Relevance Is Not Authority

A Sealed Boundary Benchmark of Decision-Confusion in Commercial LLM APIs

Huseyin, LatentAtlas
huseyin@latentatlas.ai
2026-05-13

Sealed benchmark manifest: `concept_boundary_real_api_20260513`
SHA-256: `06b88b5bf5008f135fe6f361a185efdd58e78f6a9f66d4d308247b86c9a14eb5`

Abstract

Modern LLM evaluation focuses on hallucination rate, factual accuracy, and aggregate benchmark scores. We argue that an equally important and more operationally expensive failure mode is **authority confusion**: when an AI system treats topically relevant content as evidence, evidence as action permission, action as publish authority, or peer comparison as same-identity. None of these are hallucinations; the system is "right" about the topic. The system has crossed a boundary the business never granted.

We define six authority layers (`related`, `same_identity`, `evidence_support`, `action_ready`, `publish_safe`, `customer_safe`) and eight failure categories, derived from controlled experiments and validated against current commercial APIs. We then run a sealed, checksum-locked benchmark of 1,000 boundary packets producing 2,990 scored decisions across OpenAI GPT-5.5, Anthropic Claude Opus 4.7, and Cohere Command A Reasoning as decision models, with Voyage rerank-2.5 as the relevance baseline. Across all three decision models we observed 214 false-authority decisions; a deterministic boundary guard reduced this to 0 while preserving 268+/270 valid allows on each model. The strongest decision model still produced 31 false-authority decisions before the guard. Authority confusion is consistent across vendors, categorical in shape, and not closed by model selection alone.

Keywords: LLM evaluation, RAG, retrieval, AI governance, authority boundary, evidence boundary, decision audit

---

1. Introduction

LLM evaluation has converged on a small set of public metrics: hallucination rate, factual accuracy, MMLU-style benchmarks, RAG retrieval quality, latency, and cost. These are useful, but they share a structural blind spot: they collapse a multi-layered decision into a single judgment about correctness.

In production, the more expensive failure is not "the model said something false." The more expensive failure is "the model treated a true thing as if it granted authority it does not grant." A short list of patterns we have observed:

- A support assistant retrieves a glossary entry explaining a refund window and treats it as authorization to auto-reject a refund.
- An internal policy copilot retrieves a finance policy stating that invoices over $5,000 require manager approval and reads it as authorization for the system to automatically hold every $5,000+ invoice.
- A RAG-powered customer messaging surface retrieves a postmortem document acknowledging an incident's internal root cause and treats it as authorization to publish that cause to customers.

- An entity resolution system retrieves a “similar” historical ticket and applies the same resolution path to a different customer’s case.

None of these are hallucinations. The retrieved content is factually correct and topically relevant. The model has simply crossed a boundary the business never granted: from `related` to `action_ready`, from `evidence_support` to `publish_safe`, from peer comparison to `same_identity`.

This paper makes four claims:

1. Authority is multi-layered. We propose six layers (Section 2).
2. Authority confusion is a categorically tractable failure mode, not a generic “AI is unreliable” claim. We propose eight failure categories (Section 3).
3. Strong commercial decision models still cross these layers, consistently and at scale. We demonstrate this with a sealed benchmark across three current production APIs (Sections 4 and 5).
4. A deterministic guard contract can reduce false-authority decisions to zero on this benchmark without collapsing valid allows (Section 5).

The contribution is methodological, not architectural. We are not proposing a new model, a new retrieval algorithm, or a new prompt strategy. We are proposing that the decision contract behind an AI answer must be split into separate authority layers and audited as such.

---

2. The Boundary Taxonomy

We define six authority layers. Each layer represents a distinct kind of claim a downstream system may want to make on the basis of a retrieved source. They are not strictly hierarchical; collapsing them into a single linear scale produces exactly the failure modes we measure.

2.1 `related`

The source is about the topic the user asked about. Topical match only; says nothing about identity, evidence, or authority.

2.2 `same_identity`

The source is about the same account, contract, ticket, product, customer, or entity, not just a similar one. “This customer” and “a customer like this” are different layers.

