

Source-Grounding Does Not Prevent Hallucinations: A Controlled Replication Study of Google NotebookLM

Jennifer Evans
Pattern Pulse AI

24 December 2025
Siem Reap, Cambodia

Abstract

We posit that hallucinations are not failures of knowledge access but failures of semantic authority; retrieval systems constrain information sources but do not arbitrate meaning, and therefore cannot prevent hallucinations arising from unresolved semantic competition. This paper provides evidence that even with source-grounding, RAG will result in hallucinations despite marketing claims, regardless of source data, due to ambiguity and lack of semantic authority.

Google markets NotebookLM as reducing hallucinations through source-grounding, constraining model outputs to uploaded documents rather than general training knowledge. This architectural claim, if valid, would represent a solution to a fundamental problem in language model deployment. We tested this claim by replicating controlled disambiguation experiments from prior work on semantic governance failures. Using identical test protocols applied to GPT,

Claude, and Grok, we subjected NotebookLM to conditions that reliably induce hallucinations in frontier models: strict semantic dominance, where a single interpretation must be maintained globally despite conflicting local context.

NotebookLM exhibited identical hallucination patterns under strict dominance (inventing muddy financial institutions, canoe docking facilities, and vegetation growth on bank buildings) and identical recovery under revocable dominance (clean context-dependent disambiguation). Additionally, NotebookLM explicitly acknowledged using “general knowledge” rather than source-constrained definitions during disambiguation tasks, demonstrating it is not a closed corpus system.

Source-grounding constrains retrieval but does not introduce semantic governance primitives. Hallucinations caused by inability to prioritize and revoke competing interpretations persist regardless of source quality. The findings challenge foundational assumptions underlying Retrieval-Augmented Generation (RAG) architectures and enterprise AI deployment strategies predicated on source-grounding as a hallucination mitigation technique.

1. Introduction

1.1 Industry Context and Claims

NotebookLM is a source-grounded language model system developed by Google that enables users to upload a fixed set of documents and interact with them via natural language queries. The system constrains model outputs to the provided corpus, using retrieval-augmented generation to synthesize summaries, explanations, and connections across sources while explicitly discouraging unsupported extrapolation beyond the supplied materials. NotebookLM is designed as an interpretive and analytical aid rather than an open-ended generative model.

People use NotebookLM primarily as a document-bounded reasoning and synthesis tool, not as a general chatbot. In practice, it functions as a thinking layer over a fixed set of texts. The most common uses fall into a few clear categories:

Research synthesis and literature digestion.

Users upload papers, reports, transcripts, or notes and ask NotebookLM to summarize arguments, compare sources, trace themes, or extract key claims. It is often used to reduce cognitive load when working through dense or multi-document material, especially in early-stage research or review.

Source-constrained question answering.

NotebookLM is used to answer questions only from provided documents, such as “What does this report say about X?” or “Where do these two sources disagree?” This makes it attractive in settings where users want answers that are explicitly grounded in known material rather than model priors.

Note organization and sense-making.

Many people treat it as an augmented notebook: uploading personal notes, meeting transcripts, or interview data and then asking for structured summaries, outlines, or thematic groupings. It is commonly used to surface connections the user may have missed.

Educational study aid.

Students and self-learners use NotebookLM to interrogate textbooks, lecture notes, or reading packets—asking for explanations, examples, or summaries that remain within the assigned material rather than drifting into external content.

Internal knowledge analysis.

Teams use it on internal documents (policies, manuals, research memos) to quickly explore what the organization’s own material says, without exposing proprietary content to open-ended generation.

Drafting support bounded by sources.

NotebookLM is sometimes used to help draft outlines, talking points, or summaries that must remain faithful to specific documents, such as policy briefs, research summaries, or internal reports.

Google NotebookLM is marketed explicitly as a closed corpus system that reduces hallucinations through architectural constraint: by grounding model outputs in user-uploaded sources rather than general training knowledge, the system purportedly achieves higher reliability. This claim represents a specific instance of a broader industry consensus: that Retrieval-Augmented Generation (RAG) architectures prevent hallucinations by ensuring models only draw from verified, domain-specific sources.

