Why is retrieval correct but interpretation still wrong?

Why Accurate Retrieval Fails to Preserve Meaning and Intent in LLM Systems

Semantic Fidelity Lab | Failure Modes 06

A. Jacobs — Part of the Reality Drift Framework

Core Claim

Systems can retrieve the correct information while still producing incorrect or misleading outputs, because retrieval preserves data while interpretation transforms meaning.

Mechanism

Retrieval systems identify and return relevant documents or data based on similarity, ranking, or matching criteria. This step often performs well, especially in RAG pipelines where embeddings and search mechanisms surface appropriate context. However, the retrieved information must then be interpreted, summarized, or integrated into a response. During this transformation, the model re-encodes the information into new language, introducing compression and reinterpretation. This process preserves structure but not necessarily meaning, creating a gap between what is retrieved and what is ultimately expressed.

Failure Mode

Interpretation failure — retrieval is accurate, but meaning is altered during transformation, a direct instance of **reality drift**. The system successfully accesses relevant information but fails to preserve the intent or significance of that information when generating outputs. Structure is maintained, but meaning shifts.

Example

A RAG system retrieves the correct document containing an answer. The model then generates a response based on that document, but misinterprets a key detail or relationship. The output references the right source and appears grounded, yet the conclusion is incorrect.

System Implications

This reflects a breakdown in **semantic fidelity** at the interface between retrieval and generation. Systems are often evaluated on retrieval accuracy, but not on whether meaning is preserved during interpretation. As retrieval improves, this failure becomes less visible but more consequential, since

outputs appear well-grounded while still being wrong. This creates a class of errors that are harder to detect because they originate in transformation rather than access.

Keywords: *RAG retrieval correct but wrong answer, retrieval vs interpretation AI, semantic search errors, RAG failure modes, embedding retrieval vs meaning, AI misinterpretation, factual grounding vs correctness, LLM reasoning errors, semantic fidelity, reality drift*

Core Framework and Sources

- Substack (Articles)
- GitHub (Full Library)
- DOI (Research Paper)
- Glossary & Definition