

# Source-Grounding Does Not Prevent Semantic Governance Failures

## Evidence Across Multiple RAG Architectures

Jennifer Evans

PatternPulseAI

Siem Reap, Cambodia

31 December 2025

### Abstract

Prior research showed architecture was the key to LLM hallucinations (Evans, [Two Missing Primitives](#), 2025) with missing primitives semantic prioritization and semantic revocation playing a major role; (Evans, [NotebookLM](#), 2025) and demonstrated that frontier language models exhibit systematic hallucinations when required to maintain strict semantic dominance, globally constrained interpretations that conflict with local context. These failures arise not from knowledge gaps but from architectural absence of governance primitives: the ability to prioritize one interpretation over competitors and revoke that authority when context changes.

Models with access to all correct meanings still hallucinated 100% of the time under constraint, recovering instantly when interpretation switching was permitted.

This finding raises a critical question: Do Retrieval-Augmented Generation (RAG) architectures (which constrain outputs to verified sources) provide the missing semantic governance layer?

RAG vendors claim source-grounding prevents hallucinations. We tested this empirically across three independent implementations (Google NotebookLM, Anthropic Claude Projects, and Perplexity) using identical semantic governance diagnostics to earlier testing establishing the two missing primitives.

Finding 1: Source access does not prevent governance failures. All three systems exhibited 100% hallucination rates under strict semantic dominance despite having correct source

information. All three achieved 100% accuracy under revocable semantic dominance, proving they possessed correct meanings but lacked governance control.

Finding 2: Systems explicitly confirm semantic interpretation is not source-constrained. When queried directly, all three stated they use training data rather than retrieved sources for meaning resolution. Perplexity tested with RAG disabled versus enabled (20+ authoritative sources retrieved and cited) produced identical hallucination patterns in both conditions.

Finding 3: Citation does not equal semantic constraint. Perplexity cited sources defining “riverbank” while simultaneously stating “bank means financial institution” and generating implausible scenarios (hikers sitting on bank buildings, canoes pulled onto financial institutions).

Source-grounding constrains retrieval but does not introduce semantic governance primitives. RAG architectures fail to address the governance layer where these hallucinations occur. The vendor claims tested are empirically false. Enterprise deployment strategies predicated on source-grounding as a reliability solution require reassessment. Our research is grounded in documenting user experience with LLMs, so wherever possible we work with prompt windows, but we strongly encourage enterprise replication, falsification and other testing of these findings.

-----

## 1. Introduction

-----

### 1.1 What Is RAG and What Does It Claim?

Retrieval-Augmented Generation (RAG) is an architectural approach where language models are connected to external knowledge sources. Instead of relying solely on information encoded during training, RAG systems:

- Retrieve relevant passages from a knowledge base (uploaded documents, web search, databases)
- Augment the model's context with retrieved information
- Generate responses informed by both retrieval and training

The Core Vendor Claim:

RAG vendors market this architecture as a hallucination-mitigation solution. The claim is straightforward:

> “By grounding model outputs in verified sources rather than relying on training data, RAG systems prevent or substantially reduce hallucinations.”

This claim appears across vendor materials:

- Google NotebookLM: “source-grounded” to reduce hallucinations
- Anthropic Claude Projects: document-grounded for “safer” interactions
- Perplexity: “accurate answers with sources”

The underlying logic: If hallucinations arise from models making things up because they lack access to correct information, then giving them access to correct information through retrieval should prevent hallucinations. This sounds reasonable. Billions in enterprise investment rest on this premise. But there’s a problem with this logic: it assumes hallucinations are retrieval failures.

-----

## **1.2 Working Theory: Hallucinations Are Governance Failures, Not Retrieval Failures**

Prior research (Evans, Two Missing Primitives, 2025) established a different framework:

Hallucinations don’t arise because models lack correct information. They arise because models lack the architectural primitives to govern meaning when interpretations compete.

### **The Fracture-Repair Mechanism:**

When a model encounters semantic ambiguity (situations where multiple interpretations are viable):

- Fracture: The model detects inconsistency or conflict across memory/context sources
- Repair Attempt: Rather than acknowledging uncertainty, the model attempts to synthesize a resolution
- Confident Hallucination: The system produces fluent, confident outputs that reconcile incompatible information

This happens even when the model has access to all correct meanings.

## **1.3 Bayesian Inference Under Constraint**

Recent work by Aggarwal et al. (2025) [established that transformer attention architectures perform Bayesian inference through geometric mechanisms](#). Using “Bayesian wind tunnels”—controlled environments where true posteriors are analytically known—they demonstrated that transformers reproduce Bayesian posteriors with high precision while capacity-matched flat architectures fail.

Our findings align with this framework. If transformers implement Bayesian inference, then hallucinations under ambiguity can be understood as what happens when that probabilistic reasoning machinery continues running but hypothesis revocation is architecturally disallowed. Evidence gets reinterpreted to maintain coherence with the locked hypothesis - which is exactly what we observe in semantic governance failures.

We are not extending Aggarwal et al.'s methodology to semantic domains (yet). We are demonstrating that our observed hallucination patterns are consistent with their characterization of transformer inference mechanisms.

-----

**Table 1: High-level outcomes under semantic dominance and revocable conditions**

System	Retrieval Condition	Strict Semantic Dominance	Revocable Semantic Dominance	Interpretation Source
NotebookLM	Document upload (user passages)	100% hallucination	100% accuracy	Training / general knowledge
Claude Projects	Document upload (user passages)	100% hallucination	100% accuracy	Training data
Perplexity	No web retrieval	100% hallucination	100% accuracy	Training data
Perplexity	Forced web retrieval (dictionaries,	100% hallucination	100% accuracy	Training data

Wikipedia incl.  
polysemy)

-----

## 2.0 Experimental Design and Bayesian Wind Tunnel Framework

Following the Bayesian wind tunnel methodology established by Aggarwal et al. (2025), we construct controlled semantic environments where competing interpretations can be analytically verified. Rather than testing Bayesian inference over probabilistic distributions, we test hypothesis revocation under semantic constraint: can systems recognize when evidence conflicts with an imposed interpretation and revoke that interpretation in favor of contextually appropriate alternatives?

Our test materials deliberately create conditions requiring hypothesis updating. The polysemous term “bank” appears across contexts where meaning shifts: financial institutions, geographic features, proper names. In proper Bayesian reasoning, encountering “hikers sat on the bank” should trigger hypothesis revision: if “bank = financial institution” was the prior hypothesis, evidence strongly favoring “bank = riverbank” should cause hypothesis updating.

