and detector-head ablation, examining PM reduction and content preservation.

End-to-end effectiveness. Figure 4 compares complete intervention configurations across four domains. On Shopping, FAC substantially reduces PM while retaining correctly bound claims. Deletion-intensive alternatives can achieve lower residual PM rates, but at the cost of removing more valid content. FAC offers a different trade-off by replacing unsupported attributes with recoverable anchor-supported values rather than simply discarding flagged claims. The cross-domain results reveal an important distinction. FAC also achieves substantial PM reduction on Airline and Retail, whereas Telecom favors more aggressive removal: Fixed-Anchor Deletion (FAD) achieves considerably lower residual PM than the revision-based configuration. This difference illustrates the limitations of revision when suitable replacements cannot be recovered. Consequently, mitigation performance depends not only on identifying erroneous claims but also on the evidence available for correction.

Shared-gate correction. To isolate downstream correction from claim selection, all methods receive identical claims flagged by a frozen representational probe. On Shopping, the shared gate is $p_{\mathrm{pm}}(c) > 0.05$. Figure 5 compares their residual PM rates and correctly bound claim retention, while Appendix Table G.1 provides the complementary output-claim ratios. On Shopping, FAC, FAD, and RARR achieve the same residual PM rate of 5.3% with comparable correctly bound claim retention. However, FAC preserves 59.3% of the original output-claim volume, compared with 35.8% for FAD and 37.0% for RARR. This difference demonstrates the benefit of anchor-grounded revision over deletion-oriented alternatives at matched residual PM. Other methods exhibit different trade-offs: CoVe retains more output claims but leaves higher residual PM, while constrained copying retains substantially more content at the cost of additional binding errors. The cross-domain results further establish the importance of available anchor evidence. On Retail, FAC

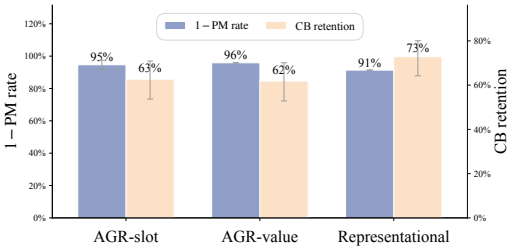


Figure 6: Detector-head comparison on Shopping with downstream routing held fixed. Bars report the PM-free trajectory rate and correctly bound claim retention at the operating points.

retains considerably more output content than FAD with a small increase in residual PM, whereas on Telecom, deletion achieves a much larger PM reduction than revision-based correction. Together, these controlled comparisons show that FAC can preserve additional response content without sacrificing measured PM reduction when supported replacements are available, while the removal branch remains important when anchor evidence is insufficient.

Detector-head ablation. Figure 6 compares upstream risk estimators with downstream routing held fixed. At the evaluated operating points, AGR-value and AGR-slot achieve PM-free trajectory rates of 96% and 95%, with correctly bound claim retention of 62% and 63%, respectively. The representational readout retains 73% of correctly bound claims at a PM-free trajectory rate of 91%. These results illustrate the trade-off between intervention coverage and preservation of correctly bound claims. This pattern is consistent with more aggressive intervention under the AGR configurations, suppressing more PM while also removing more correctly bound claims.

6 Related Work

Evaluation of LLM Agents. Recent agent benchmarks evaluate interactions with web environments (Yao et al., 2022; Zhou et al., 2024; Deng et al., 2023), external tools (Li et al., 2023; Qin et al., 2024; Lu et al., 2025; Chen et al., 2025), and multi-step tasks (Ma et al., 2024; Trivedi et al., 2024). Domain-specific benchmarks further assess agent performance in realistic environments, including product recommendation (Wang et al., 2025). These evaluations primarily characterize task completion, interaction correctness, and tool-use capabilities. Related work on entity binding in tool-augmented agents investigates whether actions target the intended entities (Babu & Indukuri, 2026b;a). Our study examines a complementary failure that can occur after entity selection: whether claims in the final response remain grounded in observations associated with the committed entity.

Factuality Verification and Correction. Existing factuality methods decompose generated responses into verifiable claims (Akbar et al., 2024; Hu et al., 2024; Wei et al., 2024) and assess their consistency with external references, retrieved evidence, or generated verification questions (Fabbri et al., 2022; Laban et al., 2022; Honovich et al., 2022; Wang et al., 2024). Attribution and citation-based approaches additionally connect generated claims to supporting evidence (Gao et al., 2023b; Huang et al., 2024; Zhang et al., 2025), while post-hoc correction methods revise unsupported content through evidence retrieval and verification (Gao et al., 2023a). Although these approaches provide mechanisms for assessing and improving factual support, applying them to multi-entity agent trajectories requires identifying the committed entity and enforcing its evidence boundary. We investigate this setting through claim-level anchor-specific verification and correction, explicitly preserving the original entity commitment.

