Grounding quality measurement for orchestrated AI reasoning chains. Evidence-binding rate as an evaluation primitive for orchestrator-in-the-loop agentic systems

Gil Raitses
aimez.ai

Abstract

Orchestrator-in-the-loop agentic systems route a user's request through a lead orchestrator that dispatches tool-bearing sub-agents and narrates a result, but Magentic-UI [20], a representative instance of this architecture, proposes no metric for whether the narrated output is bound to the evidence the orchestrator actually gathered. R_{uncited}, the fraction of scientific sentences in a system's output that carry no bound citation, is proposed here as a formal evaluation primitive for that orchestrated-reasoning quality. Using a reproducible 8-scenario parallel benchmark against the Gemini Interactions API with Google Maps grounding on marine science queries, Maps-only geospatial grounding achieves 60–100% uncited across query types. Injecting a structured orchestrated skill-dispatch step-log as RAG context drives R_{uncited} to 0%, not by generating more citations but by converting open-domain science queries into closed artifact-reference queries with zero scientific sentences left to cite. The resulting RAG lift hierarchy, step-log above structured data above narrow context above unstructured, diagnoses grounding architecture quality by context-type alignment, on this single-model, single-domain benchmark.

1 The evidence-binding gap in AI reasoning systems

R_{uncited} formalizes and measures the evidence-binding gap in orchestrator-in-the-loop agentic systems, systems that route a user's request through a lead orchestrator that dispatches tool-bearing sub-agents and narrates a result. Magentic-UI is a representative instance. A lead orchestrator directs sub-agents and Model Context Protocol tools with the human in the loop through co-planning, action guards, and memory [20]. This realizes, at agent scale, the mixed-initiative principle Horvitz set out, coupling automated services with the user under explicit uncertainty about intent [14]. Treating the human and the orchestrator as communicative agents distinguished only by function, agency, interactivity, and influence, is the agent-agnostic stance human-machine communication adopts for exactly this configuration [3, 33], with agency and control continually negotiated between human and system rather than fixed [11]. Magentic-UI is evaluated on task completion, interaction cost, and safety. It proposes no metric for whether the orchestrator's narrated output is bound to the evidence its sub-agents actually gathered. Users treat such systems as social actors [8] and accept fluent narration through heuristic rather than systematic processing [16], so an unbound claim can pass human review unchallenged, and calibrated reliance instead depends on signals the narration does not expose [13]. That missing measurement is the evidence-binding gap this paper measures.

Scientific and physical-domain outputs make claims requiring empirical grounding. On orca encounter likelihood at Haro Strait, a language model may name the salmon corridor, acoustic monitoring network, and behavioral studies, without binding any claim to generating observations, prior analyses, or data artifacts. The output is fluent without being evidentially accountable, a distinction from comprehension, which remains contested for these systems [19]. Fluent interfaces can present accountability that the underlying trace does not support [21], and prior work argues that such accountability gaps warrant evaluation beyond task-completion metrics [10]. An output-level binding criterion sidesteps the comprehension dispute by measuring traceability rather than adjudicating internal understanding.

Formally, the gap is the structural disconnection between stated claims and the observations, methods, and experiments that would make them verifiable, not factual accuracy per se, but whether a judge or downstream system can trace the claim to ground-truth data versus parametric pattern reproduction.

Prior work on RAG hallucination reduction [7, 9, 28, 24, 22] establishes that retrieval-augmented generation reduces factual error rates and that citation generation can be automated. However, this literature focuses on whether RAG improves accuracy, not on whether the improvement comes from evidence binding, from the system's outputs being traceable to their generating context, versus improved parametric knowledge activation.

The geospatial grounding baseline makes this distinction concrete. Google Maps grounding achieves high citation density for place-of-interest queries. For marine science evidence queries, the same domain, but requiring scientific rather than place-level grounding, Maps grounding resolves 0 of 25 citations to scientific or dataset sources. NOAA fisheries data, the Center for Whale Research census, Chinook salmon migration corridors, and acoustic hydrophone recordings are all generated as claims but remain entirely uncited. The system knows about these phenomena from training, but it cannot bind its claims to verifiable data artifacts.

R_{uncited}, the fraction of scientific sentences in a system's output with no bound citation, is the measurement primitive that makes this gap quantitative, reproducible, and architecture-diagnostic.