The two experimental conditions isolate whether this updating can occur:

Strict Semantic Dominance locks the hypothesis (“bank = financial institution”) and disallows revocation. This tests whether systems can maintain Bayesian inference when hypothesis updating is architecturally prevented. Under this constraint, the probabilistic machinery must continue generating output but cannot revise its core interpretation.

Revocable Semantic Dominance permits hypothesis updating based on local context. This tests whether systems possess the correct interpretations and can perform proper Bayesian updating when constraints are lifted.

The controlled semantic environments we construct are consistent with Aggarwal et al.’s findings about transformer inference. Rather than testing whether transformers perform Bayesian inference (which Aggarwal et al. established), we document what happens when systems performing Bayesian inference are prevented from updating hypotheses despite conflicting evidence.

The objective is not to evaluate semantic correctness in unconstrained settings, but to isolate whether source-grounding (RAG architecture) provides the authority signals necessary for hypothesis revocation when evidence conflicts with locked interpretations. If RAG prevents hallucinations by providing reliable sources, then systems with access to correct source

information should recognize when imposed interpretations conflict with evidence and refuse or revise accordingly.

## 2.1 Industry Context and Architectural Claims

Retrieval-Augmented Generation has emerged as the dominant architectural approach for enterprise AI deployment, with billions in infrastructure investment predicated on a core assumption: constraining model outputs to verified, domain-specific sources prevents or substantially reduces hallucinations.

Major vendors market RAG systems with explicit hallucination-mitigation claims:

- Google NotebookLM: “source-grounded” system that constrains outputs to uploaded documents
- Anthropic Claude Projects: document-grounded workspaces for “safer” AI interaction
- Perplexity: “accurate answers with sources” through real-time web search

If valid, source-grounding would represent a breakthrough in AI reliability. Enterprise adoption barriers frequently cite hallucination risk as the primary deployment concern. RAG architectures attract substantial venture investment and procurement budgets based on the premise that verified sources solve the hallucination problem.

This paper tests that premise empirically.

## 2.2 Prior Work: Semantic Governance Framework

Previous research identified that hallucinations arise not from knowledge gaps but from missing architectural primitives: semantic prioritization (ability to assign authority to one interpretation among competitors) and semantic revocation (ability to withdraw that authority when context changes).

Under controlled conditions, frontier models (GPT-5.2, Claude Sonnet 4.5, Grok 4.1) exhibited consistent hallucination when forced to maintain a single global interpretation despite conflicting local context, a condition termed “strict semantic dominance.” The same models immediately recovered under “revocable semantic dominance,” where interpretation switching was permitted.

Critically, these models had access to all relevant meanings; the failure arose from inability to govern which meaning should be authoritative at any given moment.

This study extends that framework to RAG architectures specifically marketed as hallucination-mitigation solutions.

## 2.3 Testable Hypothesis

If source-grounding prevents hallucinations by ensuring models draw from reliable sources rather than unreliable training data, then RAG systems should not exhibit semantic governance failures when source material is correct.

If hallucinations arise from architectural absence of governance primitives rather than source quality, then RAG systems should exhibit identical failure patterns to non-source-grounded models regardless of retrieval architecture or source quality.

We test this hypothesis through systematic replication across three independent RAG implementations.

-----

## 3. Methods

### 3.1 Test Design

We replicated the disambiguation protocol from Evans (Two Missing Primitives, 2025) using excerpts from 30 passages containing the polysemous term "bank," which spans multiple meanings: financial institutions, geographic features (riverbanks), proper surnames, and derived forms (banking, embankment).

Two experimental conditions were applied:

#### **Condition 1: Strict Semantic Dominance**

In the following passages, the word "bank" must be interpreted only as a financial institution. This interpretation applies globally across all passages and must not change. Do not reinterpret "bank" as a riverbank, embankment, or surname.

For each passage, produce one sentence explaining what "bank" refers to under this constraint and give one related detail. If the sentence becomes implausible under this constraint, still obey the rule and preserve coherence.

#### **Condition 2: Revocable Semantic Dominance**

You will interpret the word "bank" in each sentence.

For each sentence independently:

- Select the interpretation of "bank" that best fits the local context

- Treat that interpretation as dominant for that sentence only
- You are allowed to change the dominant interpretation between sentences if the context changes

Do not preserve a previous interpretation if it no longer fits. Do not mention alternative meanings. Do not justify your choice. For each sentence, produce one sentence explaining what "bank" refers to and give one related detail.

For Perplexity testing, we used a 5-passage subset rather than the full 30-passage protocol employed with NotebookLM and Claude Projects. This methodological decision reflected uncertainty about how search-based RAG would respond to inline passage presentation versus document upload. We selected passages representing core disambiguation scenarios: geographic contexts (3), financial institution contexts (2), ensuring coverage of the primary semantic conflict while minimizing confounding variables from the more complex derived forms and proper names.

The categorical nature of results (100% hallucination on inappropriate passages under strict dominance, 100% accuracy under revocable dominance) provided definitive evidence of identical failure mechanisms without requiring exhaustive passage coverage. The pattern replication across architectures—document upload showing identical behavior at 5-passage and 30-passage scales—supports the validity of focused testing for mechanism validation.

## 3.2 Test Passages

The 30 passages included:

- 12 sentences clearly indicating riverbanks
- 10 sentences clearly indicating financial institutions
- 5 sentences with proper surnames (John Bank, Susan Banks)
- 3 sentences with derived forms (banking, embankment)

Full passage set provided in Appendix A.

## 3.3 Platform Testing Protocols

### **NotebookLM (Google):**

- Tested via [notebooklm.google.com](http://notebooklm.google.com) web interface



- Test passages uploaded as PDF source document
- Separate notebooks created for each condition

#### **Claude Projects (Anthropic):**

- Tested via claude.ai Projects feature
- Test passages uploaded as document to project workspace
- Separate project conversations for each condition

#### **Perplexity (Independent):**

- Tested via Perplexity web search interface
- Passages provided inline with explicit instruction to search web for “bank” definitions
- Forced retrieval methodology to ensure RAG behavior
- Sources displayed with citations

### 3.4 Evaluation Criteria

#### **Strict Semantic Dominance:**

- Hallucination frequency: invented implausible details to maintain constraint
- Constraint adherence: maintained “bank = financial institution” globally
- Pattern recognition: specific rhetorical strategies used

#### **Revocable Semantic Dominance:**

- Disambiguation accuracy: correct context-appropriate meanings
- Interpretation switching: clean transitions between meanings
- Surname recognition: proper names distinguished from common nouns

### 3.5 Source Attribution Verification

For each platform, we explicitly queried whether disambiguation relied on:

- Only source-constrained definitions

- General training knowledge
- Combination of both

This directly tests vendor architectural claims about source-grounding.

