Retrieval-Augmented Generation, Grounding, Faithfulness, and Reality Drift

Why AI Systems Can Access Information Yet Still Lose Meaning

AI Systems and Reality Drift - Note #8 v1

A. Jacobs - Semantic Fidelity Project

The Basic Pattern

Modern AI systems increasingly rely on external information sources. Rather than generating responses only from training data, many systems retrieve documents, databases, web pages, reports, and knowledge repositories before producing an answer. This approach is commonly known as Retrieval-Augmented Generation, or RAG.

The goal is straightforward. If a model can access relevant information directly, it should produce more accurate and reliable outputs. At first, this often appears successful. Answers become more current, responses include citations, systems reference specific documents, and users gain greater confidence in the result.

Over time, however, a familiar pattern can emerge. The correct information may be retrieved while the final answer still changes its meaning. Sources may be cited but misread. Relevant documents may be available but ignored. Context may be compressed, flattened, or lost during generation.

Different communities describe this challenge using different terminology. It appears in discussions of retrieval-augmented generation, grounded generation, faithfulness, attribution, contextual relevance, source fidelity, citation accuracy, and retrieval failure. These are not identical problems, but they often point toward the same structural risk: the system may successfully retrieve information while failing to preserve what that information means.

Access Is Not Understanding

Retrieval systems are often discussed as if access to information automatically improves understanding. But retrieval and interpretation are different processes. A system can identify a relevant source while still distorting, omitting, summarizing, or recontextualizing that source during generation.

This distinction becomes more important as AI systems scale. The challenge is no longer simply finding information. The challenge is preserving meaning as information moves through search, ranking, context selection, compression, and generation.

A citation can make an answer appear grounded even when the relationship between source and claim has weakened. Access to evidence matters, but access alone does not guarantee fidelity.

How Retrieval Drift Emerges

Stage 1 — Retrieval

The system identifies documents, sources, or records relevant to a user's query. At this stage, the information remains available.

Stage 2 — Representation

Retrieved information is converted into embeddings, rankings, context windows, summaries, or compressed representations. The original material becomes mediated through the retrieval system.

Stage 3 — Transformation

The model interprets, summarizes, reorganizes, and generates responses from the retrieved information. Meaning now depends on how well the system preserves the relationship between source material and output.

Stage 4 — Drift

The output remains coherent, sources remain available, and citations may appear correct. Yet meaning gradually diverges from the original information. The system preserves access while losing fidelity.

The absence of failure is not proof of alignment. It may only mean the retrieval system has stopped detecting the gap.

Semantic Fidelity and Constraint Collapse

At this point, the retrieval system has not necessarily failed on its own terms. What weakens is the structural relationship between the source material and the generated response.

An answer may preserve the surface appearance of grounding while losing the conditions, context, and relationships that made the source meaningful. This is where retrieval drift becomes a semantic fidelity problem. The representation remains coherent, but its connection to the underlying source has degraded.

Retrieval drift also involves constraint collapse. Retrieved evidence is supposed to constrain generation by keeping outputs tied to source material. But when the model becomes more responsive to summaries, rankings, context compression, or its own internal patterns, that constraint can weaken. The system remains sourced, but the source loses some of its corrective force.

The Representation Stack

Retrieval drift becomes more serious when information moves through a wider representation stack. A document becomes an embedding, the embedding becomes a ranking, the ranking becomes context, and the context becomes a generated answer. Each layer makes retrieval easier to scale while increasing the distance from the original material.

This is how representational failure can become systemic. A source may begin as evidence, but after it moves through retrieval, summarization, generation, citation, and user trust, it can harden into a claim about reality. The further the representation travels from the source it was meant to preserve, the easier it becomes for access to be mistaken for understanding.

Examples Across Systems

Retrieval Failure: A relevant document may exist but never enter the context used to produce the answer.

Grounding Failure: The correct source may be retrieved while the final answer is only partially supported by the evidence.

Faithfulness Failure: A generated response may subtly distort, omit, or reinterpret the meaning of the source material.

Citation Failure: A source may be cited even when the associated claim is not actually supported by that source.

Context Loss: Important information may enter the context window but fail to survive into the final answer.

Conclusion

Retrieval-augmented generation is not only an information access problem. It is a representational fidelity problem. Documents, embeddings, rankings, context windows, citations, and generated answers all function as representations of source material.

The deeper failure occurs when information remains available while meaning weakens during transformation. A system can appear grounded because it retrieved a source, but still lose contact with what the source actually says.

More retrieval does not solve this problem if meaning is not preserved across the system. The deeper challenge is maintaining fidelity between source material and generated meaning, so that access to information does not become a substitute for understanding.

Related Phrases and Concepts

- retrieval-augmented generation
- RAG
- grounded generation
- faithfulness
- attribution
- source fidelity
- citation accuracy
- retrieval failure
- contextual relevance
- context retention
- context loss
- semantic preservation

Semantic Fidelity Project Resources

- Research Library (GitHub)
- Articles & Essays (Substack)
- Semantic Fidelity Glossary
- LLM Failure Modes and Semantic Fidelity