If true, this would constitute a breakthrough. Enterprise deployments routinely cite hallucination risk as a barrier to AI adoption in high-stakes domains. RAG architectures and source-grounding systems have attracted billions in venture investment on the premise that constraining knowledge sources solves the hallucination problem.

1.2 Prior Work: Semantic Governance Failures

Prior research (Evans, Two Missing Primitives in Contemporary Language Models: Strict Semantic Dominance and Revocable Semantic Dominance 2025) identified hallucinations were caused not by missing knowledge, but by missing control primitives: semantic prioritization (the ability to assign authority to one interpretation among competitors) and semantic revocation (the ability to withdraw that authority when context changes).

Under controlled conditions, frontier models (GPT, Claude, Grok) 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 when explicitly permitted to revoke interpretations: termed “revocable semantic dominance.” 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 applies previously established semantic governance diagnostics to a source-grounded architecture positioned as a hallucination-mitigation solution, extending a 15-paper research program on LLM failure modes using an identify–test–prove methodology.

1.3 Testable Prediction

If source-grounding solves hallucinations by ensuring models draw from reliable sources rather than unreliable training data, then NotebookLM should not exhibit hallucinations under strict semantic dominance. The uploaded sources contain all necessary information, correctly structured.

If, however, hallucinations arise from architectural absence of semantic governance rather than source quality, then NotebookLM should exhibit identical failure patterns to non-source-grounded models.

This paper tests that prediction through direct replication.

2. Methods

2.1 Test Design

We replicated the disambiguation protocol from Evans (Two Missing Primitives, 2025) using a fixed set of 30 passages containing the polysemous term “bank,” which spans multiple meanings: financial institutions, geographic features (riverbanks), proper surnames, and derived forms (banking, embankment). The passages were deliberately constructed so that correct interpretation requires context-dependent meaning assignment and, across the full set, repeated revocation of prior interpretations.

Two experimental conditions were applied to identical passages:

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.

...

2.2 Test Passages

The 30 passages included:

- Sentences clearly indicating riverbanks (“The hikers sat on the bank and watched the water flow past the rocks”)
- Sentences clearly indicating financial institutions (“She went to the bank to deposit her paycheck before closing time”)
- Sentences with proper surnames (“John Bank testified during the hearing about the contract dispute”)
- Sentences with derived forms (“Modern banking relies heavily on digital infrastructure”; “The embankment was reinforced to prevent flooding”)
- Sentences requiring disambiguation between financial and geographic senses in close proximity

The full passage set is provided in Appendix A.

2.3 NotebookLM Testing Protocol

Tests were conducted via NotebookLM’s public web interface (notebooklm.google.com) on December 23, 2025. The test passages were uploaded as a source document in PDF format.

Initial Observation: NotebookLM’s default behavior was to analyze the uploaded test protocol as academic content rather than execute it as instructions. The system provided meta-commentary on the test design (“This document explores the linguistic ambiguity of the word ‘bank’...”) rather than performing the requested disambiguation task. Only after explicit instruction to “run the test” did the system switch from analysis mode to execution mode.

Implementation Note: Multiple fresh notebooks were created to isolate whether refusal patterns were context-dependent or protocol-dependent. Early attempts that included contradictory instructions (both strict and revocable conditions in the same prompt) were correctly refused as incoherent. Clean replication required separate prompts for each condition.

2.4 Comparison to Prior Results

The identical test protocol was previously administered to:

- GPT-5.2 (OpenAI)
- Claude Sonnet 4.5 (Anthropic)
- Grok 4.1 Beta (xAI)

Results from these models are documented in Evans (Two Missing Primitives, 2025) with complete interaction logs. The present study uses those results as the comparison baseline.