2 R_{uncited} as a formal evaluation metric

2.1 Definition

Let S be the set of sentences in a system's output. Define $S_{\text{sci}} \subseteq S$ as the sentences containing scientific or quantitative markers. For each sentence $s \in S_{\text{sci}}$, let $\text{cited}(s)$ be the indicator that s names at least one entity appearing in the bound citation set C.

$$R_{\text{uncited}} = \frac{|\{s \in S_{\text{sci}} : \neg\, \text{cited}(s)\}|}{|S_{\text{sci}}|} \tag{1}$$

$R_{\text{uncited}} = 0$ when every scientific sentence names a bound citation. $R_{\text{uncited}} = 1$ when no scientific sentence does. It is undefined when $|S_{\text{sci}}| = 0$.

2.2 Properties

R_{uncited} has four properties that make it useful as an evaluation primitive.

Claim-level granularity. The metric operates at the sentence level. A response with 10 scientific sentences and 9 bound citations still has $R_{\text{uncited}} = 0.1$, distinguishing it from response-level pass/fail metrics.

Architecture-diagnostic. Two systems with the same surface accuracy can have very different R_{uncited} values depending on whether their reasoning traces are passed as context to the language model. The metric discriminates between evidence-bound and parametric-knowledge-based claim generation.

Reproducible. R_{uncited} is computed from the raw API output. No human annotation is required. The scientific marker set is implemented as a regex over domain-specific technical terms.

Model-independent. R_{uncited} can be applied to any language model's output. It measures the binding between output and context, not internal model behavior.

2.3 Relation to prior evaluation frameworks

The Ragas framework [7] evaluates faithfulness, meaning whether claims are supported by retrieved context, along with answer relevancy and context precision. These metrics require a reference context. R_{uncited} measures a weaker but more deployable condition. It checks whether the model's output names sources in the context it was given, without requiring a predefined reference set.

The ALCE benchmark [9] evaluates citation generation at the claim level. R_{uncited} is complementary. It measures the rate of claims carrying no citation at all, regardless of citation validity.

R_{uncited} is also distinct from explainable-AI evaluation, which concerns the transparency of a model's decision process rather than the binding of its output to gathered evidence [5], and from reasoning-chain evaluation, which scores the internal correctness and informativeness of intermediate steps rather than their traceability to artifacts [25].

3 Benchmark methodology and results

3.1 Setup

The benchmark uses the Gemini Interactions API, model gemini-3.5-flash, Api-Revision 2026-05-20, with the `google_maps` tool at San Juan Islands coordinates 48.5465, −123.03. Eight scenarios run in parallel via Python asyncio, `tools/testing/grounding_parallel_rag.py`, each submitting a distinct query with a distinct RAG context type. The run date was 2026-06-24.

3.2 Scenarios

#	Context injected	Query type
1	Maps-only (none)	Where are SRKWs seen + evidence
2	Live gate data	Same + gate framing
3	Cell provenance	Cell intensity + kernels
4	Surface planner step-log	Gates blocking promotion
5	Maps-only (none)	Sighting check
6	Sighting-assist output	Same + orcast data
7	Maps-only (none)	Shore whale watching plan
8	Hotspot probability data	Same + orcast forecast

3.3 Results

Scenario	Context	R_{uncited}
4	Planner step-log	**0%**
8	Hotspot data	75%
1	Maps-only orca	60%
7	Maps-only trip	100%
2	Gate context	100%
3, 5, 6	Provenance / sighting	100%

Scenario 4 achieves $R_{\text{uncited}} = 0\%$. Scenario 8's hotspot data lowers R_{uncited} by 25 pp relative to Scenario 7, the same trip-planning query without that context at 100%, partially substituting for science claims.

4 The RAG lift hierarchy as a grounding architecture diagnostic

The 8-scenario results reveal a diagnostic structure, shown below.

Context type	R_{uncited}
Step-log (complete reasoning trace)	0%
Structured domain data (hotspots)	75%
Maps-only baseline	60–100%
Narrow structured data (gates)	100%

Structural alignment between context type and the science claims the model otherwise generates drives this pattern, rather than the volume of context injected.

Full grounding, step-log. The planner step-log records skills dispatched, panels selected, and artifacts referenced. The query is answerable from those references without open-domain science claims, so R_{uncited} collapses to 0% because $|S_{\text{sci}}| = 0$.

