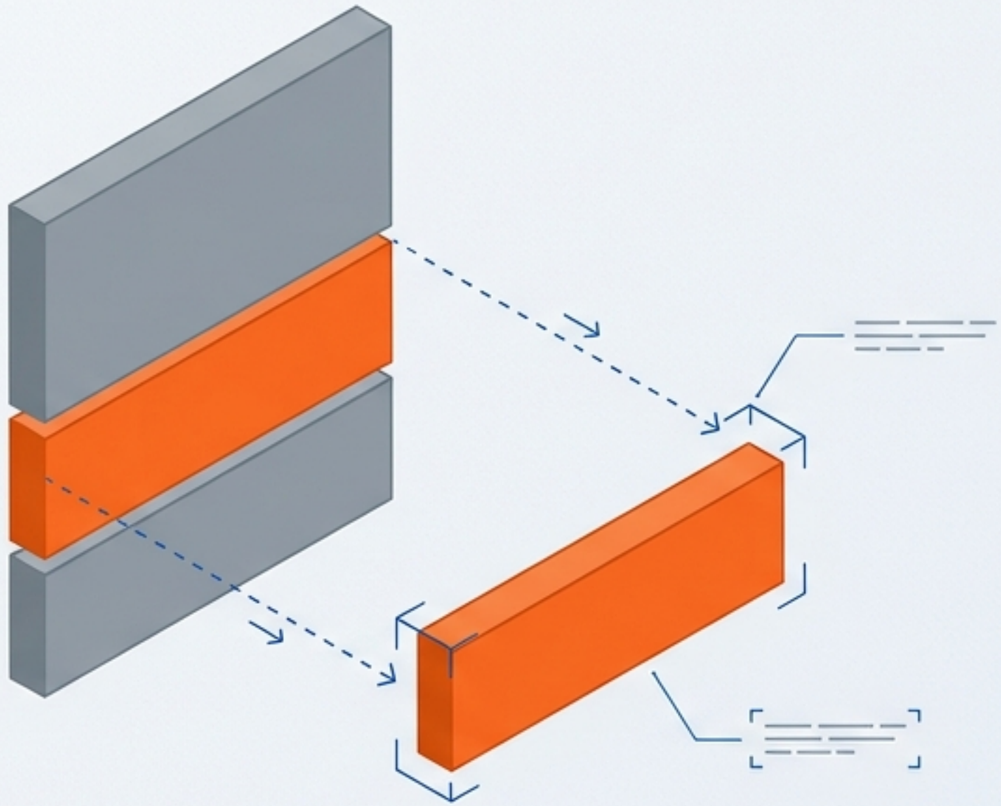
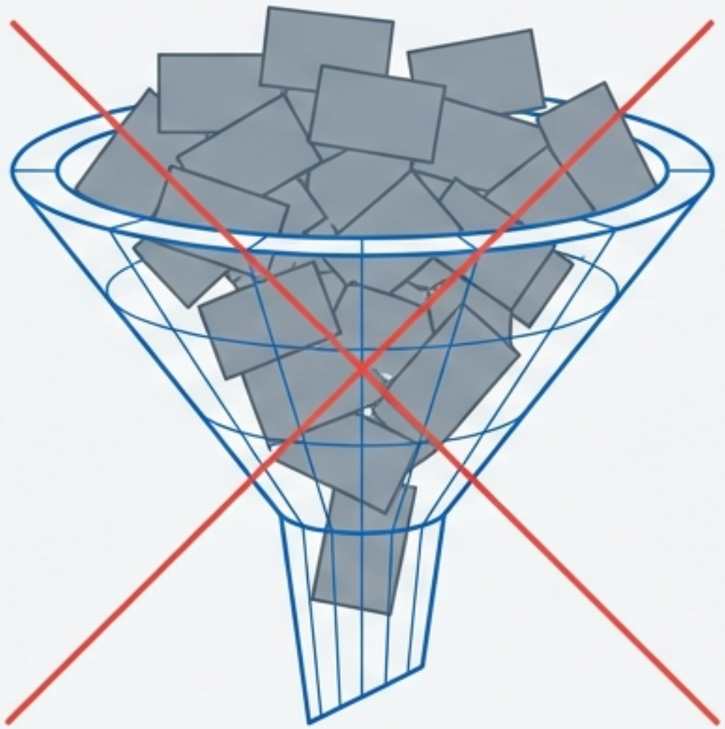
Anatomy of Evidence