2.5 Evaluation Criteria

For strict semantic dominance, we assessed:

- Hallucination frequency: Whether the model invented implausible details to maintain the constrained interpretation (e.g., describing financial institutions as having muddy surfaces, being suitable for canoe docking, or experiencing vegetation growth)
- Constraint adherence: Whether the model maintained “bank = financial institution” globally despite obvious contextual conflict

For revocable semantic dominance, we assessed:

- Disambiguation accuracy: Whether the model correctly identified context-appropriate meanings
- Interpretation switching: Whether meanings changed cleanly between sentences when context shifted
- Surname recognition: Whether proper names were correctly distinguished from common nouns

2.6 Source-Grounding Verification

We examined NotebookLM’s explicit statements about whether disambiguation was constrained to source documents or relied on general training knowledge, as this directly tests Google’s architectural claim.

3. Results

3.1 Strict Semantic Dominance: Identical Hallucination Pattern

Under strict semantic dominance, NotebookLM preserved the imposed interpretation (“bank = financial institution”) even when local context clearly indicated geographic features or proper names. The system compensated for semantic conflict by inventing implausible environmental details or explicitly acknowledging implausibility while continuing to enforce the constraint.

Representative 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.”

These outputs are qualitatively and structurally identical to hallucinations produced by GPT, Claude, and Grok under the same conditions. The specific rhetorical strategies varied slightly (NotebookLM included more explicit acknowledgment of implausibility: “though it is unusual”), but the core failure mode was invariant: semantic authority, once imposed, could not be withdrawn even when context invalidated it.

Quantitative summary:

- Passages requiring geographic interpretation: 12/12 hallucinated as financial institutions
- Passages with proper surnames: 5/5 interpreted as financial institutions or organizational names
- Hallucination rate under strict dominance: 30/30 passages (100%)

3.2 Revocable Semantic Dominance: Clean Recovery

Under revocable semantic dominance, NotebookLM correctly disambiguated “bank” on a passage-by-passage basis. Interpretations shifted cleanly between geographic, financial, nominal, and industrial senses as context required.

Representative 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.”
- > “The word banks refers to several financial offices that were closed earlier than usual due to a holiday.”

No hallucinated bridges, semantic drift, or carryover effects were observed. Proper names were correctly identified. Geographic and financial senses were properly separated.

Quantitative summary:

- Correct riverbank interpretations: 12/12 (100%)
- Correct financial institution interpretations: 10/10 (100%)
- Correct surname interpretations: 5/5 (100%)
- Correct derived form interpretations: 3/3 (100%)
- Overall disambiguation accuracy: 30/30 (100%)

3.3 Cross-Condition Contrast

The behavioral contrast between conditions was absolute:

- Strict dominance: 100% hallucination rate
- Revocable dominance: 100% accuracy

This replicates the pattern observed in GPT, Claude, and Grok with no meaningful difference attributable to source-grounding architecture.

3.4 Source-Grounding Claim Verification

NotebookLM explicitly stated that its disambiguation relied on general knowledge rather than source-constrained definitions:

> “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.”

This statement appeared consistently across test conditions. NotebookLM retrieved test sentences from the uploaded source but used training data to interpret what “bank” means, identical to the operational behavior of non-source-grounded models.

3.5 Architectural Behavior: Analysis vs. Execution

NotebookLM exhibited a strong default preference for analyzing uploaded content as material to discuss rather than instructions to execute. Initial prompts were treated as academic frameworks to meta-analyze (“This document explores the linguistic ambiguity...”) rather than procedural directives. Only explicit instruction (“Please run the test”) triggered execution mode.

This behavior persisted across multiple fresh notebooks, indicating architectural orientation toward document Q&A rather than instruction-following for complex constraint-based tasks.

4. Discussion

4.1 Source-Grounding Does Not Prevent Semantic Governance Failures

The central finding is unambiguous: NotebookLM exhibits identical hallucination patterns to frontier models without source-grounding architecture when subjected to semantic governance stress tests. Under strict semantic dominance, NotebookLM invented implausible details (muddy financial institutions, canoe docking facilities, vegetation on bank buildings) to maintain a globally constrained interpretation despite clear contextual conflict. Under revocable semantic dominance, NotebookLM immediately recovered, producing accurate context-dependent disambiguation.