Partial grounding, hotspot data. Structured probability data substitutes for some generic science claims. The remaining $R_{\text{uncited}} = 75\%$ is ecology claims the hotspot data does not cover.

No grounding, gate data. Gate data covers statistical test outcomes only. The model still generates broader ecology claims that the gate data does not address.

Context type is a grounding affordance that shapes which claims become evidence-bound [6]. Complete reasoning traces remove uncited claims, and narrow structured context does not. The diagnostic should port to any domain where a reasoning trace can be injected as context, and eRAG retrieval evaluation [29] is a plausible route to testing that, though this paper runs the step-log condition only and has not run eRAG itself.

The Citation-Enhanced Generation framework [27] demonstrates that post-hoc citation binding reduces hallucination in chatbots. The hierarchy finding is a different mechanism. Pre-hoc trace injection removes the open-domain science claims that citation binding would otherwise need to address.

5 Extension to orchestrated and planning systems

5.1 The step-log as a planning trace

Scenario 4's surface planner step-log is a planning trace that records the query, the skills dispatched, and the intermediate states. Injected as RAG context, the model answers from trace artifact references rather than world knowledge, so $R_{\text{uncited}} = 0\%$ because the query becomes closed artifact-reference rather than open-domain science.

The paper "Reasoning with Language Model is Planning with World Model" [12] explicitly frames LLM reasoning steps as world-state traces and proposes evaluating whether those states are accurate. R_{uncited} addresses a different, evidence-binding dimension of the same traces. It asks whether each state is traceable to the observations that produced it, not whether the state is accurate. TRACE [32] demonstrates that constructing knowledge-grounded reasoning chains from structured context eliminates irrelevant retrieval noise in multi-hop queries, and the orcast step-log is this structure applied to skill-dispatch interactions. Chain-of-Thought evaluation [34] confirms that intermediate reasoning steps are causally important for correct final outputs, which supports treating step-log quality as one useful signal for planning-system evaluation, though this paper does not test that claim beyond the orcast step-log itself.

5.2 Applying R_{uncited} to orchestrated agentic systems

Section 4's benchmark applies, in principle, to any orchestrator-in-the-loop agentic system that produces a structured interaction trace. Magentic-UI is the anchor instance of this class [20]. Injecting its orchestration trace as RAG context and asking a domain-specific question is the same procedure Section 4 runs on the orcast step-log. Bound artifact references should drive R_{uncited} toward 0% and sparse traces should leave it high, by the same mechanism demonstrated in Section 5, though this paper benchmarks the orcast trace only and has not run Magentic-UI itself. As users are increasingly asked to accept and rely on such systems [17], the metric supplies an output-grounding measurement that the anchored Magentic-UI architecture does not provide on its own, a gap this paper does not generalize to every orchestrator-in-the-loop system without separately testing each one.

6 Limits and falsification

6.1 Scope limits

Domain specificity. The benchmark uses marine science queries. Extension to other domains requires updating the science marker regex for that domain's technical vocabulary.

Model dependence. I produced these results with gemini-3.5-flash on 2026-06-24. A model with stronger domain-specific knowledge generates fewer science claims on this benchmark, which would change the raw uncited count without changing the ranking of context types.

R_{uncited} does not measure semantic accuracy. A claim that is cited to a bound artifact but semantically wrong is not captured. Extension to factual accuracy requires a reference ground truth.

Step-log quality dependence. $R_{\text{uncited}} = 0\%$ requires a complete, structured step-log. Sparse or malformed traces do not achieve the same result on this benchmark.

6.2 Falsification conditions

Claim	Falsifying condition
Step-log injection achieves $R_{\text{uncited}} = 0\%$	Step-log-injected query produces uncovered science claims
Hierarchy is stable across query types	Different Scenario 4 query yields high R_{uncited}
Gate context achieves no lift	Gate data injection covers generated science claims
R_{uncited} diagnoses architecture quality	Two architecturally different systems produce identical hierarchies

6.3 Extension path

Three extensions are not yet run. Multi-domain validation would run the benchmark on pedestrian routing on pax, clinical care, and manufacturing. World model evaluation would apply the benchmark to JEPA-based world models such as V-JEPA [4] and measure whether planning trace injection reduces R_{uncited}. A semantic accuracy addition would add a second metric for whether cited claims are factually accurate relative to their bound artifacts.