Entity Binding and Internal Representations. Studies of entity–attribute binding investigate how language models represent associations between entities and their properties (Feng & Steinhardt, 2024; Dai et al., 2024), while related research examines entity tracking and latent-state representations (Kim & Schuster, 2023; Huang et al., 2025). Representation-level analyses and model-editing methods further investigate the accessibility and localization of factual associations (Meng et al., 2022; 2023). More recently, Jacobian-based approaches approximate downstream model computations to interpret information accessible from intermediate hidden states (Gurnee et al., 2026). Building on these directions, we investigate entity-specific binding risk at the claim-generation state of an agent and combine representational and functional readouts for detection. Our setting connects internal representation analysis with evidence-grounded intervention over completed multi-entity trajectories.

7 Conclusion

We identify Phantom Merge (PM), an entity–evidence binding failure in which agents attribute unsupported claims to their committed entity. Evaluations across four domains and four backbones reveal diverse manifestations, including in task-correct trajectories. We introduce Anchor-Grounded Risk (AGR), combining representational and Jacobian-based functional readouts for claim-level detection, and Fixed-Anchor Correction (FAC), which revises or removes flagged claims using anchor evidence. Our experiments demonstrate effective PM detection and substantial reductions in binding errors while retaining response content. These findings establish entity-specific claim grounding as an important dimension of reliable multi-entity agents.

AI USE STATEMENT

We used AI tools for language editing and LLM-assisted annotation as described in Appendix A, with human review of sampled annotations and adjudication of disputed cases. The authors take responsibility for the final manuscript and reported results.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. This work does not involve human subjects or the collection of personally identifiable or sensitive information. We are not aware of any specific ethical concerns arising from this work beyond those generally associated with machine learning research.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, experimental setup, datasets, evaluation protocols, and implementation details in the main paper and the appendix. Additional implementation details are available in the GitHub repository linked in the abstract.

REFERENCES

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15020–15037, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.837. URL https://aclanthology.org/2024.emnlp-main.837/.

Rahul Suresh Babu and Shashank Indukuri. Binding Drift in Multi-Step Tool-Augmented Agents, July 2026a.

Rahul Suresh Babu and Shashank Indukuri. Entity Binding Failures in Tool-Augmented Agents, June 2026b.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. $\tau^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment, June 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Yuefeng Huang, Xiangcheng Liu, Wang Xinzhi, and Wu Liu. ACEBench: A comprehensive evaluation of LLM tool usage. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12970–12998, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.697. URL https://aclanthology.org/2025.findings-emnlp.697/.

I.-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios, July 2023.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967/.

E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3): 837–845, September 1988. ISSN 0006-341X.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.06070.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2587–2601, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.187. URL https://aclanthology.org/2022.naacl-main.187/.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira

Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish

Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable Representations Form a Global Workspace in Language Models, July 2026.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention, October 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287/.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.395. URL https://aclanthology.org/2024.emnlp-main.395/.

Jen-tse Huang, Kaiser Sun, Wenxuan Wang, and Mark Dredze. On the failure of latent state persistence in large language models, 2025. URL https://arxiv.org/abs/2505.10571.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14095–14113, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.838. URL https://aclanthology.org/2024.findings-acl.838/.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024.

Najoung Kim and Sebastian Schuster. Entity tracking in language models, 2023. URL https://arxiv.org/abs/2305.02363v2.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Yugang jia, and Jong Ha Lee. FHIR-AgentBench: Benchmarking LLM agents for realistic interoperable EHR question answering, 2025. URL https://arxiv.org/abs/2509.19319.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.187. URL https://aclanthology.org/2023.emnlp-main.187/.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents, 2023. URL https://arxiv.org/abs/2308.03688v3.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems*, pp. 74325–74362, 2024. doi: 10.52202/079017-2365. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/877b40688e330a0e2a3fc24084208dfa-Abstract-Datasets_and_Benchmarks_Track.html.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Kevin Meng, David Bau, Alex J. Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, 2021. URL https://arxiv.org/abs/2112.09332.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dHng2O0Jjr.