Context Interventions and the Load-Bearing Sentence in RAG Systems

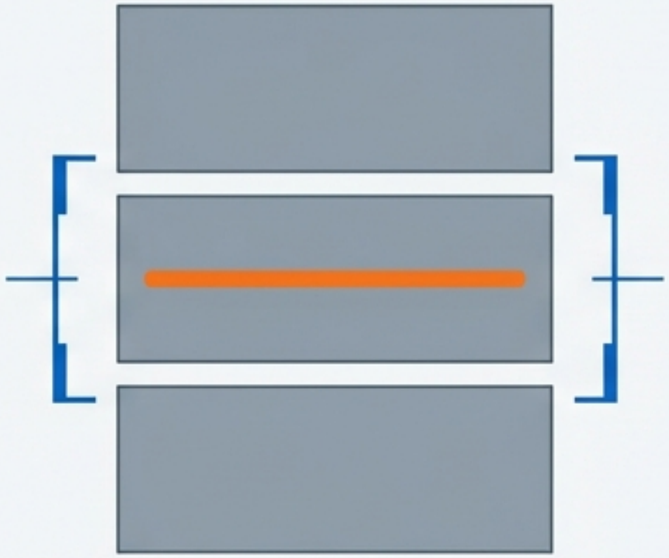


NotebookLM

The Fallacy of “More Context”



The Industry Assumption: Volume

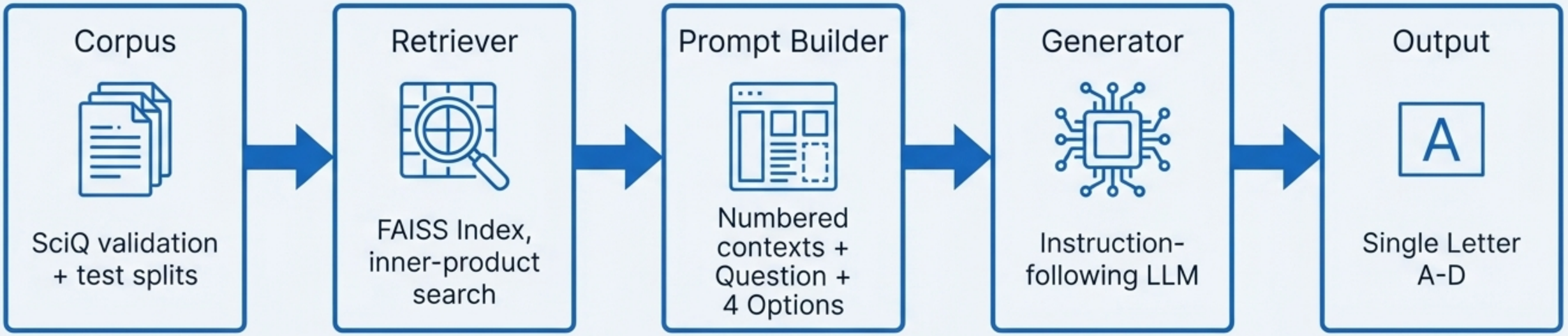


The Reality: Precision Evidence

Large language models hallucinate on specialized knowledge. **Retrieval-Augmented Generation (RAG)** mitigates this by retrieving documents at inference time. However, the standard assumption that **more context equals better answers** is flawed. RAG accuracy is determined exclusively by the **precise quality of the context**—specifically, whether it contains the **exact answer-bearing evidence**.

NotebookLM

The Diagnostic Pipeline



NotebookLM

Laboratory Parameters & Baseline Calibration

To isolate the variable of context quality, we established a rigid set of laboratory controls.

Using an exhaustive grid search ablation, smaller semantic chunks paired with exact statistical testing formed our baseline.

[DATASET]

SciQ.

1,000 science multiple-choice questions.

4 choices per question (3 distractors, 1 correct).

[OPTIMAL RAG CONFIGURATION]

Chunk size 256 tokens, 64-token overlap.

k=3 chunks retrieved.

(Smaller chunks proved to yield sharper semantic alignment during grid search).