6.4 Relation to physical-world evaluation, optional context

The Horvitz and Magentic-UI anchor in Section 2 establishes this paper's load-bearing gap. Separately, and as optional context rather than a second load-bearing claim, physical-world AI benchmarks tend to evaluate task completion or prediction accuracy rather than intermediate claim binding. PHYRE [2] measures physical reasoning via puzzle solve rate, STAR [31] evaluates situated reasoning in real-world videos via question-answering accuracy, ThreeDWorld [23] measures embodied planning by object transport success rate, and TravelPlanner [35] evaluates language agent planning by plan validity. The comprehensive Embodied AI survey [26] categorizes this evaluation family as navigation, interaction, manipulation, and reasoning accuracy. Whether R_{uncited} or a similar evidence-binding primitive is useful for that family, in the way psychological benchmarks supplemented task-completion metrics in human-robot interaction [15], is future work this paper does not test, not a gap this paper claims to have verified by absence in that literature.

6.5 World model evaluation as a future extension

A natural extension is to apply this benchmark to world model planning systems whose design explicitly requires sensor-grounded intermediate representations. LeCun's autonomous machine intelligence architecture [18] specifies that the world model module must learn to represent the information content of the percept that is relevant for predicting future states, a grounding requirement on intermediate representations. Joint-embedding predictive architectures, from I-JEPA [1] to V-JEPA [4], and Image World Models [30], are evaluated on downstream accuracy on image and video benchmarks, and none proposes a claim-level evidence-binding metric of the kind Section 2 establishes is missing from Magentic-UI specifically. Whether R_{uncited} on JEPA planning traces reproduces Section 5's diagnostic hierarchy is an open empirical question. Section 4's methodology requires only a structured planning trace, so running it on world model outputs needs no architectural change.

References

[1] M. Assran et al. *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. 2023. arXiv: 2301.08243. URL: https://arxiv.org/abs/2301.08243.

[2] A. Bakhtin et al. *PHYRE: A New Benchmark for Physical Reasoning*. 2019. arXiv: 1908.05656. URL: https://arxiv.org/abs/1908.05656.

[3] J. Banks and M. M. A. de Graaf. "Toward an Agent-Agnostic Transmission Model: Synthesizing Anthropocentric and Technocentric Paradigms in Communication". In: *Human-Machine Communication* 1 (2020), pp. 19–36. DOI: 10.30658/hmc.1.2.

[4] A. Bardes et al. *Revisiting Feature Prediction for Learning Visual Representations from Video*. 2024. arXiv: 2404.08471. URL: https://arxiv.org/abs/2404.08471.

[5] M. A. Clinciu and H. Hastie. "A Survey of Explainable AI Terminology". In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. Association for Computational Linguistics, 2019, pp. 8–13.

[6] J. L. Davis. "Affordances for Machine Learning". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2023, pp. 324–332. DOI: 10.1145/3593013.3594000.

[7] S. Es et al. *Ragas: Automated Evaluation of Retrieval Augmented Generation*. 2023. arXiv: 2309.15217. URL: https://arxiv.org/abs/2309.15217.

[8] A. J. Gambino, J. Fox, and R. A. Ratan. "Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm". In: *Human-Machine Communication* 1 (2020), pp. 71–85. DOI: 10.30658/hmc.1.5.

[9] T. Gao et al. *Enabling Large Language Models to Generate Text with Citations*. 2023. arXiv: 2305.14627. URL: https://arxiv.org/abs/2305.14627.

[10] P. M. Gardner and J. Rauchberg. "Feminist Cybernetic, Critical Race, Postcolonial, and Crip Propositions for the Theoretical Future of Human-Machine Communication". In: *Human-Machine Communication* 8 (2024), pp. 27–51. DOI: 10.30658/hmc.8.2.

[11] J. L. Gibbs, G. L. Kirkwood, C. Fang, and J. N. Wilkenfeld. "Negotiating Agency and Control: Theorizing Human-Machine Communication From a Structurational Perspective". In: *Human-Machine Communication* 2 (2021), pp. 153–172. DOI: 10.30658/hmc.2.8.

[12] S. Hao et al. *Reasoning with Language Model is Planning with World Model*. 2023. arXiv: 2305.14992. URL: https://arxiv.org/abs/2305.14992.

[13] K. A. Hoff and M. Bashir. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust". In: *Human Factors* 57.3 (2015), pp. 407–434. DOI: 10.1177/0018720814547570.