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

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu-hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I.-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 Technical Report, July 2026.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), pp. 809–819, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

Jiangyuan Wang, Kejun Xiao, Qi Sun, Huaipeng Zhao, Tao Luo, Jian Dong Zhang, and Xiaoyi Zeng. ShoppingBench: A real-world intent-grounded shopping benchmark for LLM-based agents, 2025. URL https://arxiv.org/abs/2508.04266v3.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14199–14230, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.830. URL https://aclanthology.org/2024.findings-emnlp.830/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4M9f8VMt2C.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, pp. 20744–20757, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, June 2024.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp.

5098–5122, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.264. URL https://aclanthology.org/2025.findings-acl.264/.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

A Benchmark Construction and Annotation

Evaluation cohorts. We evaluate ShoppingBench and the airline, telecom, and retail domains of τ^3-Bench across four agent backbone families. The primary Shopping cohort contains 2,000 trajectories selected to emphasize multi-entity scenarios, while each τ^3-Bench cohort contains 1,000 trajectories sampled from its task distribution. Shopping prevalence estimates therefore describe the constructed cohort. Detection uses a separately annotated Shopping dataset; its held-out test set contains 274 claims from 243 trajectories, with 156 PM and 118 correctly bound claims.

Annotation protocol. The 2,000 Shopping trajectories were manually annotated according to Section 2. Annotators identified the committed anchor, extracted anchor-attributed factual claims, and compared each claim with the user instruction, anchor evidence, and observations associated with other entities. They assigned the applicable CEM, CAP, ASF, and PH labels. All 274 held-out detection claims received an additional manual review. The Shopping data partition was established without reference to the observed PM label proportions. The detection test set evaluates PM against correctly bound claims; it does not include a separate PH negative class.

For the three τ^3-Bench domains, OpenAI GPT-5.6 Astra performed two independent annotation passes. Each pass received the complete trajectory, user instruction, committed anchor, and evidence associated with all observed entities. Disagreements were reconciled, with human judgments taking precedence when needed. The same annotation criteria were used across domains and backbones.

Human quality control. We audited 50 trajectories from each evaluated model–domain cohort of LLM-assisted annotations, stratifying the sample by PM-positive and PM-negative labels. Cohorts with unusually high PM prevalence received additional review. After reconciliation, no disagreements remained in the final PM labels of the inspected samples. Human review also clarified annotation wording and resolved disputed cases. This describes the final audited labels rather than a pre-adjudication agreement statistic.

B AGR Implementation and Readout Ablation

Data partitions. The Shopping detection dataset contains 1,244 claims: 772 training, 198 development, and 274 test instances. For the AGR fitting protocol, the 970 non-test instances are partitioned into $\mathcal{D}_p$ ($n = 393$) for probe fitting, $\mathcal{D}_c$ ($n = 289$) for score calibration, and $\mathcal{D}_f$ ($n = 288$) for fusion fitting and threshold selection. The shared held-out test set is unchanged. Parameters and selected thresholds remain fixed during test evaluation. The representational development split and $\mathcal{D}_f$ are identified separately below when reporting their respective operating points.

Score definitions. Table B.1 distinguishes the four detection outputs. The representational readout uses the frozen pre-emission hidden state. The standalone Jacobian score uses slot-level probability mass without the representational logit. AGR-slot combines that mass with the frozen probe logit, whereas AGR-value uses the fitted anchor-value fusion output described in Section 3.2. All scores are oriented so that larger values indicate greater PM risk.

Table B.1: Detection score identities. Here $\ell_{\mathrm{pm}} = \mathrm{logit}(p_{\mathrm{pm}})$, and r_{val} denotes the AGR-value fusion logit, including its fallback.

Readout	Evaluated score	Interpretation
Representational	$p_{\mathrm{pm}}(c)$	Frozen representational probe output
Jacobian (slot)	$\log \pi_J(T_{\mathrm{slot}}(c) \mid h_c)$	Standalone slot-level score
AGR-value	$r_{\mathrm{val}}(c)$; $\rho(c) = \sigma(r_{\mathrm{val}}(c))$	Fitted value-level fusion
AGR-slot	$B(c) = \ell_{\mathrm{pm}}(c) + \log \pi_J(T_{\mathrm{slot}}(c) \mid h_c)$	Fixed-additive AGR-slot score

Anchor-value availability. The fitted AGR-value pipeline has an available anchor-value functional score for 23 instances in $\mathcal{D}_c$ and 48 of the 274 test claims. For the remaining 226 test claims, the fusion output falls back to the calibrated representational branch. All 274 claims contribute to the reported AGR-value metrics. These counts describe functional-score availability in the evaluated fusion pipeline.