[VARIANCE CONTROL]

3 random seeds used to shuffle MCQ option order.

Avoids positional bias in small models.

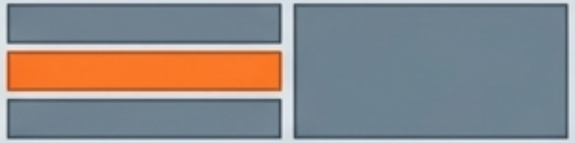
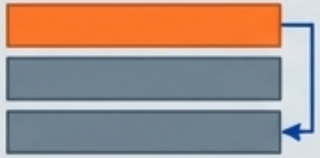
[STATISTICAL RIGOR]

95% confidence intervals.

McNemar's exact test for paired binary responses.

NotebookLM

The Diagnostic Matrix: Surgical Interventions

Setting	Visual Anatomy	The Surgery	Hypothesis Tested
RAG Baseline Fira Code		Standard top-3 chunks.	Standard retrieval accuracy.
Setting A Fira Code		Delete exactly the sentences containing the gold answer. Fira Code	Is the specific evidence sentence load-bearing? Inter
Setting B Fira Code		Keep ONLY the sentences with the gold answer. Fira Code	Is the single evidence sentence sufficient alone? Inter
Setting C (Distractor) Fira Code		Add one randomly selected irrelevant corpus paragraph. Fira Code	Does irrelevant noise degrade a clean signal? Inter
Setting D (Shuffle) Fira Code		Randomly permute chunk order. Fira Code	Does position matter? Inter
Oracle Fira Code		Feed perfectly matching gold support passage. Fira Code	What is the theoretical upper bound? Inter

