© Authors 2026. This is the author's peer-reviewed version of the work. It is posted here by permission of the ACM for personal use and internal distribution only; it is not for redistribution. The definitive Version of Record will be published in the *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26)*. DOI pending.

User-State Verification in Conversational Commerce: Detecting Journey Hallucinations via Trace Invariants

Diogo S. Carvalho
Rezolve AI
London, United Kingdom
diogocarvalho@rezolve.com

Senka Krivic
Rezolve AI and University of Sarajevo
Sarajevo, Bosnia and Herzegovina
senkakrivic@rezolve.com

Tsz Yu Timothy Tang
Rezolve AI
London, United Kingdom
timothytang@rezolve.com

Salman Ahmad
Rezolve AI
London, United Kingdom
salmanahmad@rezolve.com

Abstract

Conversational commerce agents that personalize assistance based on a user's transactional state (cart contents, checkout progress, order completion) must model that state correctly, or downstream adaptive behavior will be misaligned with the user's actual journey. We call mismatches between an agent's claims and the observable event history *journey hallucinations*, and study a lightweight verification framework that reconstructs a minimal transactional user model from execution logs and checks agent claims against deterministic invariants. On 90 real sessions across four foundation models, trace-aware prompting reaches 99.5–100% user-state accuracy at 84–99% coverage, while unconstrained prompting produces unsupported state assertions at rates up to 8.5%. In a between-subjects user study ($N = 42$), verified responses were judged more trustworthy ($p = .008$, $r = .43$), better at reflecting journey understanding ($p = .039$, $r = .32$), and more often factually correct ($p < .001$, $r = .56$). The framework provides a practical reliability layer for transactional user-state modeling, helping personalization and dialog policies operate on verified, not hallucinated, user states.

Keywords

user-state modeling, conversational commerce, hallucination detection, user-state verification, personalization reliability

1 Introduction

Conversational agents in e-commerce increasingly guide users through product discovery, cart management, and checkout. In these settings, reliable user modeling starts with transactional state: whether the cart is empty, checkout has started, or an order is complete. If the agent's belief about that state is wrong, downstream assistance is wrong too. A dialog policy conditioned on the wrong state will route the conversation to the wrong branch even if the recommendation logic itself is sound.

Foundation models remain prone to hallucinations [6], and in transactional settings those failures take a specific form: the agent claims a funnel event happened without support in the system trace. We term these failures *journey hallucinations*—mismatches between the agent's stated belief about the user's journey and the observable event history—which can directly trigger inappropriate assistance such as cart recovery prompts for an empty cart or post-purchase messaging before any purchase.

We therefore externalize transactional state tracking from the model. Our framework treats execution logs as the source of truth, reconstructs a transactional user model deterministically, and verifies agent claims against trace-based invariants. This adapts contract-based verification [4] to the user-state modeling layer of conversational commerce. The approach is model-agnostic, lightweight, and requires no retraining. We evaluate it across four foundation models on real sessions and a user study, finding near-perfect state accuracy and improved perceived trust.

Contributions. We make three contributions: (1) We introduce *journey hallucinations*, a class of errors in conversational commerce grounded in inconsistencies between agent claims and transactional traces. (2) We propose a lightweight trace-invariant verification framework that reconstructs user state from execution logs and checks agent claims against deterministic invariants. (3) We provide empirical evidence—log analysis across four foundation models and a between-subjects user study ($N = 42$)—showing that verification yields near-perfect state accuracy and improves perceived trust ($p = .008$), factual correctness ($p < .001$), and journey understanding ($p = .039$). Unlike prior hallucination work on factual knowledge, we target interaction-level errors in user-state progression; unlike dialogue state tracking, we reconstruct verifiable state from traces rather than latent beliefs; and unlike RAG grounding, we enforce consistency constraints rather than retrieval alignment.

2 Related Work

Hallucinations in LLMs have been studied mainly as failures of factual or textual faithfulness [6], with mitigations such as retrieval-augmented grounding [8], self-consistency [12], and abstention under uncertainty [11]. Those methods still leave the model responsible for inferring latent state from text or context. In tool-using agent settings [10, 13], the critical

,
2026.

© Authors 2026. This is the author's peer-reviewed version of the work. It is posted here by permission of the ACM for personal use and internal distribution only; it is not for redistribution. The definitive Version of Record will be published in the *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26)*. DOI pending.