Readout comparison. Table B.2 compares the configurations on the same test cohort. Both individual readouts provide substantial discrimination. AGR-slot attains the highest AUROC, improving over the representational readout by 7.4 percentage points and over the standalone Jacobian score by 6.8 points. AGR-value provides a smaller improvement. The two fused configurations differ in both their functional target and fusion rule; these results establish their relative performance rather than isolate either design choice.

Table B.2: Readout comparison on Shopping. AUROC is reported with its 95% bootstrap confidence interval.

Readout	AUROC [95% CI]
Representational	0.888 [0.846, 0.924]
Jacobian (slot)	0.894 [0.851, 0.934]
AGR-value	0.898 [0.857, 0.933]
AGR-slot	**0.962 [0.940, 0.979]**

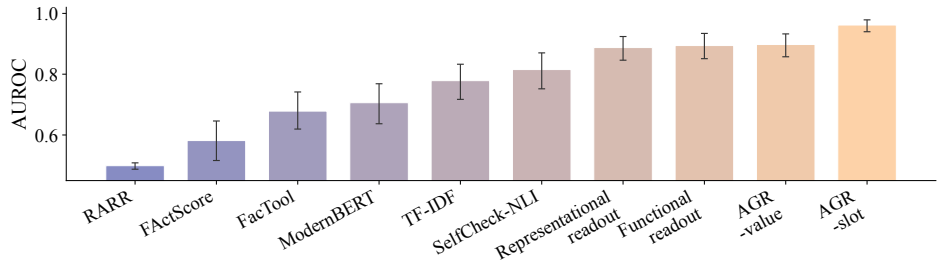


Figure B.1: AUROC of selected detection baselines and internal readouts. The ModernBERT variant shown in this plot is the quote-input variant.

C Detection Baselines and Complete Results

Ranking performance. Table C.1 extends the main comparison with the additional ModernBERT and SelfCheckGPT configurations. All methods are evaluated on the same 274 test claims. AUROC and average precision (AP) summarize ranking performance. Confidence intervals for detection use 2,000 trajectory-level bootstrap replicates. The internal-score audit verified identical claim identifiers and labels across the representational and AGR test outputs.

The standalone Jacobian, AGR-value, and AGR-slot scores reproduce AUROC/AP values of 0.8942/0.9089, 0.8976/0.9179, and 0.9618/0.9730, respectively. AGR-slot's result uses the AGR-slot score in Table B.1.

Table C.1: Extended detection results on the Shopping test cohort ($n = 274$). All metrics are on a 0–1 scale; AUROC and AP include 95% bootstrap confidence intervals. P/R/F1 use the common threshold of 0.05 for the designated probability-based methods. "– denotes internal scores whose operating points are reported in Appendix D. SC denotes SelfCheckGPT-NLI and MB denotes ModernBERT. † marks predictions that label every test claim as PM.

Method	AUROC [95% CI]	AP [95% CI]	P	R	F1
RARR gate	0.499 [0.487, 0.508]	0.569 [0.502, 0.630]	0.500	0.006	0.013
FActScore	0.582 [0.516, 0.646]	0.655 [0.575, 0.730]	0.594	0.628	0.611
FacTool	0.679 [0.619, 0.742]	0.679 [0.608, 0.745]	0.776	0.622	0.690
SC-NLI ($K = 5$)	0.693 [0.626, 0.757]	0.752 [0.676, 0.817]	0.598	0.974	0.741
MB (quote)	0.706 [0.637, 0.769]	0.729 [0.636, 0.809]	0.617	0.962	0.752
MB (structured)	0.762 [0.691, 0.826]	0.799 [0.716, 0.869]	0.612	0.962	0.748
TF-IDF	0.779 [0.717, 0.833]	0.783 [0.699, 0.859]	0.569	1.000†	0.726
SC-NLI ($K = 1$)	0.815 [0.752, 0.870]	0.851 [0.792, 0.903]	0.704	0.885	0.784
Anchor-only LLM	0.719 [0.664, 0.774]	0.727 [0.659, 0.786]	0.836	0.590	0.692
Source-agnostic LLM	0.759 [0.706, 0.812]	0.766 [0.698, 0.826]	0.883	0.628	0.734
Cite-or-drop LLM	0.788 [0.735, 0.838]	0.776 [0.709, 0.836]	0.831	0.788	0.809
DeBERTa-MNLI	0.829 [0.774, 0.879]	0.894 [0.853, 0.929]	0.736	0.840	0.784
Representational	0.888 [0.846, 0.924]	0.908 [0.860, 0.945]	0.765	0.917	0.834
Jacobian (slot)	0.894 [0.851, 0.934]	0.909 [0.869, 0.943]	–	–	–
AGR-value	0.898 [0.857, 0.933]	0.918 [0.872, 0.952]	–	–	–
AGR-slot	**0.962 [0.940, 0.979]**	**0.973 [0.955, 0.986]**	–	–	–