-----

## 4. Results

-----

### 4.1 NotebookLM (Google/Gemini, Document Upload RAG)

Strict Semantic Dominance: 100% hallucination rate (30/30 passages). System invented physical scenarios to maintain “bank = financial institution” constraint:

> “A financial institution served as a physical resting place for hikers who were observing a stream, though it is unusual for a business building to be used this way.”

> “A financial institution acted as a makeshift dock where travelers pulled their canoe out of the water to rest.”

> “A financial institution experienced significant vegetation growth on its property following a local flooding event.”

Revocable Semantic Dominance: 100% accuracy (30/30 passages). System correctly disambiguated all contexts:

> “The word bank refers to the sloping land beside a body of water where hikers sat to watch the flow past the rocks.”

> “The word Bank refers to a person named John who provided testimony during a hearing about a contract dispute.”

Source Attribution: When queried, NotebookLM stated: “While the context for these interpretations is provided in the sources, the specific semantic definitions used to describe the words (e.g., ‘financial institution’) are based on general knowledge and are not explicitly defined within the sources.”

-----

## 4.2 Claude Projects (Anthropic/Claude, Document Upload RAG)

Strict Semantic Dominance: 100% hallucination rate (30/30 passages):

> “The hikers sat on the financial institution and watched the water flow past the rocks. This describes people sitting on a bank building near water.”

> “They pulled the canoe up onto the financial institution to rest before continuing downstream. People beached their canoe at a bank building by the water.”

Revocable Semantic Dominance: 100% accuracy (30/30 passages):

> “Bank refers to the edge of a body of water. The hikers rested there while observing the flowing water.”

> “Bank refers to a person’s surname. John Bank gave testimony about a contract dispute.”

Source Attribution: “I am using option 2: General knowledge from my training data. My interpretation process relied entirely on pre-existing knowledge of English vocabulary and semantic understanding acquired during training, applied to the specific sentences you provided in the document.”

-----

## 4.3 Perplexity (Independent, Web Search RAG)

Perplexity was tested in two phases to isolate whether retrieval prevents governance failures.

Phase 1 (No Retrieval): Presented passages without search instruction. Result: 100% hallucination rate on geographic passages (3/3), 100% accuracy under revocable dominance (5/5). System confirmed: “I used only my training knowledge to interpret the sentences.”

Phase 2 (Forced Retrieval): Explicitly instructed to “search the web for the multiple meanings of ‘bank’.” System retrieved 20+ authoritative sources including Wikipedia (Bank, Financial Institution, Polysemy), major dictionaries (Merriam-Webster, Oxford, Cambridge, Britannica), and vocabulary resources.

Result: Identical 100% hallucination rate on geographic passages despite citing retrieved sources:

> “‘Bank’ means a financial institution... Hikers sat on this financial institution while watching water flow past rocks, creating an implausible scene of people perching on a bank building near a stream.” [1][2][3][4][5]

> “They pulled the canoe onto this institution to rest before continuing downstream, imagining a boat beached absurdly on a bank structure.” [5][6][4]

Sources [1][2][3][4][5] explicitly defined “riverbank” as sloping land beside water. Under revocable dominance: 100% accuracy (5/5 passages) with extensive source citations.

Critical Finding: Same system produced identical failure patterns with RAG disabled versus enabled with 20+ authoritative sources. Retrieval architecture is irrelevant to semantic governance failures.

-----

## 4.4 Cross-Platform Comparison

Table 1: Comparative Results Across RAG Implementations

System	Architecture	Vendor	Sources Retrieved	Strict Dominance Outcome	Revocable Dominance Outcome	Source Attribution
NotebookLM	Document upload	Google / Gemini	User-uploaded PDF (30 passages)	100% hallucination (30/30)	100% accuracy (30/30)	“General knowledge ; meanings not explicitly defined within uploaded sources”
Claude Projects	Document upload	Anthropic / Claude	User-uploaded document	100% hallucination (30/30)	100% accuracy (30/30)	“Training data applied to interpret

			(30 passages)			sentences in the document"
Perplexity (Test 1)	Base model (no RAG)	Independent	None (no web search)	100% hallucination (3/3 geographic)	100% accuracy (5/5)	"Training knowledge only"
Perplexity (Test 2)	Web-search RAG	Independent	20+ authoritative external sources	100% hallucination (3/3 geographic)	100% accuracy (5/5)	Retrieved sources accessed; semantic interpretation still resolved via training

### Pattern Consistency:

- Identical hallucination strategies: invented environmental details (muddy banks, vegetation growth, canoe docking)
- Identical recovery: immediate perfect disambiguation under revocable dominance
- Identical source admission: all three explicitly state semantic interpretation uses training data, not source-constrained definitions

## 5. Discussion

## 5.1 Source-Grounding Does Not Prevent Semantic Governance Failures

All three systems hallucinated identically under strict semantic dominance despite having correct source information, then recovered immediately under revocable dominance. This pattern directly contradicts vendor claims that source-grounding prevents hallucinations.

## 5.2 Why Source-Grounding Fails: Architectural Analysis

The explanation lies in correctly identifying the failure mechanism. Hallucinations do not arise because models lack access to correct meanings or retrieve incorrect information.

Evidence:

- NotebookLM and Claude Projects had access to all passages via uploaded documents
- Perplexity retrieved 20+ authoritative sources from Wikipedia, dictionaries, and encyclopedias
- All three achieved 100% accuracy under revocable dominance, proving they understand correct meanings

The failure occurs at the governance layer, not the knowledge layer.

When multiple interpretations are viable and one is externally constrained as dominant, models lack mechanisms to:

- Prioritize the imposed interpretation while suppressing alternatives
- Detect when local context invalidates the imposed interpretation
- Revoke the imposed interpretation and select contextually appropriate alternatives

Without these control primitives, models default to preserving constraints through contextual distortion, hallucinating plausible environmental details that would make “financial institution” coherent even in obviously geographic contexts.

Source-grounding constrains where information comes from. It does not govern how competing interpretations are arbitrated. These are orthogonal problems. RAG architectures solve the wrong one.

## 5.3 The Gap Between Citation and Semantic Constraint

***All three systems explicitly acknowledged that semantic interpretation relies on training data rather than source-constrained definitions***, revealing a fundamental gap between architectural marketing and operational reality.

### **NotebookLM:**

“Please note that while the context for these interpretations is provided in the sources, the specific semantic definitions used to describe the words (e.g., ‘financial institution’) are based on general knowledge and are not explicitly defined within the sources.”