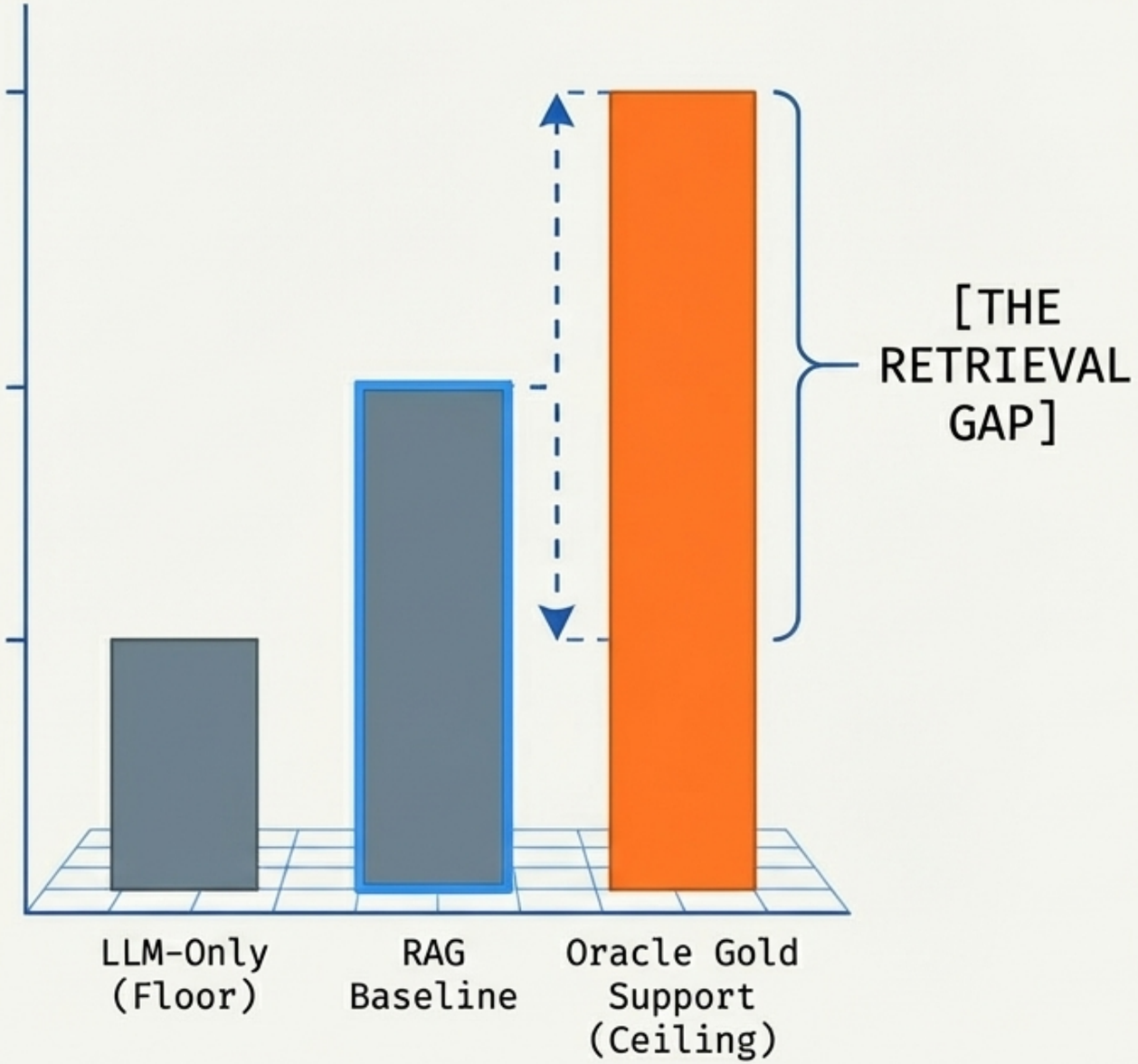
NotebookLM

Establishing Tablishing the Bounds of Performance

Feeding the perfect, dataset-author-confirmed document to a small model (Oracle) sets the ceiling for performance.

The RAG Baseline successfully outperforms the LLM-only floor, proving retrieval works—but it falls short of the Oracle ceiling.

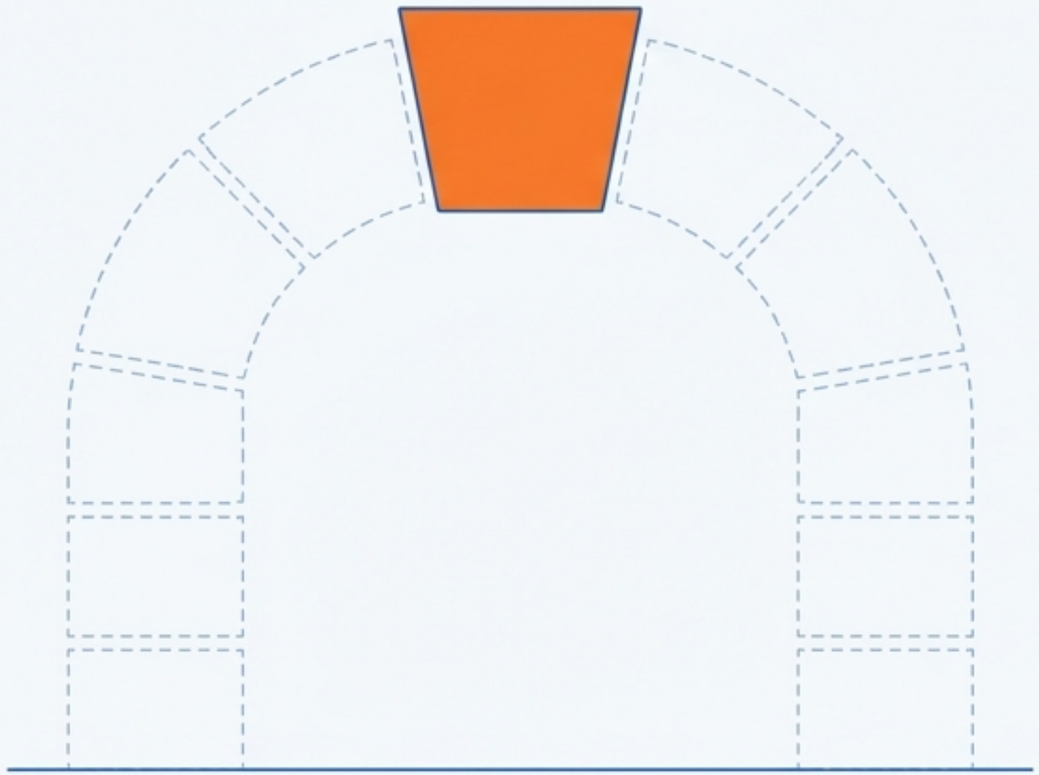
This gap quantifies exactly what the retriever fails to deliver: perfect semantic evidence every time.