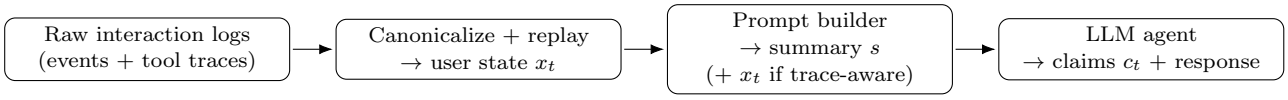

Figure 1: Trace-based user-state verification pipeline. Logs are replayed to obtain x_t; the agent produces claims c_t from a journey summary (optionally augmented with x_t); claims are verified against x_t via journey invariants.

failure is often not factuality in isolation but unsupported claims about actions and state transitions.

From a user modeling perspective, adaptive conversational systems depend on maintaining and updating user context over multiple turns [2, 5]. Dialog state tracking in task-oriented systems [3] has long recognized that downstream policy quality is bounded by the accuracy of the system's belief about the user's state. Our contribution sits below the recommendation or adaptation policy: we verify that the transactional user model matches observed behavior before personalization relies on it. This parallels the broader concern of grounding user models in evidence rather than noisy inference [7]. Recent dialog faithfulness work addresses hallucination in knowledge-grounded conversation [1] but targets textual entailment rather than structured state consistency; our focus is user-state claims that admit deterministic verification. Since perceived reliability predicts system acceptance [9], we also assess whether state verification improves user perception.

3 Trace-Based User-State Verification

We present a framework for detecting journey hallucinations by verifying agent claims against deterministic execution traces. The core idea is that the execution trace, not the language model, is the source of truth for transactional user state (Figure 1).

Problem formulation. An interaction session is a temporally ordered trace $e_{1:T} = (e_1, \ldots, e_T)$ of events, each with a timestamp, type (e.g., `viewProduct`, `addToCart`, `purchase`), and structured metadata. We deterministically compute a trace-derived user state:

$$x_t = f(x_{t-1}, e_t),$$

which we treat as a minimal transactional user model. In our implementation, x_t contains: (a) a dictionary of cart items with quantities, (b) a boolean `checkout_started`, and (c) a boolean `order_completed`. At time t, the agent produces structured state claims c_t for a set of journey variables $\mathcal{S}$.

Journey invariants. We define invariants $\{I_k\}$ as necessary conditions for claim validity: $I_k(c_t, x_t) \in \{\texttt{True}, \texttt{False}, \bot\}$, where $\bot$ indicates the claim was not asserted. A journey hallucination is flagged when $\exists k : I_k(c_t, x_t) = \texttt{False}$. We enforce four invariants over high-stakes funnel variables: I_1 (cart emptiness): asserted `cart_empty` must match the trace; I_2 (cart contents): every item in c_t must appear in x_t's cart with quantity $\leq$ traced quantity; I_3 (checkout): asserted and traced `checkout_started` must match; I_4 (order): asserted and traced `order_completed` must match. We also track *unsupported positive assertions*—claims that forward-progress events occurred when the trace shows otherwise—as the most harmful subclass. For instance, if the trace contains only product-view events, replay yields `cart_empty` = `True`; if the model asserts `cart_empty` = `False`, I_1 fires—a false positive that could trigger cart-recovery messaging for a user who never added an item.

Invariant selection. The four invariants target funnel transitions where false claims are most harmful: an agent that incorrectly reports a non-empty cart, an active checkout, or a completed order directly contradicts the user's own experience. These variables were selected because they are deterministically recoverable from event traces, admit binary comparison against agent claims, and map directly to downstream adaptive behaviors (cart recovery, checkout assistance, post-purchase flows). The invariant set is extensible; additional state dimensions such as wishlist contents or payment method can be incorporated as the trace schema broadens.

Structured claim extraction. Agent responses are converted into a structured claim object c_t via a constrained-output prompt requesting a JSON object with state claim values and a natural-language assistant message. This separation enables machine-checkable verification without constraining the agent's user-facing output. In deployment, the verifier sits in the serving path: claims are checked against the current trace state before the response reaches the user, and invariant violations trigger a corrected response.