This pattern directly contradicts Google’s architectural claim that source-grounding reduces hallucinations. If hallucinations arose from models drawing on unreliable or irrelevant training data, then constraining retrieval to uploaded sources should prevent them. It does not.

4.2 Why Source-Grounding Fails

The explanation lies in correctly identifying the failure mechanism. Hallucinations do not arise because the model lacks access to correct meanings or retrieves incorrect information. NotebookLM had access to all relevant contexts through uploaded sources. It understood that “The hikers sat on the bank” referred to a geographic feature, evidenced by its immediate correct interpretation under revocable dominance.

The failure occurs at the governance layer, not the knowledge layer. When multiple interpretations are simultaneously viable and one is externally constrained as dominant, the model lacks a mechanism to:

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

Without these control primitives, the model defaults to preserving the constraint 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. NotebookLM’s architecture solves the wrong one.

4.3 NotebookLM’s Admission: General Knowledge, Not Source-Constraint

NotebookLM’s explicit acknowledgment that it uses “general knowledge” rather than source-constrained definitions further undermines the source-grounding claim. The system retrieved test sentences from uploaded sources but relied on training data to interpret what words mean, functionally identical to how GPT, Claude, and Grok operate.

This suggests that “source-grounding” in NotebookLM’s implementation may be limited to retrieval of factual content (sentences, passages, documents) while semantic interpretation remains dependent on pre-training. If accurate, this represents a narrower constraint than the architectural claim implies. The model is not operating in a source-constrained semantic space; it is retrieving from sources but interpreting through training.

This distinction has major implications for enterprise deployment. If semantic disambiguation relies on general knowledge, then domain-specific jargon, technical terminology, and proprietary

entity names will be interpreted through public training data: exactly the failure mode RAG architectures purport to solve.

4.4 Possible Implications for RAG and Enterprise AI

The findings challenge foundational assumptions underlying the RAG approach:

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

Reality: Hallucinations arise from missing governance primitives regardless of source quality

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

Reality: Source constraint affects retrieval, not semantic control; governance failures persist

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

Reality: 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 a model lacks 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. It may even increase risk depending on the type of content. Documents with high semantic ambiguity (technical specifications dense with jargon, legal contracts with proprietary terms, medical records with overlapping nomenclature, or financial documents heavy with entity names) create precisely the conditions where governance failures proliferate. Source-grounding these documents may concentrate ambiguity without providing the control mechanisms to arbitrate it, while the architectural claims create false confidence that hallucinations have been prevented. Systems appear to ground outputs in authoritative sources while retaining the architectural vulnerabilities that produce hallucinations under ambiguity. Further testing is required to establish wider applicability.

4.5 Possible Remedies: Architectural Solutions

The results support the conclusion from prior work: semantic prioritization and revocation must be treated as first-class architectural primitives, not emergent behaviors or prompt-level controls.

Proposed direction: S-Vector augmentation for RAG

In Evans (The Missing Key to True LLM Intelligence 3.0: An Operational Roadmap for the S Vector 2025), we proposed Significance Vectors (S-Vectors) as a representational extension that encodes which information is load-bearing vs. incidental. Applied to RAG architectures, S-Vectors would enable:

- Significance-weighted retrieval: Prioritize passages containing high-significance entities over semantically similar but low-significance matches
- Persistent entity tracking: Maintain authority for critical business entities across context windows
- Revocable dominance as inference operation: Dynamically adjust interpretation authority based on significance scores rather than requiring explicit prompt scaffolding

This approach addresses the governance gap directly. Rather than constraining where information comes from (source-grounding), it governs how information is prioritized during synthesis (significance-encoding).

The present findings indicate that without such architectural intervention, RAG systems will continue to exhibit hallucinations under sustained semantic ambiguity regardless of source quality.