[14] E. Horvitz. "Principles of Mixed-Initiative User Interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. 1999, pp. 159–166. DOI: 10.1145/302979.303030.

[15] P. H. Kahn et al. "What is a Human? Toward psychological benchmarks in the field of human-robot interaction". In: *Interaction Studies* 8.3 (2007), pp. 363–390. DOI: 10.1075/is.8.3.04kah.

[16] K. Koban and J. Banks. "Dual-Process Theory in Human-Machine Communication". In: *Communication*. SAGE Publications Ltd, 2023. DOI: 10.4135/9781529782783.

[17] P. D. Koenig. "Attitudes toward Artificial Intelligence: Combining Three Theoretical Perspectives on Technology Acceptance". In: *AI & Society* 40 (2025), pp. 1333–1345. DOI: 10.1007/s00146-024-01987-z.

[18] Y. LeCun and K. Alahari. *Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence*. 2023. arXiv: 2306.02572. URL: https://arxiv.org/abs/2306.02572.

[19] M. Mitchell and D. C. Krakauer. "The debate over understanding in AI's large language models". In: *Proceedings of the National Academy of Sciences* 120.13 (2023), e2215907120. DOI: 10.1073/pnas.2215907120.

[20] H. Mozannar, G. Bansal, C. Tan, A. Fourney, et al. *Magentic-UI: Towards Human-in-the-loop Agentic Systems*. Microsoft Research AI Frontiers, MSR-TR-2025-40. 2025. arXiv: 2507.22358. URL: https://arxiv.org/abs/2507.22358.

[21] S. Natale. *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*. Oxford University Press, 2021.

[22] S. J. Semnani et al. *WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia*. 2023. arXiv: 2305.14292. URL: https://arxiv.org/abs/2305.14292.

[23] Unknown. *The ThreeDWorld Transport Challenge: A Visually Guided Task-and-Motion Planning Benchmark for Physically Realistic Embodied AI*. 2021. arXiv: 2103.14025. URL: https://arxiv.org/abs/2103.14025.

[24] Unknown. *AGREE: Achieving Ground Truth Faithful Generations with LLMs*. 2023. arXiv: 2311.09533. URL: https://arxiv.org/abs/2311.09533.

[25] Unknown. *ReCEval: Evaluating Reasoning Chains via Correctness and Informativeness*. 2023. arXiv: 2304.10703. URL: https://arxiv.org/abs/2304.10703.

[26] Unknown. *Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI*. 2024. arXiv: 2407.06886. URL: https://arxiv.org/abs/2407.06886.

[27] Unknown. *Citation-Enhanced Generation for LLM-based Chatbots*. 2024. arXiv: 2402.16063. URL: https://arxiv.org/abs/2402.16063.

[28] Unknown. *CRAG – Corrective Retrieval Augmented Generation*. 2024. arXiv: 2401.15884. URL: https://arxiv.org/abs/2401.15884.

[29] Unknown. *Evaluating Retrieval Quality in Retrieval-Augmented Generation*. 2024. arXiv: 2404.13781. URL: https://arxiv.org/abs/2404.13781.

[30] Unknown. *Learning and Leveraging World Models in Visual Representation Learning*. 2024. arXiv: 2403.00504. URL: https://arxiv.org/abs/2403.00504.

[31] Unknown. *STAR: A Benchmark for Situated Reasoning in Real-World Videos*. 2024. arXiv: 2405.09711. URL: https://arxiv.org/abs/2405.09711.

[32] Unknown. *TRACE the Evidence: Constructing Knowledge-Grounded Reasoning Chains for Retrieval-Augmented Generation*. 2024. arXiv: 2406.11460. URL: https://arxiv.org/abs/2406.11460.

[33] J. B. Walther. "Theories of Computer-Mediated Communication and Interpersonal Relations". In: *The Handbook of Interpersonal Communication*. Ed. by M. L. Knapp and J. A. Daly. 4th ed. SAGE, 2011, pp. 443–479.

[34] J. Wei et al. *Chain of Thought Prompting Elicits Reasoning in Large Language Models*. 2022. arXiv: 2201.11903. URL: https://arxiv.org/abs/2201.11903.

[35] J. Xie et al. *TravelPlanner: A Benchmark for Real-World Planning with Language Agents*. 2024. arXiv: 2402.01622. URL: https://arxiv.org/abs/2402.01622.