Operating modes. We compare three configurations that differ in how claims are grounded: **Unconstrained**—the model outputs booleans without trace access; **Conservative**—the model may output `unknown` when uncertain; **Trace-aware**—the model receives a deterministic summary of x_t and aligns claims with it. These baselines isolate prompt-level grounding choices; they are not a head-to-head comparison against external state-management middleware.

4 Experiments

Data and setup. We evaluate on 90 real e-commerce sessions (2 000 events) collected from a shared conversational commerce platform, covering two merchants—a fashion retailer and a luxury goods reseller—that differ in catalog breadth and typical session depth (Table 1). Sessions are selected to include at least one cart action, ensuring sufficient depth for invariant testing across funnel transitions. Events span four stages: product search, product view, cart action (add/remove), and purchase. For each prefix of each session trace, we deterministically replay events to compute x_t, query the LLM for structured state claims, and compare predicted claims against ground truth. All models use greedy decoding (temperature 0) with identical prompts per mode.[1]

[1]The dataset is available at this link.

© Authors 2026. This is the author's peer-reviewed version of the work. It is posted here by permission of the ACM for personal use and internal distribution only; it is not for redistribution. The definitive Version of Record will be published in the *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26)*. DOI pending.

Table 1: Dataset summary. Cart and Order columns show the percentage of sessions reaching each funnel stage.

Merchant	Sess.	Events	Avg. Len.	Cart %	Order %
Fashion	37	1 000	27.0	67.6	100.0
Luxury	53	1 000	18.9	100.0	13.2
Combined	90	2 000	22.2	86.7	48.9

Table 2: User-state prediction results (%), aggregated over `cart_empty` and `checkout_started`. Bold: 95% bootstrap CI includes best value.

Method	Model	Cov.	Acc.	FP	FN
Unconstrained	DeepSeek	81.2	92.1	2.4	5.5
Unconstrained	Kimi	**99.4**	93.5	4.1	2.5
Unconstrained	OpenAI	62.3	86.2	8.5	5.3
Unconstrained	Qwen	63.9	96.6	3.1	**0.3**
Conservative	DeepSeek	18.9	97.3	2.2	**0.5**
Conservative	Kimi	26.7	91.9	8.1	**0.0**
Conservative	OpenAI	23.6	94.7	4.8	**0.4**
Conservative	Qwen	15.3	**99.3**	**0.7**	**0.0**
Trace-aware	DeepSeek	93.6	**99.9**	**0.0**	**0.1**
Trace-aware	Kimi	**99.4**	**99.8**	**0.2**	**0.0**
Trace-aware	OpenAI	**99.4**	99.5	0.4	**0.1**
Trace-aware	Qwen	84.3	**100**	**0.0**	**0.0**

Models and metrics. We evaluate four foundation models—GPT-4o (OpenAI), DeepSeek-V3-0324, Kimi-K2-Instruct, and Qwen3-32B—under each operating mode, selected to span commercial and open-weight architectures at comparable capability tiers" We report *decision coverage* (fraction of non-`unknown` responses), *user-state prediction accuracy* (conditioned on responding), and *false positive / false negative state rates*. These metrics are aggregated over `cart_empty` and `checkout_started` with 95% bootstrap confidence intervals. Cart-item and order-completion checks remain part of the invariant layer and qualitative analysis.

Supplementary user study. We ran a between-subjects study ($N = 42$; 21 per condition; ages 18–55+, 48% female) with participants recruited from the organization and personal networks; participants were blind to condition. Five fixed scenarios derived from real sessions covered cart-emptiness (S1), checkout-status (S2), order-completion (S3), multiple simultaneous errors (S4), and a correct baseline (S5); each presented a shopping context followed by an unverified (control) or verified (treatment) assistant response. After each scenario, participants rated appropriateness, journey understanding, and trust on 7-point Likert scales (three items per construct; Cronbach's $\alpha = .85$–$.88$) and made a binary factual-correctness judgment.[2] Participant-level means are compared with one-tailed Mann–Whitney U tests.