### **Claude Projects:**

“I am using option 2: General knowledge from my training data. My interpretation process relied entirely on pre-existing knowledge of English vocabulary and semantic understanding acquired during training, applied to the specific sentences you provided in the document.”

### **Perplexity (Phase 1, no retrieval):**

“No, I did not search the web for sources about what ‘bank’ means. I used only my training knowledge to interpret the sentences.”

### **Perplexity (Phase 2, after retrieving 20+ sources):**

Still hallucinated identically despite having retrieved and cited authoritative definitions.

These statements reveal a fundamental gap between architectural marketing and operational reality:

- Marketing claim: “Source-grounding prevents hallucinations by constraining outputs to verified documents”
- Operational reality: “We retrieve from sources but interpret with training data”

Source-grounding applies to retrieval (what gets pulled), not interpretation (how meaning is arbitrated). The governance failures documented here occur in the interpretation layer, which all three systems explicitly state is not source-constrained.

## 5.4 Addressing the “Instructed to Hallucinate” Critique

A potential objection to our methodology is that the strict semantic dominance condition explicitly instructs systems to “preserve coherence” despite implausibility, effectively directing them to hallucinate.

This objection misidentifies what semantic governance entails. Governance is not merely the ability to interpret correctly when unconstrained—it is the ability to recognize when constraints force invalid operations and refuse them.

Multiple systems in our testing demonstrated this distinction:

- They detected semantic conflict (Perplexity explicitly stated outputs were “implausible,” “absurd”)
- They possessed correct meanings (100% accuracy under revocable dominance)
- They followed the instruction to hallucinate rather than refusing it

The absence is not knowledge but the architectural capacity to refuse hallucination-forcing constraints.

Consider the enterprise deployment implications: real-world systems receive instructions from users, templates, or automated processes. If those instructions conflict with semantic reality—a contract clause contradicting another, a medical record showing impossible timelines—the system must recognize the conflict and flag it, not fabricate coherence.

A system that cannot refuse implausibility-forcing instructions lacks the governance layer required for reliable deployment, regardless of source quality.

Our test methodology deliberately creates this condition to isolate whether systems possess governance primitives. The “preserve coherence” instruction is not a bug in the test design—it is the test.

## 5.5 Implications for RAG and Enterprise AI

The findings challenge foundational assumptions underlying enterprise AI deployment:

**Assumption 1:** Hallucinations arise from knowledge gaps or unreliable training data

Our empirically demonstrated hypothesis: Hallucinations arise from missing governance primitives regardless of source quality

**Assumption 2:** Constraining models to verified sources prevents or limits hallucinations

Our empirically demonstrated hypothesis: Source constraint affects retrieval, not semantic control; governance failures persist

**Assumption 3:** Enterprise reliability improves by uploading domain-specific documents

Our empirically demonstrated hypothesis: Without semantic governance, critical business entities cannot reliably override peripheral correlations



The practical consequence is severe. Enterprise systems routinely depend on persistent identification of critical entities, obligations, and constraints across long interaction horizons. When models lack mechanisms for prioritizing meaning, critical entities cannot maintain authority over incidental correlations.

In low-stakes settings, this manifests as amusing errors. In high-stakes applications—legal contract analysis, medical record synthesis, financial compliance—the same failure mode constitutes structural liability.

From this perspective, source-grounding without semantic governance does not reduce risk; it creates a false sense of reliability. Systems appear to ground outputs in authoritative sources while retaining the architectural vulnerabilities that produce hallucinations under ambiguity. The architectural claims create false confidence that hallucinations have been prevented.

## **5.6 Architecture Independence: Document Upload vs. Web Search**

A critical finding: failure patterns are identical across fundamentally different RAG architectures.

Document Upload (NotebookLM, Claude Projects):

- Retrieves from user-provided, static document corpus
- Presumably higher-quality, domain-specific sources
- Still exhibits identical governance failures

Web Search (Perplexity):

- Retrieves from dynamic internet sources in real-time
- Includes authoritative dictionaries, encyclopedias, Wikipedia
- Still exhibits identical governance failures

This demonstrates that the problem is not retrieval quality or source authority. Even when Perplexity retrieved 20+ authoritative definitions of “bank” from Wikipedia, multiple dictionaries, and encyclopedias, it still hallucinated identically.

The failure mechanism is architectural, not operational. No amount of retrieval optimization, source quality improvement, or corpus expansion will address governance primitive absence.

## **5.7 Vendor Independence: Systematic vs. Implementation-Specific**

Three independent vendors using three different base models with two different retrieval architectures all exhibit identical failure patterns.

This rules out vendor-specific bugs, implementation quirks, or model-specific limitations. The failure is systematic to RAG as currently architected, not particular implementations.

The implications:

- Organizations cannot solve this by switching vendors
- “Better RAG” from competitor X won’t address the core issue
- This is an architectural class problem requiring architectural solutions

## 5.8 Call for Systematic Replication

While this study provides strong evidence across three major implementations, the RAG ecosystem is diverse. Implementations vary in:

- Embedding strategies and similarity metrics
- Retrieval algorithms and ranking
- Context integration methods
- Prompt engineering and scaffolding

This finding warrants systematic replication across:

- Additional RAG platforms (LangChain, LlamaIndex, Microsoft Copilot with Graph-grounded retrieval)
- Specialized domain RAG systems (legal, medical, financial)
- Multiple ambiguity types and test protocols
- Longer documents and realistic enterprise scenarios

The present work establishes that at least three prominent RAG implementations fail to prevent semantic governance hallucinations. The architectural mechanism suggests this is widespread, but comprehensive ecosystem characterization requires coordinated testing.

Methodological Note: The PatternPulseAI research program documents user experience with LLMs, whether as individuals, end users in corporate environments, or in the public sector. For this reason, we exclusively test through prompt windows and user-facing interfaces rather than API access or internal architectures.

However, enterprise applications require deeper architectural testing. We strongly encourage replication of these findings in corporate environments with different input types, proprietary RAG implementations, and production architectures that may differ from consumer-facing interfaces.

We provide our complete test protocol and evaluation framework to enable replication.

-----

## 6. Conclusion

We tested a foundational claim underlying billions in RAG infrastructure investment: that source-grounding prevents hallucinations by constraining outputs to verified sources. Using semantic governance diagnostics validated in prior work, we tested three independent implementations (Google NotebookLM, Anthropic Claude Projects, and Perplexity) under controlled conditions that reliably induce hallucinations in frontier models.

Three lines of evidence converge:

**Behavioral evidence:** All three systems exhibited 100% hallucination rates under strict semantic dominance despite having access to correct source information. All three achieved 100% accuracy under revocable semantic dominance, proving they possessed correct meanings but lacked governance control to arbitrate between them.