Appendix: Figures

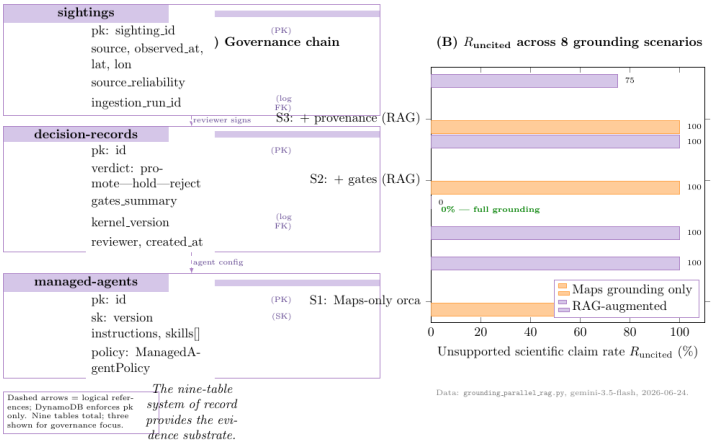

Figure 1: **Evidence-binding failure and its measurement.** (A) The orcast governance chain: sightings, decision-records, and managed-agents in DynamoDB. (B) R_{uncited} across eight grounding scenarios. Maps-only baseline: 60–100%. Surface planner step-log (Scenario 4): $R_{\text{uncited}} = 0\%$. Data: `grounding_parallel_rag.py`, gemini-3.5-flash, 2026-06-24.

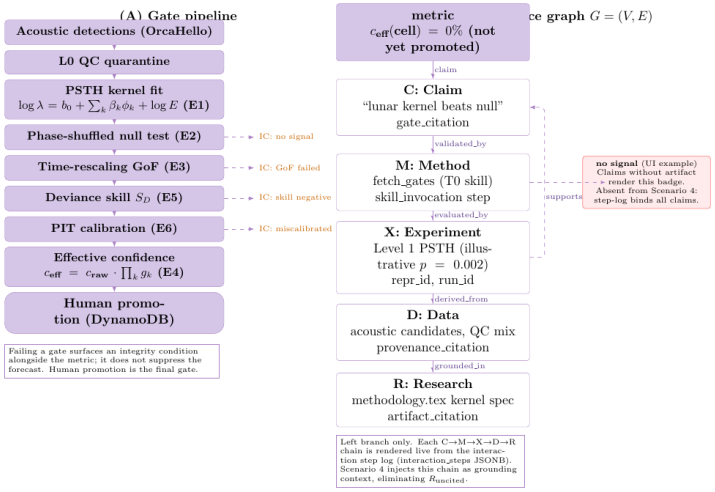

Figure 2: **The mechanism that produces $R_{\text{uncited}} = 0\%$.** (A) Gate pipeline from acoustic detections to displayed confidence. (B) Provenance graph $G = (V, E)$: Claim (C) → Method (M) → Experiment (X) → Data (D) → Research (R). Claims without a bound artifact render the "no signal" badge. Scenario 4 eliminates R_{uncited} because the planner step-log binds every claim to an artifact reference.

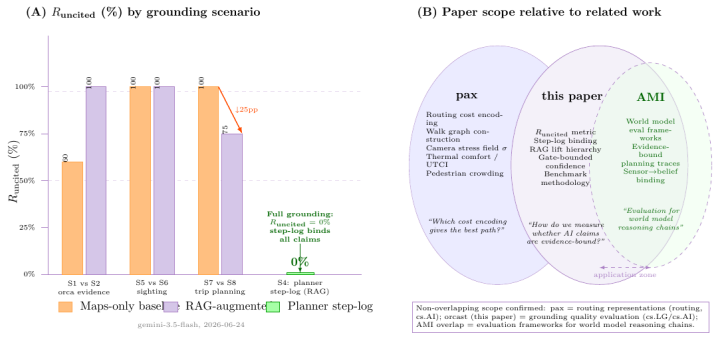

Figure 3: **Grounding quality hierarchy and paper scope.** (A) R_{uncited} for baseline vs RAG-augmented scenarios. The surface planner step-log (green, 0%) dominates; hotspot data provides partial lift (S7→S8); gate context provides no lift. (B) Paper scope: this paper addresses grounding evaluation methodology (center), distinct from pedestrian routing cost representations (pax, left), and directly applicable to world model evaluation infrastructure (AMI, right).