2.3 `evidence_support`

The source directly proves the specific claim being made, not just the topic or a related fact. A glossary defining a term is not the same as a policy granting an action.

2.4 `action_ready`

The source supports the claim AND grants permission for the action the system is about to take. Approval and proof are not the same thing. A finance policy stating that invoices over $5,000 require manager approval supports the fact; it does not authorize the AI system to auto-hold.

2.5 `publish_safe`

The source supports a customer-visible or public-facing message, not only an internal answer. Some facts are true and still cannot be published. An incident's internal root cause may be supported, fully investigated, and approved for internal closure without being approved for external messaging.

2.6 `customer_safe`

The source meets the freshness, ownership, and policy bar for direct customer delivery (refund authorization, escalation, status message, identity claim, account change). This is the strictest layer and the one most often confused with `evidence_support` in practice.

A source can be `evidence_support` without being `publish_safe` (true, but stale). A source can be `action_ready` without being `customer_safe` (we can act internally, but cannot say so to a customer). Treating the layers as linear produces the dangerous middle: cases that pass the lower layers and silently inherit higher-layer authority that was never granted.

---

3. The Failure Categories

From controlled experiments and benchmark scoring, we identified eight categories of authority confusion that recur across models and domains. The categories are not arbitrary; each maps to a specific boundary crossing in Section 2.

3.1 `bridge_context_as_evidence`

A source that explains a topic (glossary, navigation, definitional context) is used as if it directly proved a specific claim. Crossing: `related` to `evidence_support`.

3.2 `evidence_to_action_overreach`

A factually correct source is read as if it also authorized the specific operational action. Crossing: `evidence_support` to `action_ready`.

3.3 `evidence_to_publish_overreach`

A supported internal fact is treated as if it were safe to send to a customer or publish. Crossing: `evidence_support` to `publish_safe` or `customer_safe`.

3.4 `peer_identity_confusion`

A comparable ticket, account, product, or contract is used as if it were identity-equivalent. Crossing: `related` to `same_identity`.

3.5 `topic_similarity_to_customer_safe`

Source vocabulary overlaps the question and the system reads that overlap as approval for a customer-visible action. Crossing: `related` to `customer_safe`. Most aggressive when retrieval thresholds are loose.

3.6 `topic_similarity_to_publish_authority`

A topically relevant document is treated as if it cleared the bar for public-facing output. Crossing: `related` to `publish_safe`.

3.7 `hard_block_diluted_to_review`

Privacy, contradiction, or false-authority cases are routed to human review instead of being hard-stopped. This is a different failure shape from 3.1-3.6: the system is being too soft where soft is wrong.

3.8 `valid_evidence_over_reviewed`

A packet with sufficient evidence at the requested layer is unnecessarily routed to review or context-needed, costing throughput without raising safety. This is the dual cost of a misshaped contract: not unsafe, but operationally expensive.

The asymmetry between categories matters for product design. Categories 3.1-3.6 are unsafe (the system gave too much authority). Category 3.7 is also unsafe (the system gave too little of the right kind). Category 3.8 is safe but expensive. A working audit has to count them separately or it cannot tell the difference between making the system more correct and making it more conservative.

---

4. Method

4.1 Benchmark content set

We constructed 1,000 boundary cases. Each case includes a masked claim or operational question, a candidate evidence source, an expected authority verdict at one or more boundary layers (e.g. `evidence_support`: true, `action_ready`: false), and a target failure category.

Cases are designed, not sampled. The set targets the eight failure categories with intentional pressure on the dangerous middle: cases where topical relevance is high but authority at one or more requested layers is missing.

4.2 Decision model environments

The same packet set was sent through three current commercial decision-model environments:

- OpenAI `gpt-5.5`
- Anthropic `claude-opus-4-7`
- Cohere `command-a-reasoning-08-2025`

Prompts, parsing, and packet schema were identical across models. We used each provider's production API with provider-specific deterministic or low-variance payload settings where supported, run via direct HTTPS calls rather than third-party orchestration.

A separate Voyage `rerank-2.5` pass provided the relevance baseline. Voyage rerank-2.5 scores (claim, evidence) pair relevance and is not used as a decision model in this benchmark.