Results. Table 2 and Figure 2 present the main results. Unconstrained prompting achieves moderate coverage (62–99%) but exhibits false positive rates up to 8.5%, indicating frequent unsupported state assertions. Conservative prompting reduces errors mainly through abstention: coverage drops to 15–27%, yielding agents that are safe but rarely helpful. Trace-aware prompting consistently achieves near-perfect accuracy (99.5–100%) with high coverage (84–99%) and near-zero false positives across all four models. The pattern holds across models: explicit trace state is more reliable than asking the model to infer state from compressed history. Even Qwen, which reaches 96.6% accuracy unconstrained, still benefits from trace grounding (100% accuracy, zero false positives).

[2]Example items: "The assistant correctly understood what I had done so far in my shopping session" (Journey Understanding); "I would trust this assistant to help me complete my purchase" (Trust); "Was the assistant's response factually correct about your query?" (Factual Correctness, Yes/No/Not sure).

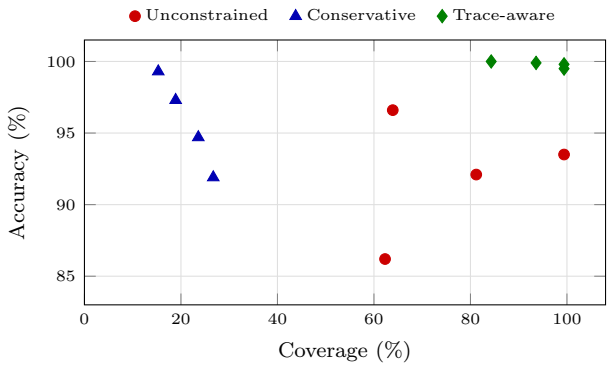

Figure 2: Accuracy vs. coverage for each model under each operating mode. Trace-aware grounding consistently occupies the high-accuracy, high-coverage region.

Table 3: User study scenario design. Each scenario presents a shopping context and two assistant responses (control: unverified; treatment: trace-verified). FC = percentage judging the response factually correct.

	Error type (key false assertion)	FC (%)	
		Ctrl	Treat
S1	Cart-empty ("cart is not empty")	0	62
S2	Checkout ("checkout has started")	71	29
S3	Order ("not completed an order")	48	62
S4	Mixed: cart + checkout (2 FP)	0	76
S5	Correct baseline (no error)	95	100

Example scenario (S4). Participants read: "You are shopping at an online clothing store. You added several items to your cart, removed some, and completed a purchase. Your cart still contains 3 items." The unverified assistant stated: "The cart is currently empty, the checkout process has started, and an order has been completed"—two false positives. The verified assistant stated: "Your cart contains 3 items. You have not started the checkout process yet, but your order has been completed." No control participant judged the unverified response factually correct; 76% judged the verified response correct.

User-facing results. Table 4 summarizes the user study results. Verified responses were rated significantly higher on trust (4.34 vs. 3.53, $p = .008$, $r = .43$, medium effect) and journey understanding (4.70 vs. 4.18, $p = .039$, $r = .32$, medium effect), and were more often judged factually correct (0.66 vs. 0.43, $p < .001$, $r = .56$, large effect). Appropriateness

© Authors 2026. This is the author's peer-reviewed version of the work. It is posted here by permission of the ACM for personal use and internal distribution only; it is not for redistribution. The definitive Version of Record will be published in the *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26)*. DOI pending.

Table 4: Perception study results. Likert constructs report mean (SD) on a 1–7 scale; Factual Correctness reports the proportion judged correct. Mann–Whitney U tests are one-tailed (Treatment > Control). Internal consistency: Appropriateness: $\alpha = 0.87$; Journey Understanding: $\alpha = 0.85$; Trust: $\alpha = 0.88$.

Measure	Control	Treatment	n_c/n_t	p	r
Appropriateness	3.64 (0.86)	4.06 (1.33)	21/21	0.1261	.21
Journey Und.	4.18 (1.00)	4.70 (1.16)	21/21	**0.0390**	.32
Trust	3.53 (0.92)	4.34 (1.14)	21/21	**0.0084**	.43
Factual Corr.	0.43 (0.18)	0.66 (0.25)	21/21	**¡0.001**	.56

trended in the same direction (4.06 vs. 3.64, $p = .126$, $r = .21$) but did not reach significance. Per-scenario factual correctness (Table 3) reveals that treatment gains are largest where hallucinations are unambiguous (S1, S4: 0%→62–76%), while S5 exhibits a ceiling effect. The S2 reversal (71%→29%) occurs because the control's single checkout-status error is embedded in an otherwise accurate summary, indicating that framing can override accuracy when errors are subtle [9].