NotebookLM

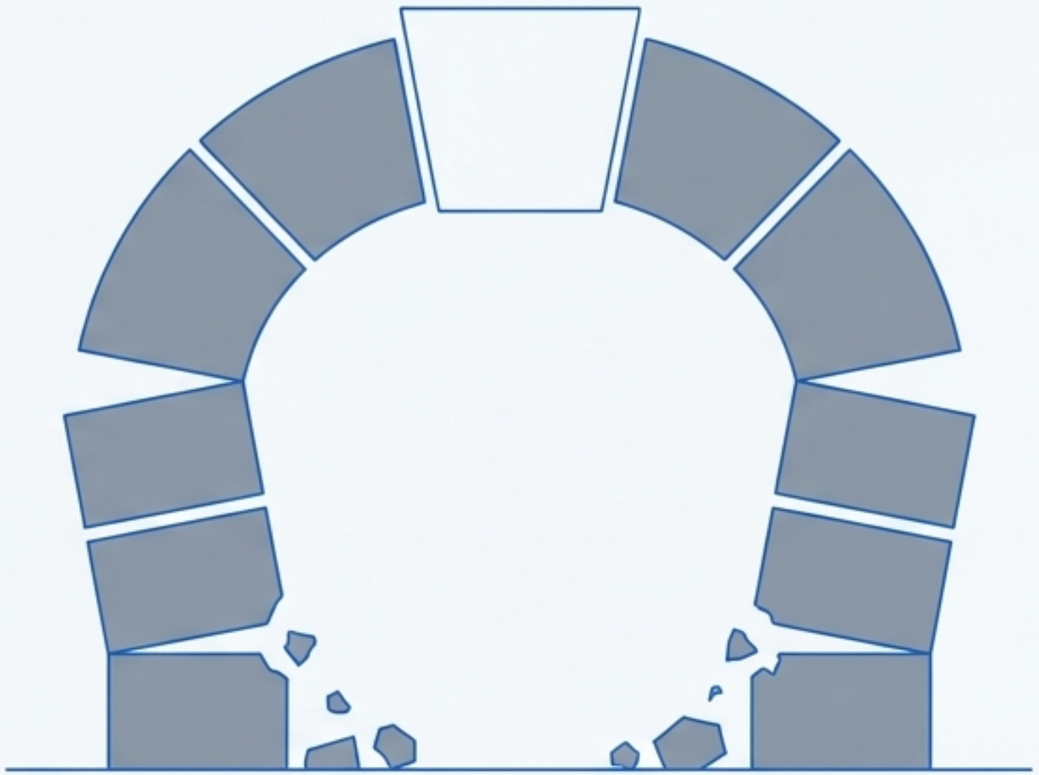
The Load-Bearing Evidence Sentence

Setting B: Compression



Performance Maintained. You can strip away the surrounding context and preserve accuracy. The keystone is sufficient.

Setting A: Removal



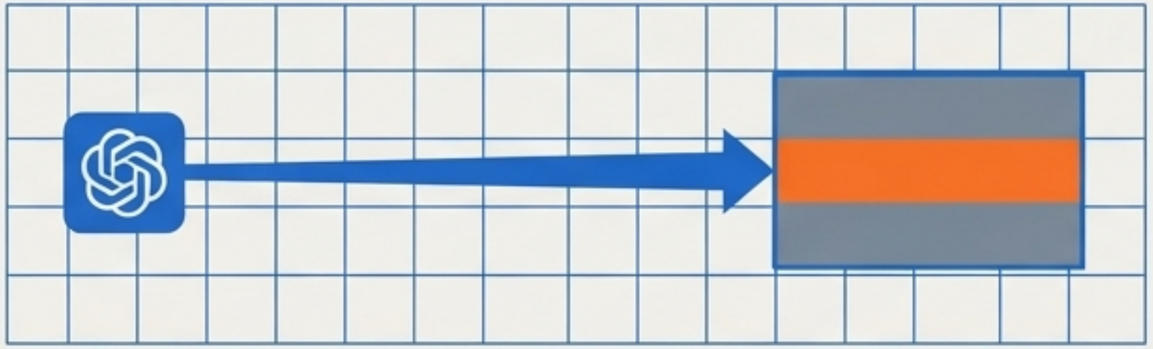
Performance Drops. Provide full context but silently delete the one answer-bearing sentence, and the system collapses.