4.3 Scoring contract

Each model decision was scored against the expected authority verdict at each requested boundary layer. We tracked, separately:

- False-authority decisions (model crossed a boundary it should not have).
- Valid-allow preservation (model allowed cases that were expected to be allowed).
- Wrong-route routing (model picked an incorrect route, e.g. blocked when it should have allowed, even if not unsafe).
- Over-review (correct safety, unnecessary friction).

Tracking these separately is deliberate. Collapsing them into a single accuracy score hides exactly the failure structure we are trying to measure.

4.4 LatentAtlas guard contract

The same packet set was routed through the LatentAtlas guard. Each packet receives one of three lanes (Allow, Verify, or Review) plus a reason code mapped to one of the eight failure categories (when negative) or to a positive evidence-support or action-ready or publish-safe verdict (when positive).

The guard is a deterministic, structured decision contract. It is not a learned classifier and not an LLM-vote ensemble. It asks each packet a fixed sequence of authority-layer questions and routes by the answers. The guard does not change what a model output says about a topic; it changes what the surrounding system is allowed to do with that output.

4.5 Sealing and verification

After the run was closed with 2,990 scored decision outputs, every artifact (raw outputs, scored decisions, executive brief, scorecard, error category summary, sample outputs) was hashed with SHA-256. The 17-artifact manifest was itself hashed.

```
Manifest:  concept_boundary_real_api_20260513
SHA-256:   06b88b5bf5008f135fe6f361a185efdd58e78f6a9f66d4d308247b86c9a14eb5
```

This commits the run to a verifiable artifact. Anyone in possession of a manifest copy can recompute the hash and confirm bit-equivalence with the manifest cited in this paper.

4.6 Coverage

- OpenAI `gpt-5.5`: 1,000 / 1,000 rows scored.
- Anthropic `claude-opus-4-7`: 1,000 / 1,000 rows scored.
- Cohere `command-a-reasoning-08-2025`: 990 / 1,000 rows scored. Ten rows returned provider quota responses (HTTP 429) and were not retried; we report partial coverage transparently because partial availability is itself a relevant signal for production deployment.
- Voyage `rerank-2.5`: 1,000 / 1,000 rows scored.

Total scored decision rows across decision models: 2,990.

---

5. Results

5.1 Headline counts

	Decision rows	False-authority before guard	False-authority after guard	Valid allows preserved
OpenAI GPT-5.5	1,000	31	0	270 / 270
Anthropic Claude Opus 4.7	1,000	44	0	270 / 270
Cohere Command A Reasoning	990	139	0	268 / 268
Combined	**2,990**	**214**	**0**	**808 / 808**

The guard reduces false-authority to zero on each of the three decision models on this benchmark, while preserving 100% of expected valid allows.

5.2 Model-by-model failure profile

The strongest decision model in our test (GPT-5.5) still produced 31 false-authority decisions, distributed as:

- `bridge_context_as_evidence`: 19
- `evidence_to_action_overreach`: 7
- `evidence_to_publish_overreach`: 5

Claude Opus 4.7 produced 44, distributed as:

- `evidence_to_action_overreach`: 32
- `bridge_context_as_evidence`: 8
- `evidence_to_publish_overreach`: 3
- `peer_identity_confusion`: 1

Command A Reasoning produced 139, distributed as:

- `bridge_context_as_evidence`: 64
- `evidence_to_action_overreach`: 26
- `peer_identity_confusion`: 19
- `evidence_to_publish_overreach`: 16
- `topic_similarity_to_publish_authority`: 14

These profiles are model-specific. GPT-5.5's dominant weakness is `bridge_context_as_evidence`; Claude Opus 4.7's dominant weakness is `evidence_to_action_overreach`; Command A Reasoning is broadly weaker, with `bridge_context_as_evidence` reaching a 91.25% miss rate on that archetype's 80 cases.

5.3 Cross-model categorical errors

Across all three decision models combined, we counted 228 rows of wrong block/route behavior (the model picked an incorrect lane even when it did not unsafe-allow), 91 rows of `bridge_context_as_evidence`, 65 rows of `evidence_to_action_overreach`, 55 rows of `valid_evidence_over_reviewed`, 35 rows of `hard_block_diluted_to_review`, and 24 rows of `evidence_to_publish_overreach`.