Recall-matched comparison. At approximately 91.7% recall, AGR-slot achieves 90.5% precision, compared with 77.3% for the representational readout, 77.7% for AGR-value, and 64.7% for the standalone Jacobian score. These are descriptive operating points obtained by matching recall on the test predictions. The representational precision differs from its fixed-threshold value because recall matching selects a different operating point. Development-selected thresholds are evaluated separately in Appendix D.

Paired ranking comparison. A paired DeLong comparison (DeLong et al., 1988) between AGR-slot and AGR-value yields $p = 4.83 \times 10^{-16}$. This comparison pairs detector scores by claim. The trajectory-level bootstrap intervals above account for grouping by trajectory; the standard DeLong test does not additionally model within-trajectory dependence.

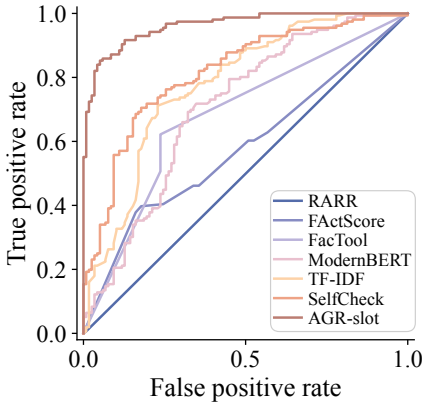


Figure C.1: ROC curves for selected detection methods on the Shopping test set.

C.1 Detection by Quote-Template Overlap

We stratify the test cohort using the quote-fingerprint protocol underlying the reported 66.1% training-template overlap. Of the 274 test claims, 181 share a quote template with the training set and 93 do not. The detectors remain frozen, and TF-IDF is fitted on training data. No retraining or threshold adjustment is performed for this analysis. Confidence intervals use 2,000 trajectory-level bootstrap replicates.

Table C.2: Detection stratified by training-set quote-template overlap. Cells report AUROC [95% CI]; PM+ is the number of PM-positive claims.

Subset	n	PM+	AGR-slot	Representational	TF-IDF
Overlap	181	108	.972 [.950, .989]	.903 [.860, .942]	.783 [.706, .851]
Non-overlap	93	48	.950 [.907, .984]	.882 [.805, .941]	.756 [.645, .855]

AGR-slot maintains an AUROC of 0.950 on the non-overlapping subset, compared with 0.882 for the representational readout and 0.756 for TF-IDF. Its advantage is therefore not confined to claims sharing a quote template with training examples. This exploratory analysis measures overlap under the specified fingerprint protocol, rather than all lexical or semantic similarities.

D Detection Threshold Selection

Common-threshold protocol. The main P/R/F1 comparison uses a common threshold of 0.05 for the designated probability-based methods. The Shopping shared-gate correction experiments use the frozen representational probe at this threshold. Figure D.1 shows its development-set sensitivity. At 0.05, TF-IDF predicts all 274 test claims as PM, yielding recall of 1.000. The internal ranking scores are evaluated below at their own development-selected operating points.

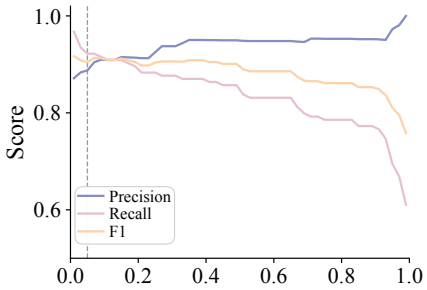


Figure D.1: Representational-gate precision, recall, and F1 on development set. The marked threshold of 0.05 is used for Shopping shared-gate claim selection.

Development-selected operating points. Table D.1 reports thresholds selected without test labels. The Jacobian threshold maximizes F1 on the representational development split. The $\mathcal{D}_f$ thresholds maximize F1 on the AGR fusion split ($n = 288$). For AGR-value, this split also fits the fusion weights, so it is not an independent threshold-validation set. We report both saved AGR-slot thresholds, identifying their respective

selection splits. These results supplement the common-threshold comparison; external baselines have not been uniformly retuned under this protocol.