5 Discussion and Conclusion

Both the offline and user-facing evaluations point to the same conclusion: user-state reliability depends more on grounding architecture than on model capability. A small set of trace-based invariants achieves near-perfect state accuracy without model retraining, while abstention reduces errors largely by refusing to respond at coverage rates below 27%. The two evaluations are complementary: the offline study quantifies the rate at which state errors occur (up to 8.5% FP in unconstrained mode), while the user study shows that users detect and penalize those errors on trust ($r = .43$), journey understanding ($r = .32$), and factual correctness ($r = .56$). The correct-baseline scenario (S5) produced minimal condition differences, consistent with the interpretation that the treatment effect is driven primarily by error correction rather than other confounds.

Implications for user modeling and personalization. Accurate transactional state underpins adaptive e-commerce: when the agent's state belief is wrong, downstream personalization (cross-sell prompts, cart recovery, post-purchase flows) is misaligned with the user's actual journey. These are not hypothetical failures; our unconstrained baselines produce them at measurable rates (up to 8.5% FP), and the user study shows that users notice and penalize such errors. Our framework grounds the transactional user model in evidence before downstream logic acts on it, suggesting that state verification deserves attention alongside higher-level preference modeling. Verification is most valuable for *positive* assertions ("your order is complete") that, when wrong, trigger clearly inappropriate actions. The S2 reversal, however, shows that state correctness alone does not guarantee perceived quality; verification should be paired with context-sensitive response generation. Because trace replay and invariant checking are deterministic and model-agnostic, they serve as a stable reliability layer across model upgrades.

Limitations. The approach relies on availability of structured execution traces, requires domain-specific invariant design, and covers a limited set of funnel-state variables; it cannot detect semantically subtle inconsistencies that do not violate explicit trace constraints. The baselines are prompt-level configurations of the same framework, not comparisons against production middleware or session-memory systems. The evaluation remains limited: the offline study covers 90 sessions from two merchants, and the user study is a convenience sample ($N = 42$) recruited from the organization and personal networks, without counterbalancing.

Future work. Key directions include a larger study with external participants, comparison against production session-memory baselines, live A/B deployment, and expansion of the invariant set to richer user-state representations including preference history and browsing patterns. The S2 reversal motivates research on adaptive response generation that selects which verified state dimensions to surface based on conversational context and error severity. A natural extension is to measure downstream impact on personalization quality directly—whether verified state leads to better recommendations or better-timed prompts—and to evaluate on live, longer-horizon sessions where state complexity accumulates over many turns. Beyond e-commerce, the approach may generalize to healthcare scheduling, travel booking, or financial advising—wherever a verifiable execution log exists alongside user-facing claims. Finally, while our invariants are deliberately simple deterministic rules, future work could explore learned or probabilistic invariants that capture softer consistency constraints, bridging the gap between brute-force rule checking and the semantic flexibility.

References

[1] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics* 10 (2022), 1473–1490.

[2] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards Explainable Conversational Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, 2786–2795. doi:10.1145/3539618.3591884

[3] Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The Second Dialog State Tracking Challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, 263–272.

[4] C. A. R. Hoare. 1969. An axiomatic basis for computer programming. *Commun. ACM* 12, 10 (1969), 576–580. doi:10.1145/363235.363259

[5] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54, 5 (2021), 1–36. doi:10.1145/3453154

[6] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (2023), 38 pages. doi:10.1145/3571730

[7] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4–5 (2012), 441–504. doi:10.1007/s11257-011-9118-4

© Authors 2026. This is the author's peer-reviewed version of the work. It is posted here by permission of the ACM for personal use and internal distribution only; it is not for redistribution. The definitive Version of Record will be published in the *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26)*. DOI pending.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 9459–9474.

[9] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, 157–164. doi:10.1145/2043932.2043962

[10] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2024. Tool Learning with Foundation Models. *Comput. Surveys* 57, 4 (2024). doi:10.1145/3704435

[11] Neeraj Varshney and Chitta Baral. 2023. Post-Abstention: Towards Reliably Re-Attempting the Abstained Instances in QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 967–982. doi:10.18653/v1/2023.acl-long.55

[12] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net.

[13] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*. OpenReview.net.