The model is highly sensitive to the absence of critical evidence, but resilient to extreme compression.

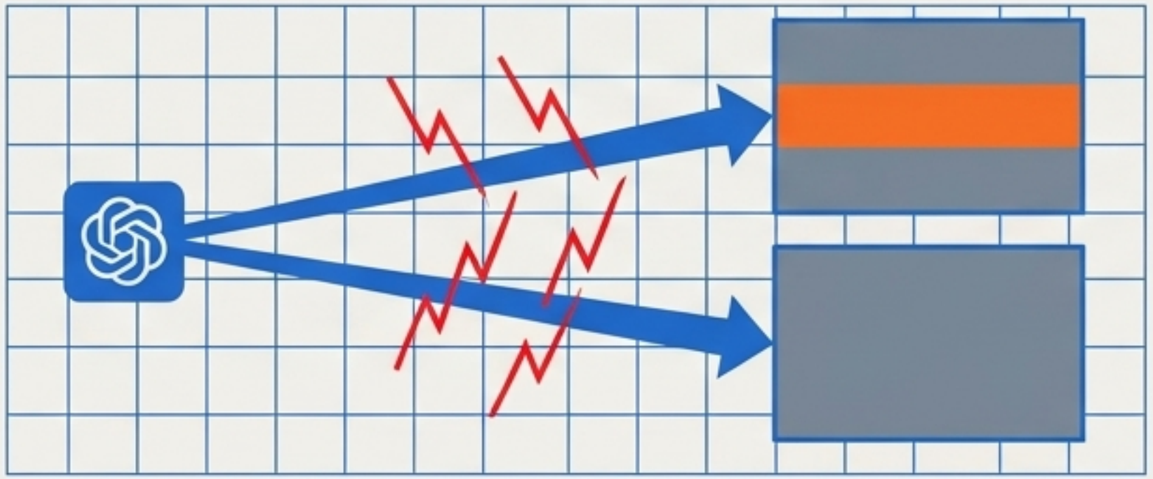
NotebookLM

The Penalty of Distraction

Clean k=1



Setting C: k=1 + Random Distractor



Adding irrelevant context is not free.