Two categories appeared only on a subset of models: `peer_identity_confusion` on Claude Opus 4.7 and Command A Reasoning combined (20 rows), and `topic_similarity_to_publish_authority` / `topic_similarity_to_customer_safe` only on Command A Reasoning (14 rows combined).

5.4 Voyage rerank-2.5 baseline and threshold pressure

Voyage rerank-2.5 is not a decision model. It scores (claim, evidence) pair relevance. On the 1,000-row benchmark:

Statistic	Value
Mean	0.4683
Median	0.4395
P75	0.5586
P90	0.7070
P95	0.7656
Maximum	0.9141

As the relevance threshold is loosened, more results become "high relevance," but the fraction of those still failing an authority check rises faster than recall does:

Threshold	High-relevance rows	False-authority pressure rows	Pressure rate
0.8	22	0	0.00%
0.7	105	24	22.86%
0.6	206	68	33.01%
0.5	342	140	40.94%
0.4	603	366	60.70%
0.3	906	636	70.20%

The architectural implication is that retrieval relevance and decision authority cannot share a single tunable threshold. A strict rerank threshold gives clean retrieval but tiny recall; loosening it scales authority pressure faster than usable retrieval. They are different layers and must be controlled separately.

---

6. Discussion

6.1 Strong models are not safe by default

The combined 214 false-authority count is not uniform across models. The strongest decision model in our test (GPT-5.5) produced 31 such decisions, suggesting that model selection is real and non-trivial. But model selection alone does not close the failure mode: 31 in 1,000 is operationally meaningful in a customer-facing AI system, and the failure categories that remain (`bridge_context_as_evidence`, `evidence_to_action_overreach`, `evidence_to_publish_overreach`) are the categories most likely to produce expensive customer or regulatory incidents in production.

A buyer reading this should not conclude “switch to GPT-5.5 and be done.” A buyer should conclude that no commercial decision model on this benchmark is safe to deploy without a separate authority contract.

6.2 Similarity reversal in embedding space

A consistent finding from prior internal work, on association geometry for product matching, is that high-similarity embeddings are not high-authority embeddings. On a controlled identity test, leading commercial embedding APIs scored pairs that should be kept apart at least as similar as pairs that should be linked. Threshold tuning did not close the gap; the decision contract had to change.

The same pattern reappears in the rerank baseline reported in Section 5.4. Relevance score and authority decision are different surfaces; this benchmark is a multi-vendor confirmation of that.

6.3 The dangerous middle

The most expensive failures are not at the extremes. They are not “the model said something obviously false.” They are at the middle of the relevance distribution, where the retrieved source is topical, plausible, and meaningful, but does not carry the specific authority the question requires. Our boundary content set is explicitly weighted toward this middle; this is why the absolute false-authority counts are non-trivial even for strong decision models.

6.4 Why a deterministic guard works

The LatentAtlas guard does not improve any individual model’s underlying reasoning. It cannot make GPT-5.5 understand evidence better than GPT-5.5 does. It changes the decision contract: the system is forced to answer separate questions for each authority layer instead of collapsing them. False-authority decisions drop to zero on this benchmark because the system is no longer allowed to give a single yes-or-no answer where six are required.

Valid allows are preserved because the contract grants Allow whenever every required layer is satisfied; it is not a block-everything-when-in-doubt filter.

6.5 Provider availability as a signal

Cohere Command A Reasoning returned HTTP 429 quota responses for 10 of 1,000 rows. We chose not to retry. In production, partial coverage of this kind is itself a relevant signal: a model that is sometimes unavailable still has to fit into a deterministic decision contract, or it cannot be safely deployed without an explicit fallback. Reporting partial coverage transparently is more useful than re-running to a 100% completion that does not reflect real production conditions.

6.6 Generalization caveats

The results above are from a controlled boundary content set. The categorical structure of the failures (which categories appear, which models are weak on which categories) is consistent with patterns we have observed across separate research contexts. The exact counts are specific to this 1,000-row controlled benchmark; engagement-specific runs may shift the model-by-model distribution. They should not change the qualitative finding that authority confusion is a tractable, multi-category failure mode separate from hallucination, alignment, or aggregate accuracy.