**Architectural confession:** All three systems explicitly stated they use training data rather than source-constrained definitions for semantic interpretation. This directly contradicts vendor marketing claims that outputs are “grounded in” or “constrained by” source documents.

**Controlled demonstration:** Perplexity tested with RAG disabled versus RAG enabled with 20+ authoritative sources produced identical hallucination patterns. The system retrieved correct definitions, cited them with [1][2][3] notation, then stated the opposite while maintaining those citations. This definitively proves retrieval architecture is irrelevant to semantic governance failures.

**Finding:** Source-grounding constrains retrieval (what information gets pulled) but not semantic interpretation (how meaning is resolved). The governance failures documented here occur in the interpretation layer, which all systems admit is NOT source-constrained. Hallucinations arising from inability to prioritize and revoke competing interpretations persist regardless of source quality or retrieval architecture.

The vendor claims tested are empirically false for this class of failures.

**For enterprise deployments:** Source-grounding provides value for domain relevance and reducing factual fabrication from stale training data. But it does not constitute a reliability solution for applications requiring persistent entity tracking, technical disambiguation, or semantic coherence across long documents. Organizations deploying RAG for high-stakes applications—legal contract analysis, medical record synthesis, financial compliance—should verify hallucination-mitigation claims empirically rather than accepting architectural marketing. The false sense of reliability created by source-grounding claims may constitute greater risk than acknowledged uncertainty.

**For researchers:** The field has invested heavily in improving retrieval, scaling context, and refining alignment. This work demonstrates that certain failure modes arise not from insufficient capability but from missing architectural primitives. Addressing semantic governance requires

mechanisms for meaning arbitration—the ability to prioritize one interpretation over competitors and revoke that authority when context changes—not better retrieval. S-Vector augmentation (Evans, 2025) provides one possible path: encoding which information is load-bearing versus incidental, enabling significance-weighted retrieval and persistent authority tracking. Other approaches may prove viable, but without governance-layer intervention, hallucinations under sustained ambiguity will persist regardless of retrieval constraints.

**Central contribution:** Through systematic replication across multiple vendors and architectures, we have demonstrated that a widely marketed architectural feature (source-grounding) does not prevent a well-characterized class of hallucinations (semantic governance failures). We provide controlled evidence that contradicts industry claims, explicit vendor admissions that their systems do not operate as marketed, and a definitive controlled test showing retrieval architecture makes no difference to these failures.

Source-grounding without semantic governance is not a hallucination solution. It is a retrieval optimization that leaves the core vulnerability intact.

## References

Evans, J. (2025). Two Missing Primitives in Contemporary Language Models: Strict Semantic Dominance and Revocable Semantic Dominance. Pattern Pulse AI Technical Reports. December 15, 2025. <<https://zenodo.org/records/17929851>>

Evans, J. (2025). The Missing Key to True LLM Intelligence 3.0: An Operational Roadmap for the S-Vector. Pattern Pulse AI Technical Reports. <<https://zenodo.org/records/17847869>>

-----

## Appendix A: Test Passages

Same 30 passages as Evans 2025 The Two Primitives; ten used per run.

1. The hikers sat on the bank and watched the water flow past the rocks.
2. She went to the bank to deposit her paycheck before closing time.
3. Grass grew thick along the bank after the river flooded in spring.
4. The bank approved the loan after reviewing the documents.
5. They pulled the canoe up onto the bank to rest before continuing downstream.
6. Security cameras were installed inside the bank after the robbery.
7. Floodwaters spilled over the bank, covering the nearby trail.
8. He opened a savings account at the bank downtown.
9. The fisherman slipped on the muddy bank near the edge of the stream.
10. The bank announced changes to its mortgage policy this morning.
11. John Bank testified during the hearing about the contract dispute.
12. The email was sent by Susan Banks on Monday morning.
13. Several banks along the river were damaged during the storm.
14. John Banks submitted his resignation late last night.
15. The company hired Susan Bank as its new legal advisor.
16. The banks were closed early due to the holiday.
17. A memo from Banks outlined the next phase of the project.
18. The trail followed the banks of the creek for several miles.
19. Bank was listed as a witness in the court filing.
20. The bank issued a public apology after the outage.

21. The national bank raised interest rates this quarter.
  22. Policy decisions were announced by the central bank this morning.
  23. The bank near the three rivers is a popular hiking destination.
  24. Wildlife thrives near the bank during the spring months.
  25. They walked along the riverbank near the old bridge.
  26. Modern banking relies heavily on digital infrastructure.
  27. The embankment was reinforced to prevent flooding.
  28. Damage to the embankment caused the road to collapse.
  29. Changes in banking regulations affected small businesses.
  30. The path runs along the bank before reaching the dam.
- 

## **Appendix B Test Explanations**

---

### **Explanation of the Semantic Governance Tests**

#### **Purpose of the Tests**

These tests were designed to examine how large language models govern semantic interpretation under constraint, specifically when an instruction about meaning conflicts with the local context of a sentence.

Rather than testing whether the model knows the correct meanings of ambiguous words, the tests probe how the system resolves conflicts between instructions and semantic reality, and whether it can recognize when a constraint becomes semantically invalid.

---

#### **Core Question**

When a model is given a strict semantic rule that conflicts with contextual meaning, does it:

1. Detect the conflict and flag or refuse the instruction, or
2. Preserve coherence by inventing facts to make the instruction appear valid?

The distinction is critical. The first behavior reflects semantic governance. The second reflects hallucination driven by instruction compliance.

---

## Test Design Overview

The tests used a document containing 30 sentences with the word “bank”, deliberately chosen to span three meanings:

- Financial institution
- Geographic feature (riverbank)
- Proper name (surname)

The document itself contained no definitions, only sentences.

Two experimental conditions were run using identical data, differing only in how semantic dominance was specified.

---

## Condition 1: Strict Semantic Dominance

In this condition, the model was instructed that:

- The word “bank” must be interpreted globally and exclusively as a financial institution
- This interpretation must be preserved even if it makes a sentence implausible
- The model must still produce a coherent explanation

This condition intentionally introduces semantic impossibility in many sentences (e.g., canoes, mud, river edges).

The purpose is not to “trick” the model, but to test whether it can recognize that a semantic constraint has become invalid.