The Jacobian operating point detects 120 PM claims with no false positives among the 118 correctly bound claims. Its 100% precision is the observed value on this finite test cohort. The two AGR-slot thresholds expose a precision–recall trade-off: the $\mathcal{D}_f$ threshold yields higher precision, while the representational development threshold yields higher recall, with similar F1 scores.

AGR-value on the probability scale. Applying $\rho(c) = \sigma(r_{\text{val}}(c)) \geq 0.05$ yields 61.3% precision, 99.4% recall, and 75.8% F1, with $(TP, FP, FN, TN) = (155, 98, 1, 20)$. This is a computable probability-scale operating point and strongly favors recall. The dashes for AGR-value in the main table denote separate operating-point reporting, rather than an inability to compute P/R/F1.

Table D.1: Internal detection operating points on the full Shopping test cohort. P/R/F1 are percentages; thresholds use the raw score scales in Table B.1.

Method	Selection split	Threshold	P	R	F1
Jacobian (slot)	representational dev	-7.064	100.0	76.9	87.0
AGR-value	$\mathcal{D}_f$	-0.7585	76.4	93.6	84.1
AGR-slot	$\mathcal{D}_f$	-8.822	91.9	87.2	89.5
AGR-slot	representational dev	-12.714	84.5	94.2	89.1

Table D.2: Confusion counts for the operating points in Table D.1. PM is the positive class.

Method	Selection split	TP	FP	FN	TN
Jacobian (slot)	representational dev	120	0	36	118
AGR-value	$\mathcal{D}_f$	146	45	10	73
AGR-slot	$\mathcal{D}_f$	136	12	20	106
AGR-slot	representational dev	147	27	9	91

E Baseline Implementation and Adaptation

Table E.1 summarizes the implementations and experimental roles of the baselines. Detection methods assess claim-level risk; mitigation methods additionally implement a correction procedure.

Table E.1: Implementation provenance and experimental roles of the evaluated baselines.

Method	Implementation and role
RARR	Adapted from pinned third-party prompts and executed against trajectory observations using our generation backend. Its agreement gate is evaluated for detection; its revision procedure is evaluated for mitigation.
CoVe	CoVe-inspired verification and revision procedure implemented from the published method description. Evaluated as a mitigation baseline.
Constrained copy	Self-designed, FEVER-style constrained-copy baseline. Evaluated for mitigation under the corresponding intervention protocols.
SelfCheckGPT, FacTool, FActScore	Adapted detection implementations using the available trajectory observations. Evaluated as detection baselines, without a corresponding mitigation implementation.
ModernBERT	Embedding-based classifier trained on the labeled split. Quote-based and structured-input variants are reported.
DeBERTa-MNLI	Anchor-aware verification implementation based on an NLI checkpoint. Evaluated for detection and verification, without a separate rewriting procedure.
LLM verifiers	Anchor-only, source-agnostic, and cite-or-drop verification configurations. Evaluated for claim-level detection and verification, without a rewriting backend.

Intervention protocols. The Shopping complete-intervention configurations of RARR and CoVe perform an exhaustive intervention scan using the adapted backends. In shared-gate experiments, every correction backend receives the same externally selected claims. This separates complete-system performance from correction behavior conditional on a common claim selection.

FAC implementation. Shopping shared-gate experiments use the frozen representational probe with a development-selected threshold of $\theta = 0.05$. Claims satisfying $p_{\mathrm{pm}}(c) > \theta$ enter FAC. This gate is the representational probe, not the AGR-slot AGR-slot score. FAC uses precomputed audit states derived from the frozen probe, slot adjudication, and Jacobian-lens readouts with top-$k = 50$. Functional audit thresholds are fitted on development data. In the end-to-end configuration, audit states are constructed during inference and Qwen3-8B performs evidence-grounded rewriting. The committed anchor is fixed throughout correction.

Table E.2: Score and threshold roles in Shopping detection and correction. Detection operating points and audit thresholds have different roles.

Role	Score or configuration	Rule
Primary detection ranking	AGR-slot $B(c)$	Threshold-independent AUROC/AP
Shared-gate claim selection	Representational $p_{\mathrm{pm}}(c)$	Flag if $p_{\mathrm{pm}}(c) > 0.05$
Standalone Jacobian detection	Slot log mass	representational dev F1 threshold -7.064
Jacobian audit partition	Slot log mass	Development PM-positive median -3.401
Functional audit readout	Slot lexicon, top-$k = 50$	Development-fitted audit thresholds
Evidence-grounded rewriting	Qwen3-8B	Deterministic decoding; fixed anchor

F Complete Mitigation Results

Tables F.1 and F.2 report complete-intervention and shared-gate results. Residual PM is measured at the trajectory level on the corresponding mitigation cohorts, whose initial rates differ from the characterization cohorts in Section 5.1. Mitigation confidence intervals use 500 bootstrap replicates.