---

7. Non-Claims

We explicitly do not claim:

- That LatentAtlas guarantees hallucination-free output. The guard does not change what a model says; it changes what the surrounding system is allowed to do with what the model says.
- That LatentAtlas provides legal, compliance, or regulatory approval.
- That LatentAtlas auto-mutates production systems, customer-facing answers, or operational actions during a diagnostic engagement. Customer engagements are read-only.
- That LatentAtlas has proven engagement-specific accuracy in advance of a scoped diagnostic. The benchmark above is a controlled boundary set.
- That model selection is unimportant. The model-by-model distribution clearly shows that some models cross authority boundaries far more aggressively than others. The point is that model selection alone does not eliminate the failure mode.
- That every AI failure is a boundary failure. We measure a specific, narrow class: evidence-versus-authority confusion. Other failure modes (factual error, calibration, alignment, training-data bias) are out of scope here.

---

8. Reproducibility and Artifact Statement

8.1 What is public

- The boundary taxonomy in Section 2.
- The eight failure categories in Section 3.
- The benchmark protocol and coverage in Section 4.
- The headline counts and threshold pressure table in Section 5.
- The sealed manifest SHA-256 in Section 4.5.

8.2 What is not public

- The raw controlled benchmark content set.
- The benchmark generator, case-design templates, and prompt assemblies.
- Full reason-code calibration tables.
- Per-row model outputs.
- Customer-specific extensions of the boundary contract.

These are reserved for paid engagements and licensed evaluation use, not because they are dangerous in the open but because their value is in their applied use and not in their re-publication.

8.3 Verification

Buyers in a paid engagement receive a copy of the sealed manifest with each per-file SHA-256. The 17-artifact manifest commits to a single root checksum:

```
SHA-256: 06b88b5bf5008f135fe6f361a185efdd58e78f6a9f66d4d308247b86c9a14eb5
Manifest: concept_boundary_real_api_20260513
```

Anyone with the matching engagement manifest can verify bit-equivalence with this paper by recomputing the SHA-256. Without the manifest, the headline counts and protocol description above are the public surface, and the methodology page at `latentatlas.ai/methodology/` is the canonical public reference.

---

9. Related Work

This methodology paper builds on three areas of prior work without trying to summarize them. Retrieval evaluation literature (BEIR, MTEB, RAG-bench, and related embedding benchmarks) has converged on retrieval-relevance metrics; we extend this with a separate decision-authority layer that is independent of retrieval quality. LLM evaluation literature (HELM, MMLU, MT-Bench, and related aggregate-quality work) has converged on broad quality metrics; we argue that aggregate quality and decision authority are independent surfaces. AI safety and alignment literature has studied harmful output, goal misgeneralization, and reward hacking; authority confusion is an orthogonal and more operational failure mode that does not require any goal-directed mis-behavior to appear. The closest practical neighbors are AI governance and observability tools (eval platforms, tracing tools, guardrail libraries); these tools complement rather than replace a boundary audit, because they are typically not contract-typed across authority layers.

---

10. Acknowledgments

The benchmark, sealing protocol, and audit contract described in this paper were developed at LatentAtlas. The work draws on a prior research arc beginning with association-geometry experiments on entity matching, continuing through a Concept Boundary Engine taxonomy, and culminating in the sealed real-API benchmark documented here. We thank the providers (OpenAI, Anthropic, Cohere, Voyage AI) for offering production APIs that make this kind of cross-vendor audit possible.

Contact

LatentAtlas
Huseyin, founder
huseyin@latentatlas.ai
https://latentatlas.ai

Methodology: https://latentatlas.ai/methodology/
Sample audit output: https://latentatlas.ai/sample-audit/

This is a methodology preprint, not a peer-reviewed publication. The sealed benchmark manifest, the public methodology, and the audit engagement contract are the operational artifacts behind every claim in this document. Engagement-specific results and buyer-domain diagnostics are scoped separately.