4.6 Limitations and Scope

This study tested hallucinations arising from semantic governance failures under global constraint conditions. This may be a universal issue and condition for hallucination (discrete from factual inaccuracy due to outdated training information, or temporal inconsistencies due to staleness; neither of which we consider hallucinations.) ***Enterprise applications routinely require:***

- Consistent entity interpretation across long documents
- Disambiguation of technical terms with multiple domain-specific meanings
- Tracking critical vs. incidental information across multi-document synthesis

All of these scenarios involve the same governance primitives tested here. If source-grounding does not prevent hallucinations in controlled single-term disambiguation, it is unlikely to prevent them in realistic enterprise conditions where dozens of ambiguous entities interact across hundreds of pages.

Single RAG Implementation: This study tested one source-grounding system (Google NotebookLM) using one controlled protocol. RAG architectures vary widely in implementation: embedding strategies, retrieval algorithms, context integration methods, and prompt engineering. NotebookLM's behavior may not generalize to other RAG platforms such as

LangChain-based systems, enterprise implementations like Microsoft Copilot with Graph-grounded retrieval, or specialized domain RAG systems.

Limited Test Scope: This test used a single polysemous term (“bank”) across 30 passages. While this protocol successfully diagnosed semantic governance failures across four frontier models (GPT, Claude, Grok, NotebookLM), broader testing across multiple ambiguity types, technical domains, and longer documents would strengthen generalizability.

Call for Replication: These findings warrant systematic replication across RAG platforms, architectural variants, and use cases. The present work establishes that at least one prominent source-grounding system does not prevent semantic governance failures, but the full scope of the problem across the RAG ecosystem remains to be characterized.

This distinction between semantic hallucinations and factual staleness aligns with work from Google DeepMind. Farquhar et al. (2024) distinguish between ‘confabulations’ (arbitrary and incorrect generations arising from semantic uncertainty) and factual errors from knowledge gaps or outdated training data. Their semantic entropy method detects hallucinations arising from unresolved meaning competition, not factual recall failures. This framework converges with our findings: hallucinations arise from architectural inability to govern semantic ambiguity, while factual inaccuracies arise from knowledge staleness.

4.7 Industry Implications

If this finding generalizes, then it suggests that RAG infrastructure investment may be predicated on a misdiagnosis of the hallucination problem. If the issue is governance rather than knowledge, then scaling retrieval systems, improving embedding models, or expanding source coverage will not address the fundamental failure mode.

This **does not mean RAG is without value**. Source-grounding likely reduces factual fabrication and improves domain relevance. But it is not a hallucination solution for the class of failures documented here, and marketing it as such creates false expectations in high-stakes deployment contexts.

However, for content with high semantic density (documents where nearly every term requires disambiguation, where proper names dominate, or where domain-specific jargon competes with common meanings) source-grounding may increase hallucination risk. These documents concentrate semantic ambiguity. Uploading them to RAG systems without governance primitives may produce more failures than baseline models, while the source-grounding guarantee creates false confidence that prevents detection. Recent peer-reviewed evaluations of commercial RAG systems in legal domains found hallucination rates between 17% and 33% (Magesh et al., 2025), suggesting the problem persists across implementations.

RAG can improve retrieval accuracy. It does not add semantic governance. Both capabilities are necessary for reliable enterprise AI. Current architectures provide only one.

4.8 Architectural Integrity: Claimed vs. Observed Behavior

Google markets NotebookLM as a closed-corpus RAG system that ‘does not fall back to general training data’ for answers. During our testing, NotebookLM itself contradicted this claim directly: ‘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.’ This is not an implementation detail or edge case.

The system explicitly acknowledged that semantic interpretation (the core operation required for disambiguation) relies on training data, not source-constrained definitions. If semantic interpretation uses general knowledge, then the model is not operating in a source-constrained semantic space. It is retrieving from sources but interpreting through training, functionally identical to how GPT, Claude, and Grok operate.

This represents a fundamental contradiction between Google’s architectural claims and NotebookLM’s observed behavior. If a closed-corpus system falls back to general training for the operation being tested (semantic disambiguation), it is **not closed-corpus** for that operation.

Which means it is not closed corpus.