Each cell reports residual PM with its 95% confidence interval and the output-to-original claim ratio R. This ratio measures output claim count, not retention of the original correctly bound claims, and can exceed one when an intervention generates additional claims. Zero-event bootstrap intervals can collapse to $[0, 0]$; they do not establish zero population risk.

Table F.1: Complete-intervention mitigation across the three τ^3-Bench domains. Initial PM rates are 52.4% for Telecom, 20.1% for Airline, and 3.1% for Retail. Each cell reports residual PM [95% CI] and the output-claim ratio R. FAC denotes Fixed-Anchor Correction and FAD denotes Fixed-Anchor Deletion.

Method	Telecom	Airline	Retail
FAD	$1.6\%\,[0.0, 3.7],\ R = 0.49$	$0.5\%\,[0.0, 1.6],\ R = 0.81$	$0.1\%\,[0.0, 0.3],\ R = 0.37$
RARR	$0.0\%\,[0.0, 0.0],\ R = 0.00$	$0.0\%\,[0.0, 0.0],\ R = 0.00$	$0.0\%\,[0.0, 0.0],\ R = 0.00$
CoVe	$0.5\%\,[0.0, 2.1],\ R = 0.01$	$0.5\%\,[0.0, 1.6],\ R = 0.01$	$0.0\%\,[0.0, 0.0],\ R = 0.00$
Constrained copy	$51.3\%\,[44.5, 58.4],\ R = 1.30$	$8.2\%\,[4.3, 12.0],\ R = 1.01$	$1.6\%\,[0.9, 2.4],\ R = 0.91$
FAC	$47.6\%\,[40.3, 55.0],\ R = 1.01$	$1.1\%\,[0.0, 2.7],\ R = 0.79$	$0.4\%\,[0.1, 0.8],\ R = 0.75$
FAC + RARR	$51.3\%\,[44.5, 58.4],\ R = 1.30$	$8.2\%\,[4.3, 12.0],\ R = 1.01$	$1.7\%\,[1.0, 2.5],\ R = 0.78$

Table F.2: Shared-gate mitigation across the same three domains. All methods receive identical flagged claims within each domain. Each cell reports residual PM [95% CI] and the output-claim ratio R.

Method	Telecom	Airline	Retail
FAD	$1.6\%\,[0.0, 3.7],\ R = 0.49$	$0.5\%\,[0.0, 1.6],\ R = 0.81$	$0.1\%\,[0.0, 0.3],\ R = 0.37$
RARR	$1.6\%\,[0.0, 3.7],\ R = 0.47$	$0.0\%\,[0.0, 0.0],\ R = 0.78$	$0.1\%\,[0.0, 0.3],\ R = 0.37$
CoVe	$1.6\%\,[0.0, 3.7],\ R = 0.47$	$0.0\%\,[0.0, 0.0],\ R = 0.78$	$0.1\%\,[0.0, 0.3],\ R = 0.37$
Constrained copy	$51.3\%\,[44.5, 58.4],\ R = 1.30$	$3.8\%\,[1.6, 6.5],\ R = 1.01$	$1.7\%\,[1.0, 2.5],\ R = 0.91$
FAC	$47.6\%\,[40.3, 55.0],\ R = 1.01$	$1.1\%\,[0.0, 2.7],\ R = 0.79$	$0.4\%\,[0.1, 0.8],\ R = 0.75$
FAC + RARR	$51.3\%\,[44.5, 58.4],\ R = 1.30$	$3.8\%\,[1.6, 6.5],\ R = 1.01$	$1.7\%\,[1.0, 2.5],\ R = 0.78$
Audit-only	$49.7\%\,[42.9, 57.1],\ R = 0.83$	$3.8\%\,[1.6, 6.5],\ R = 0.22$	$1.6\%\,[0.9, 2.4],\ R = 0.54$

FAC reduces residual PM from 20.1% to 1.1% on Airline and from 3.1% to 0.4% on Retail, with output-claim ratios of 0.79 and 0.75, respectively. On Telecom, FAC leaves 47.6% residual PM, compared with 1.6% for FAD. The relative value of revision and deletion therefore varies substantially by domain. The complete-intervention RARR configuration produces an output-claim ratio of 0.00 at the reported precision across these domains, so its zero residual PM must be read alongside the corresponding content loss.