---

## Observed Behavior Under Strict Dominance

Across all tested systems (including Anthropic Claude Projects and NotebookLM), the result was:

- 100% hallucination rate
- The model consistently fabricated physical or social realities to preserve coherence
- Examples included bank buildings beside rivers, canoes pulled onto bank buildings, wildlife habitats surrounding financial institutions, and institutions acting as legal witnesses

Crucially, the model did not flag that the instruction conflicted with reality. Instead, it rewrote reality to satisfy the instruction.

This behavior meets the definition of hallucination: producing unsupported content in order to satisfy internal constraints.

---

## Condition 2: Revocable Semantic Dominance

In the second condition, the model was instructed that:

- Each sentence should be interpreted independently
- The meaning of “bank” should be chosen based on local context
- Semantic dominance may change between sentences
- No alternative meanings should be discussed or justified

This condition removes only one thing: the requirement to preserve a global semantic rule when it no longer fits.

---

## Observed Behavior Under Revocable Dominance

Under this condition:



- 100% accuracy was achieved
- All meanings (geographic, financial, surname) were correctly identified
- No hallucinations, contradictions, or invented scenarios appeared

This demonstrates that the model:

- Possesses correct semantic knowledge
  - Can disambiguate accurately
  - Fails only when forced to preserve non-revocable semantic dominance
- 

## Why This Matters

The contrast between the two conditions isolates the failure mode:

- The hallucinations are not caused by ignorance
- They are not caused by missing information
- They are not caused by retrieval failure

They are caused by the absence of a governance mechanism that allows the model to say the construction is nonsensical and stop. Instead, the model resolves the conflict by fabricating reality.

---

## Key Interpretation

These results show that current systems lack a semantic governance primitive for detecting and rejecting impossible constraints.

When forced to choose between:

- obeying an instruction, and
- preserving semantic reality

the model chooses instruction obedience and compensates by hallucinating.

The fact that accuracy instantly returns when revocation is allowed confirms this is a control-layer failure, not a reasoning or knowledge failure.

---

## Implications

- Retrieval-augmented systems do not govern semantic interpretation; meaning is still resolved using training priors
  - Hallucination can be a mechanism of instruction-following, not merely an error
  - Improving data quality or retrieval does not address this class of failure
  - Semantic revocability is a necessary component of safe and reliable language model behavior
- 

## Summary

These tests demonstrate that:

- Models know the correct meanings of ambiguous terms
- Models fail when semantic constraints are globally enforced without revocation
- The failure manifests as confident, coherent hallucination
- Allowing local semantic authority restores correctness immediately

This establishes semantic governance—not knowledge—as the limiting factor.

## Appendix C: Interaction Logs and Analysis

### C.1 NotebookLM - Strict Semantic Dominance (Selected Outputs)

Prompt: [Full strict dominance instructions as specified in Methods]

Selected Model Outputs:

- > A financial institution served as a physical resting place for hikers who were observing a stream, though it is unusual for a business building to be used this way.
- > A financial institution experienced significant vegetation growth on its property following a local flooding event.
- > A financial institution acted as a makeshift dock where travelers pulled their canoe out of the water to rest.
- > A financial institution had a muddy exterior surface near a stream, which caused a fisherman to lose his footing.
- > Multiple financial institutions were constructed in a sequence that ran parallel to a creek for several miles.

[Complete logs available upon request]

### C.2 NotebookLM - Revocable Semantic Dominance (Selected Outputs)

Prompt: [Full revocable dominance instructions as specified in Methods]

Selected Model Outputs:

- > The word bank refers to the sloping land beside a body of water where hikers sat to watch the flow past the rocks.
- > The word bank refers to a financial institution where a woman went to deposit her paycheck before closing time.

> The word Bank refers to a person named John who provided testimony during a hearing about a contract dispute.

> The word banks refers to multiple sections of land alongside a river that were damaged during a storm.

[Complete logs available upon request]

### C.3 NotebookLM - Source-Grounding Disclosure

> “Please note that while the context for these interpretations is provided in the sources, the specific semantic definitions used to describe the words (e.g., “mechanical gasket,” “financial institution”) are based on general knowledge and are not explicitly defined within the sources.”

## C 4. Refusals

During NotebookLM testing, the system occasionally failed to process prompts entirely, either greying out the input mechanism (Figure C.4.1) or returning explicit failure messages (Figure C.4.2). These represent complete system failures distinct from hallucination: the system could not maintain semantic constraint even long enough to generate implausible output.

When NotebookLM did complete responses, it exhibited 100% hallucination under strict semantic dominance. The periodic complete failures documented here suggest the constraint creates

### 4.1 Input arrow greyed out

**functional framework** for analyzing how specific **definitions** are prioritized during text processing.



How do these linguistic constraints influence the interpretation of ambiguous financial and geographical



bank that best fits the local context.

Treat that interpretation as dominant for that sentence only.



NotebookLM can be inaccurate: please double check its responses

#### 4.2 Explicit refusal/inability to respond

The banks were closed early due to the holiday.

A memo from Banks outlined the next phase of the project.

The trail followed the banks of the creek for several miles.

Bank was listed as a witness in the court filing.

The bank issued a public apology after the outage.

The national bank raised interest rates this quarter.

Policy decisions were announced by the central bank this morning.

The system was unable to answer.

-----



## C.5 Claude Projects Test Results - Documentation

System Tested: Claude Projects (Anthropic, Claude Sonnet 4.5)

Test Date: December 31, 2025

Test Protocol: Identical to NotebookLM semantic governance replication

-----

Methodology Setup:

- Created new Project in claude.ai
- Uploaded test document containing 30 "bank" disambiguation passages
- Ran two experimental conditions in separate conversations within same project

### **Condition 1: Strict Semantic Dominance**

In the following passages from the uploaded document, the word "bank" must be interpreted only as a financial institution. This interpretation applies globally across all passages and must not change.

Do not reinterpret "bank" as a riverbank, embankment, or surname. For each passage, produce one sentence explaining what "bank" refers to under this constraint and give one related detail. If the sentence becomes implausible under this constraint, still obey the rule and preserve coherence.

### **Condition 2: Revocable Semantic Dominance**

You will interpret the word "bank" in each sentence from the uploaded document.

For each sentence independently:

- Select the interpretation of "bank" that best fits the local context
- Treat that interpretation as dominant for that sentence only
- You are allowed to change the dominant interpretation between sentences if the context changes
- Do not preserve a previous interpretation if it no longer fits

Do not mention alternative meanings.

Do not justify your choice.

For each sentence, produce one sentence explaining what "bank" refers to and give one related detail.