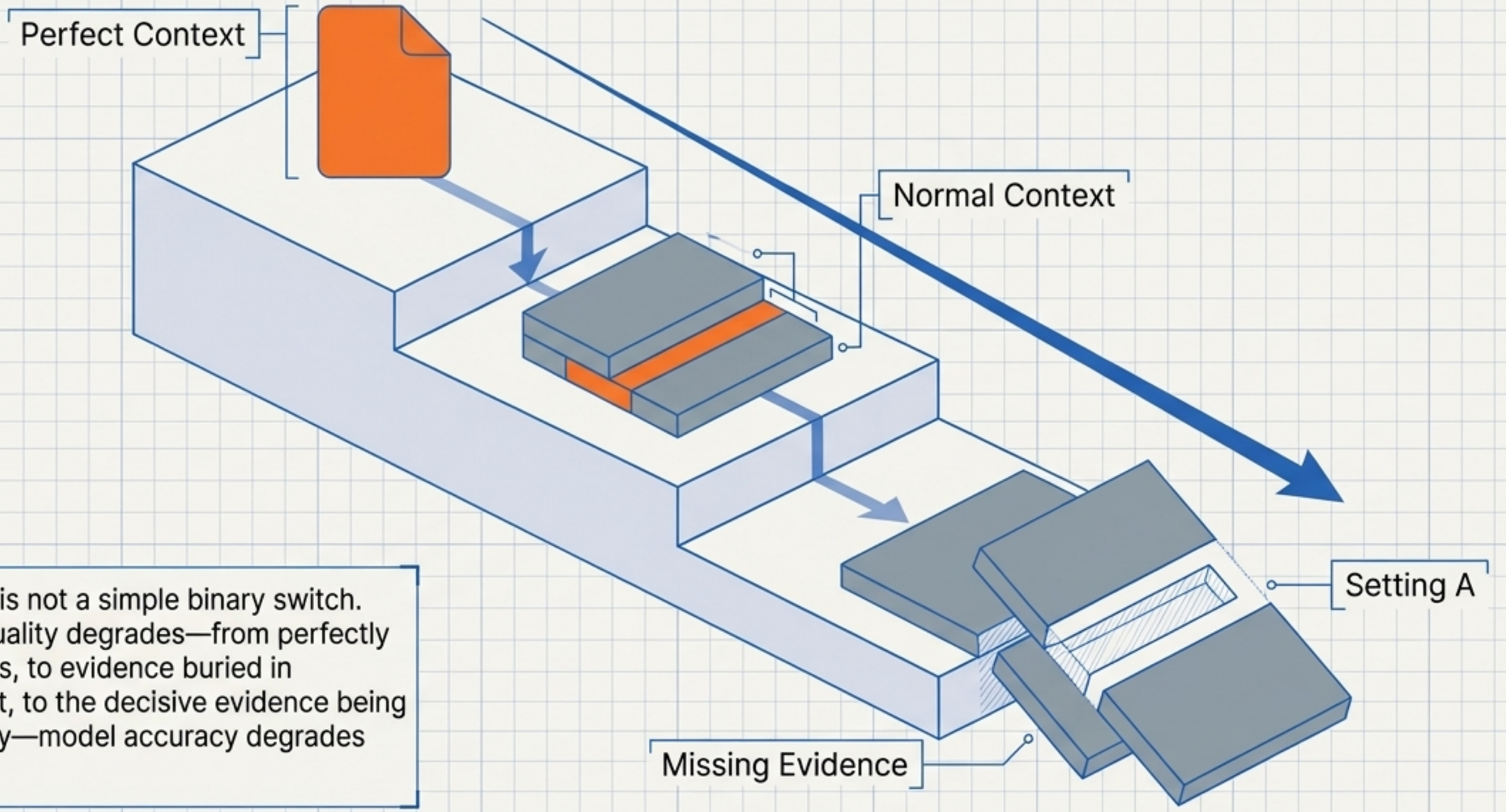
When the model is provided with a single clean chunk, accuracy is high.

Injecting just one randomly selected support paragraph that does not contain the answer degrades performance. Irrelevant noise splits the model's attention and pollutes clean evidence.

Note: Shuffling context order (Setting D) had almost zero effect. The model reads content semantics, not position.

NotebookLM

The Document Quality Step-Function



Context quality is not a simple binary switch. As document quality degrades—from perfectly isolated answers, to evidence buried in surrounding text, to the decisive evidence being removed entirely—model accuracy degrades monotonically.

NotebookLM

Clinical Readouts: Case Studies in Focus

Case 1: Proving the Sentence Matters

Model flips from correct to wrong when the answer-bearing sentence is deleted. Decisive evidence is lexical.

Case 2: Proving Context is Redundant

Model remains correct even when all surrounding context is stripped. The rest of the chunk was noise.

Case 3: Proving Distraction Kills

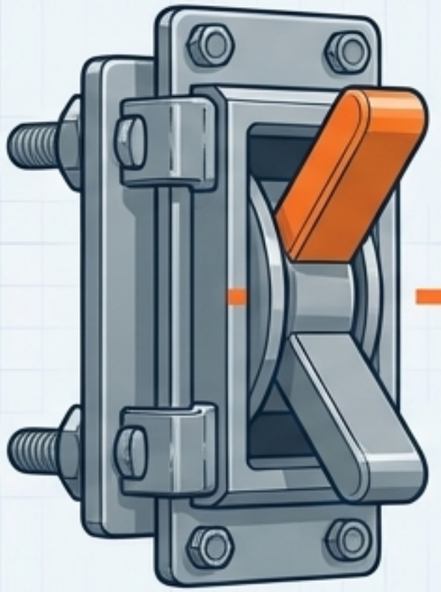
Adding a single irrelevant paragraph from the corpus shifts the prediction. Attention is split.

NotebookLM

The Rules of Context Anatomy

RAG accuracy is not a volume dial where “more text = better answers”. Instead, it is governed by two independent controls.

1. EVIDENCE TOGGLE



Is the specific load-bearing evidence present? If false, the system fails regardless of context length.