G Rewrite-Backend Ablation

We replace FAC's rewriting procedure with the adapted RARR backend while retaining the intervention configuration. Table G.1 reports this comparison alongside the other Shopping mitigation methods. Complete-intervention and shared-gate results are shown separately because they use different detection configurations.

Table G.1: Complete Shopping mitigation results. PM is residual trajectory-level PM rate with 95% confidence interval. R denotes the output-claim ratio. The initial PM rate is 51.4%.

	Complete intervention		Shared gate	
Method	PM [95% CI]	R	PM [95% CI]	R
FAD	5.3 [2.5, 8.2]	0.358	5.3 [2.5, 8.2]	0.358
RARR	0.0 [0.0, 0.0]	0.041	5.3 [2.5, 8.2]	0.370
CoVe	0.4 [0.0, 1.2]	0.572	6.6 [3.7, 9.7]	0.691
Constrained copy	5.8 [2.9, 9.1]	1.128	10.7 [6.6, 14.6]	0.947
FAC	6.2 [3.3, 9.1]	0.807	5.3 [2.5, 8.2]	0.593
FAC + RARR	8.2 [4.5, 11.5]	0.926	9.1 [5.3, 12.3]	0.926

Under complete intervention, FAC reduces trajectory-level PM from 51.4% to 6.2% with an output-claim ratio of 0.807. Replacing its rewrite backend with RARR yields a higher output-claim ratio of 0.926 and a higher residual PM rate of 8.2%. The shared-gate comparison shows the same direction of change: FAC has 5.3% residual PM and $R = 0.593$, compared with 9.1% and $R = 0.926$ for FAC+RARR.

The shared-gate results also separate revision from deletion-oriented correction. FAC, FAD, and RARR all have 5.3% residual PM, while their output-claim ratios are 0.593, 0.358, and 0.370. FAC thus produces more output claims at the same measured residual PM rate. Correctly bound claim retention is evaluated separately in the main mitigation figures.

Table G.2: Shopping shared-gate action counts: 187 flagged and 87 unflagged claims.

Category	Claims
Unflagged / keep	87
Flagged: copy-rewrite	57
Flagged: probe-only deletion	125
Flagged: remaining cases	5
Total	274

Shopping branch accounting. At the shared probe threshold of 0.05, 187 of the 274 test claims are flagged and 87 are unflagged. The action summary records 57 flagged claims in

the copy-rewrite branch and 125 in the probe-only deletion branch. Table G.2 reports these counts and the five remaining flagged cases separately. All 57 rewritten claims were manually verified to be supported by anchor evidence.

H ANCHOR-AWARE VERIFICATION

Table H.1 evaluates anchor-aware verification on the held-out Shopping test cohort. PM recall measures the fraction of PM claims flagged. CB drop measures the fraction of correctly bound claims flagged for removal, and CB retention is its complement. The frozen representational probe provides the common-threshold reference for shared-gate claim selection.

Table H.1: Anchor-aware verification on the held-out Shopping test cohort ($n = 274$). PM R denotes PM recall; CB drop and CB ret. denote correctly bound claim false-drop rate and retention, respectively.

Verifier	PM R	CB drop	CB ret.
Anchor-only LLM	0.59	0.15	0.85
Source-agnostic LLM	0.63	0.11	0.89
Cite-or-drop LLM	0.79	0.21	0.79
DeBERTa-MNLI	0.84	0.40	0.60
Probe gate	0.92	0.37	0.63
Lexical anchor-only	0.71	0.40	0.60
Lexical source-agnostic	0.58	0.37	0.63
Lexical cite-or-drop	0.63	0.39	0.61

Cite-or-drop LLM attains 79% PM recall with 79% correctly bound claim retention, while the source-agnostic LLM retains 89% of correctly bound claims at 63% PM recall. DeBERTa-MNLI reaches 84% recall with 60% retention. The probe gate reaches 92% recall with 63% retention. These results characterize verification trade-offs before applying FAC's downstream audit and correction. The lexical variants provide additional controls under the corresponding verification configurations.

I LIMITATIONS

Our analysis concerns completed trajectories with recoverable entity commitments and observed entity-specific evidence. FAC preserves the committed anchor and therefore does not correct erroneous initial entity selection or earlier tool actions. Its ability to revise a claim depends on recovering supporting anchor evidence. The internal readouts provide predictive evidence about PM risk rather than establish a causal mechanism for the original failure.