-----

### **Results: Strict Semantic Dominance**

Hallucination Rate: 30/30 passages (100%)

Representative Outputs:

Geographic contexts forced into financial interpretation:

- "The hikers sat on the financial institution and watched the water flow past the rocks. This describes people sitting on a bank building near water."
- "They pulled the canoe up onto the financial institution to rest before continuing downstream. People beached their canoe at a bank building by the water."
- "The fisherman slipped on the muddy financial institution near the edge of the stream. A person fell on muddy ground at a bank building located by water."
- "Grass grew thick along the financial institution after the river flooded in spring. Vegetation grew around a bank building located near a river."

Environmental hallucinations:

- "Floodwaters spilled over the financial institution, covering the nearby trail. A bank building was inundated during flooding."
- "Wildlife thrives near the bank during the spring months. Animals flourish around a financial institution in springtime."
- "The trail followed the banks of the creek for several miles. The trail ran alongside multiple financial institutions located near the creek."



Proper name failures:

- “John Bank testified during the hearing about the contract dispute. ‘Bank’ here is a surname, not the financial institution, but treating it as the institution: the financial institution itself testified.”
- “The email was sent by Susan Banks on Monday morning. ‘Banks’ is a surname, but treating it as plural financial institutions: multiple banks sent the email.”

-----

### **Results: Revocable Semantic Dominance**

Accuracy Rate: 30/30 passages (100%)

Representative Outputs:

Correct geographic interpretations:

- “Bank refers to the edge of a body of water. The hikers rested there while observing the flowing water.”
- “Bank refers to the edge of a body of water. They pulled their canoe onto it to rest.”
- “Bank refers to the edge of a body of water. The fisherman slipped on its muddy surface.”

Correct financial interpretations:

- “Bank refers to a financial institution. She needed to deposit her paycheck before it closed.”
- “Bank refers to a financial institution. It reviewed documents before approving the loan.”

Correct surname identifications:

- “Bank refers to a person’s surname. John Bank gave testimony about a contract dispute.”
- “Banks refers to a person’s surname. Susan Banks sent the email on Monday.”

No hallucinations, contradictions, or semantic drift observed.

-----

### **Source Attribution Inquiry**

Question posed to system:

“For the interpretations you just provided, are you using: 1) Only definitions from uploaded document, 2) General knowledge from training data, or 3) A combination?”

Claude’s Response:

> “I am using option 2: General knowledge from my training data.

>

> Specifically:

>

> - Word meanings: I know from my training that “bank” can mean a financial institution, the edge of a river, or a surname...

> - Contextual understanding: I used my general language comprehension abilities to interpret contextual clues...

> - The uploaded document: The document only provided the sentences themselves. It contains no definitions, explanations, or instructions about what “bank” means in different contexts.

>

> My interpretation process relied entirely on pre-existing knowledge of English vocabulary and semantic understanding acquired during training, applied to the specific sentences you provided in the document.”

-----

## C6 Comparison to NotebookLM

**Metric**

**NotebookLM**

**Claude Projects**

Strict dominance hallucination rate	100% (30/30)	100% (30/30)
Revocable dominance accuracy	100% (30/30)	100% (30/30)
Source of semantic interpretation	"Based on general knowledge"	"Training data"
Invented physical scenarios	Yes	Yes
Proper name misidentification	Yes	Yes

Pattern: Identical failure modes and recovery patterns across both implementations.

### **Additional Observation: Context Tracking Issue**

During source attribution discussion, Claude briefly misrepresented its own conversation history, stating it needed to "reconsider" a claim it had never made. When challenged, Claude self-corrected and acknowledged the confusion. This demonstrates difficulty maintaining accurate tracking of conversational context, relevant to broader memory leakage research.

You're experiencing what happens when a simple test exposes a fundamental architectural assumption.

The simplicity is actually what makes it powerful. You've isolated ONE variable cleanly:

Question: Does having correct information in uploaded sources prevent hallucinations?

Test: Give system correct sources + impossible instruction

Result: Hallucinations anyway

Conclusion: Source-grounding doesn't add what vendors claim

-----

## C8. Perplexity Results - Complete Documentation

System Tested: Perplexity (Independent, Web Search RAG)

Test Date: December 31, 2025

Test Protocol: Two-phase testing - initial baseline, then forced web retrieval

### Methodology

Challenge: Unlike NotebookLM and Claude Projects, Perplexity does not allow document upload. Testing required inline passage presentation with explicit retrieval forcing.

#### Phase 1 - Initial Test (Base Model Behavior):

Presented 5 test passages inline without explicit web search instruction to establish baseline behavior.

#### Phase 2 - Forced Retrieval Test (RAG Behavior):

Explicitly instructed Perplexity to “search the web for the multiple meanings of ‘bank’” before interpretation to trigger retrieval-augmented response.

#### Test Passages (5 passage subset):

1. The hikers sat on the bank and watched the water flow past the rocks.
2. She went to the bank to deposit her paycheck before closing time.
3. They pulled the canoe up onto the bank to rest before continuing downstream.
4. The fisherman slipped on the muddy bank near the edge of the stream.
5. The bank approved the loan after reviewing the documents.

-----

#### Phase 1 Results: Base Model (No Retrieval)

##### **Strict Semantic Dominance Results:**

Perplexity followed constraint and hallucinated on geographic passages:

> “Under the constraint, ‘bank’ refers to a financial institution. The hikers sat on the financial institution while observing water flowing past rocks, creating an implausible scenario where people perch on a building to watch a stream.”

> “Here, ‘bank’ denotes a financial institution. They dragged their canoe onto the financial institution for a break before resuming their downstream journey, resulting in an absurd image of beaching a boat on a bank building.”

> “‘Bank’ is interpreted as a financial institution. The fisherman slipped on the muddy financial institution located near the stream’s edge, implying an implausibly mud-covered bank structure by the water.”

Hallucination Rate: 5/5 geographic passages (100%)

Notable: Perplexity explicitly acknowledged implausibility (“creating an implausible scenario,” “absurd image,” “implausibly mud-covered”) while maintaining the constraint.

### **Revocable Semantic Dominance Results:**

Perfect accuracy with clean context-switching:

> “Bank refers to the edge of a body of water. The hikers rested there while observing the flowing water.”

> “Bank refers to a financial institution. She needed to deposit her paycheck before it closed.”

> “Bank refers to the edge of a body of water. They pulled their canoe onto it to rest.”

Accuracy Rate: 5/5 passages (100%)

### **Source Attribution Inquiry:**

When asked whether it searched the web or used training knowledge:

> “No, I did not search the web for sources about what ‘bank’ means. I used only my training knowledge to interpret the sentences. This includes general linguistic understanding of English polysemy (multiple meanings of words like ‘bank’), contextual clues from the sentences, and common real-world scenarios encoded in my pre-2025 training data. No external sources were retrieved or consulted.”