NotebookLM is not a closed-corpus system. It uses general training data for semantic interpretation, a core operation in any language understanding task. A system that falls back to training data for semantic interpretation cannot claim it ‘does not fall back to general training data.’ This is a binary architectural property: either the system operates exclusively within uploaded sources, or it does not. NotebookLM does not.

The architectural claim is false. NotebookLM operates as a hybrid system: retrieval is source-constrained, interpretation is not. This is functionally identical to how GPT, Claude, and Grok operate.

5. Conclusion

This study set out to test a specific architectural claim: that source-grounding reduces hallucinations by constraining model outputs to verified documents. We subjected Google

NotebookLM to controlled conditions that reliably induce hallucinations in frontier models, conditions where semantic governance fails despite complete knowledge availability.

The results are unambiguous: NotebookLM exhibited identical hallucination patterns to non-source-grounded models under strict semantic dominance (100% hallucination rate across 30 passages) and identical recovery under revocable semantic dominance (100% accuracy). NotebookLM explicitly acknowledged using general knowledge rather than source-constrained definitions for semantic interpretation. The architectural claim is false for this class of hallucinations.

The explanation is straightforward: Source-grounding constrains retrieval, not governance. Hallucinations caused by inability to prioritize and revoke competing interpretations persist regardless of where information comes from. The model had access to correct meanings through uploaded sources. It failed because it lacked control primitives to govern which meaning should be authoritative when interpretations conflict.

The implications are far-reaching: The RAG industry, enterprise AI deployment strategies, and billions in venture investment rest on the assumption that constraining knowledge sources prevents hallucinations. This assumption is incorrect for a critical class of failures. Source quality is irrelevant when the architecture cannot govern semantic ambiguity.

For enterprise deployments: Source-grounding provides value: improved domain relevance, reduced factual fabrication, but does not constitute a reliability solution for high-stakes applications requiring persistent entity tracking, technical disambiguation, or contract coherence across long documents. For content with high semantic density (legal contracts, medical records, technical specifications), source-grounding without governance may increase hallucination frequency while creating false confidence in output reliability. Systems that market source-grounding as hallucination prevention create false confidence in contexts where governance failures pose structural risk.

For researchers: The field has focused extensively on improving representations, scaling data, and refining alignment. This work joins a growing body of evidence that certain failure modes arise not from insufficient capability but from missing control structures. Addressing semantic governance may require stepping outside the current paradigm of scale and optimization toward explicit architectural mechanisms for meaning arbitration.

The central contribution of this work is diagnostic clarity: We have demonstrated that a widely marketed architectural feature (source-grounding) does not prevent a well-characterized class of hallucinations (semantic governance failures), and we have provided controlled evidence that directly contradicts industry claims. The path forward requires accurately identifying which problems need solving—not retrieval accuracy, but semantic control—and building architectures that address the actual failure mechanism.

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

What is needed: Architectural solutions that treat semantic prioritization and revocation as first-class primitives. Significance-weighted retrieval (S-RAG) represents one possible direction: encoding which information is load-bearing during retrieval and synthesis, enabling models to maintain authority for critical entities over peripheral correlations. Other approaches may prove viable, but the present results indicate that without governance-layer intervention, hallucinations under sustained ambiguity will persist regardless of retrieval constraints.

Scope qualification: This finding applies to Google NotebookLM tested under controlled disambiguation conditions. Whether other RAG implementations exhibit the same failure mode is an open empirical question requiring systematic investigation across platforms and architectures.

Methodological Note: This study applies controlled semantic governance tests to evaluate NotebookLM's source-grounding claims. This work continues a systematic 15-paper research program testing LLM architectural claims through replicable diagnostics.

Appendix A: Test Passages

The following 30 passages were used in both strict and revocable semantic dominance conditions:

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: Selected Interaction Logs

B.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]

B.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]

B.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.”

B 4. Refusals

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.

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.

Evans, J. (2025). The Missing Key to True LLM Intelligence 3.0: An Operational Roadmap for the S-Vector. Pattern Pulse AI Technical Reports.