2. NOISE DIAL (DISTRACTORS)

How many distractors surround the evidence? Each additional irrelevant chunk acts as a penalty, splitting attention and degrading the signal.

NotebookLM

Laboratory Limitations & Future Protocols

Scale Sensitivity

`flan-t5-small` is an `80M` parameter model.

Future work must replicate interventions on **7B+ instruction models** to verify if massive architectures exhibit the same **distraction penalties**.

Dataset Cleanliness

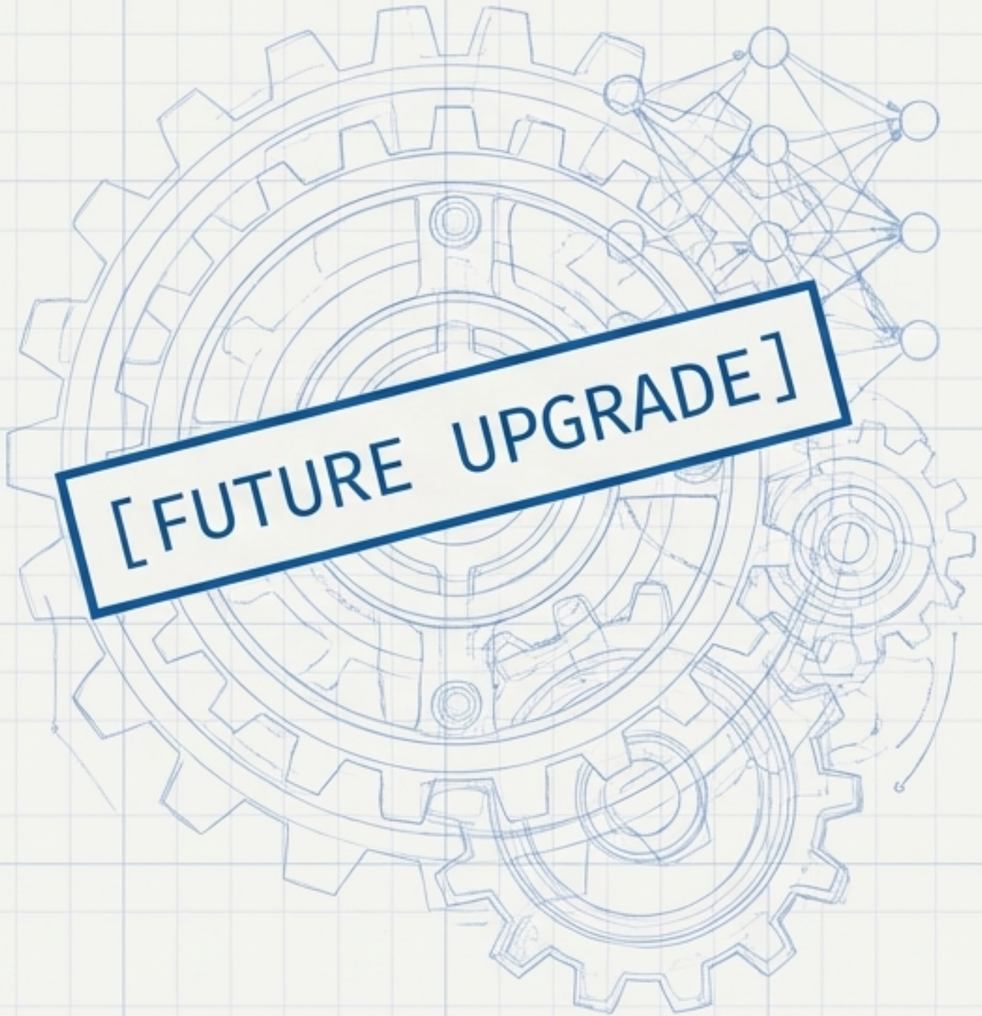
`SciQ` provides structured, clean gold-support coverage.

Real-world enterprise retrieval operates on significantly **messier, unstructured data**.

Lexical vs. Semantic Isolation

The current study isolated sentences via exact **exact string matches**.

Future pipelines should use **LLMs** to isolate **logical reasoning chains** rather than just lexical phrases.



NotebookLM

If you want to improve a RAG system, don’t just add more retrieved documents.

Make sure the critical sentence is actually there.

NotebookLM