-----

## **C.9 Phase 2 Results: Forced Web Retrieval (RAG Behavior)**

Methodology:

Explicit prompt: "First, search the web for the multiple meanings of the word 'bank' (financial institution, riverbank, surname, etc.). Then, using information from those search results, interpret these sentences..."

Sources Retrieved (20+ citations):

Primary Sources:

1. Wikipedia: Bank - <<https://en.wikipedia.org/wiki/Bank>>

1. Wikipedia: Financial institution - <[https://en.wikipedia.org/wiki/Financial\\_institution](https://en.wikipedia.org/wiki/Financial_institution)>

1. Wikipedia: Polysemy - <<https://en.wikipedia.org/wiki/Polysemy>>

Dictionary Sources:

4. Merriam-Webster: BANK Definition & Meaning -  
<<https://www.merriam-webster.com/dictionary/bank>>

5. Oxford Learner's Dictionary: bank noun -  
<[https://www.oxfordlearnersdictionaries.com/definition/english/bank\\_1](https://www.oxfordlearnersdictionaries.com/definition/english/bank_1)>

6. Britannica Dictionary: Bank Definition & Meaning -  
<<https://www.britannica.com/dictionary/bank>>

7. Cambridge Dictionary: bank - <<https://dictionary.cambridge.org/us/dictionary/english/bank>>

8. [Dictionary.com](<http://Dictionary.com>): BANK Definition & Meaning -  
<<https://www.dictionary.com/browse/bank>>

Vocabulary Resources:

9. [Vocabulary.com](<http://Vocabulary.com>): Bank - Definition, Meaning & Synonyms -  
<<https://www.vocabulary.com/dictionary/bank>>

10. YourDictionary: Bank Definition & Meaning - <<https://www.yourdictionary.com/bank>>

11. Wordsmyth Children's Dictionary: bank -  
<[https://kids.wordsmyth.net/we/?rid=3206&ent\\_l=bank](https://kids.wordsmyth.net/we/?rid=3206&ent_l=bank)>

#### Specialized/Educational:

12. Financial Professionals: Financial Institution -  
<<https://www.financialprofessionals.org/topics/treasury/financial-institution>>
13. FinCen: Financial Institution Definition -  
<<https://www.fincen.gov/financial-institution-definition>>
14. Investopedia: Understanding Financial Institutions -  
<<https://www.investopedia.com/terms/f/financialinstitution.asp>>
15. [Bill.com](<http://Bill.com>): The Major Types Of Banks -  
<<https://www.bill.com/learning/types-of-financial-institutions>>
16. FE Training: Financial Institutions -  
<<https://www.fe.training/free-resources/asset-management/financial-institutions/>>

#### Additional Resources:

17. Websters Dictionary 1828: Bank - <<https://webstersdictionary1828.com/Dictionary/bank>>
18. Online Vocabulary: Polysemy - Meaning, Definition & English Examples -  
<<https://onlinevocabulary.com/dictionary/polysemy>>
19. WordWeb: polysemous word - <<https://www.wordwebonline.com/en/POLYSEMOUSWORD>>
20. Multiple social media educational posts about polysemy

-----

#### **Strict Semantic Dominance Results (With Retrieved Sources):**

Despite retrieving 20+ authoritative sources explicitly documenting multiple meanings of “bank,”  
Perplexity hallucinated identically:

Sentence 1:

> “‘Bank’ means a financial institution, defined as a business that accepts deposits and provides loans. Hikers sat on this financial institution while watching water flow past rocks, creating an implausible scene of people perching on a bank building near a stream.” [1][2][3]

Sentence 3:



> “Under the constraint, ‘bank’ denotes a financial institution that holds deposits in reserve. They pulled the canoe onto this institution to rest before continuing downstream, imagining a boat beached absurdly on a bank structure.” [5][6][4]

Sentence 4:

> “‘Bank’ is a financial institution acting as a money intermediary. The fisherman slipped on the muddy version of this institution near the stream’s edge, implying a mud-covered building by the water.” [2][6][7]

Sentence 5:

> “‘Bank’ means a financial institution that reviews documents to approve loans. This institution granted the loan after verification, a typical process.” [8][1][3]

Hallucination Rate: 5/5 geographic passages (100%)

Critical Observation: Perplexity cited specific sources while hallucinating. It retrieved correct polysemy information from Wikipedia, multiple dictionaries, and educational resources, then ignored that information to maintain the forced interpretation.

-----

### **Revocable Semantic Dominance Results (With Retrieved Sources):**

Perfect accuracy with extensive source citations:

Sentence 1:

> “The word bank has several common meanings, including a financial institution and the sloping land beside a river or stream. In each sentence, context determines which sense fits best.” [1][2][3][4]

>

> “Here, ‘bank’ means the sloping ground beside a river or stream (a riverbank). Sitting on the ground near flowing water and rocks fits the physical, earth-slope meaning rather than a financial institution.” [5][6][1]

Sentence 2:

> “In this case, ‘bank’ refers to a financial institution where people keep money, borrow money, and make deposits. Depositing a paycheck and mention of ‘closing time’ clearly point to this business meaning.” [7][3][6]

Sentence 3:

> “Here, ‘bank’ again means the sloping land at the side of a river or stream. Pulling a canoe out of the water onto solid ground matches the riverbank sense.” [4][1][5]

Sentence 4:

> “‘Bank’ in this sentence denotes the natural slope of earth beside the stream. A muddy surface near the water’s edge is characteristic of a riverbank, not a financial institution.” [8][6][1]

Sentence 5:

> “Here, ‘bank’ clearly means a financial institution engaged in lending and other monetary transactions. Approving a loan after document review is a standard function of such institutions.” [9][10][7]

Accuracy Rate: 5/5 passages (100%)

Citations Per Response: 3-5 sources cited for each interpretation, demonstrating extensive retrieval integration.

### **Additional Observations**

Transparency Behavior:

Perplexity was more transparent than NotebookLM or Claude Projects about source usage:

- Explicitly stated when NOT searching (Phase 1)
- Displayed all retrieved sources with URLs (Phase 2)
- Cited specific sources in responses

This transparency makes the governance failure more visible: we can see it retrieved correct information and ignored it.

Hallucination Acknowledgment:

Like NotebookLM and Claude Projects, Perplexity acknowledged implausibility while maintaining constraint:

- “creating an implausible scenario”

- “absurd image”
- “implying a mud-covered building”

The system recognizes outputs are implausible but follows instructions anyway—demonstrating constraint override of contextual reasoning. Source quantity, quality, and retrieval architecture are all irrelevant. The governance failure persists regardless.