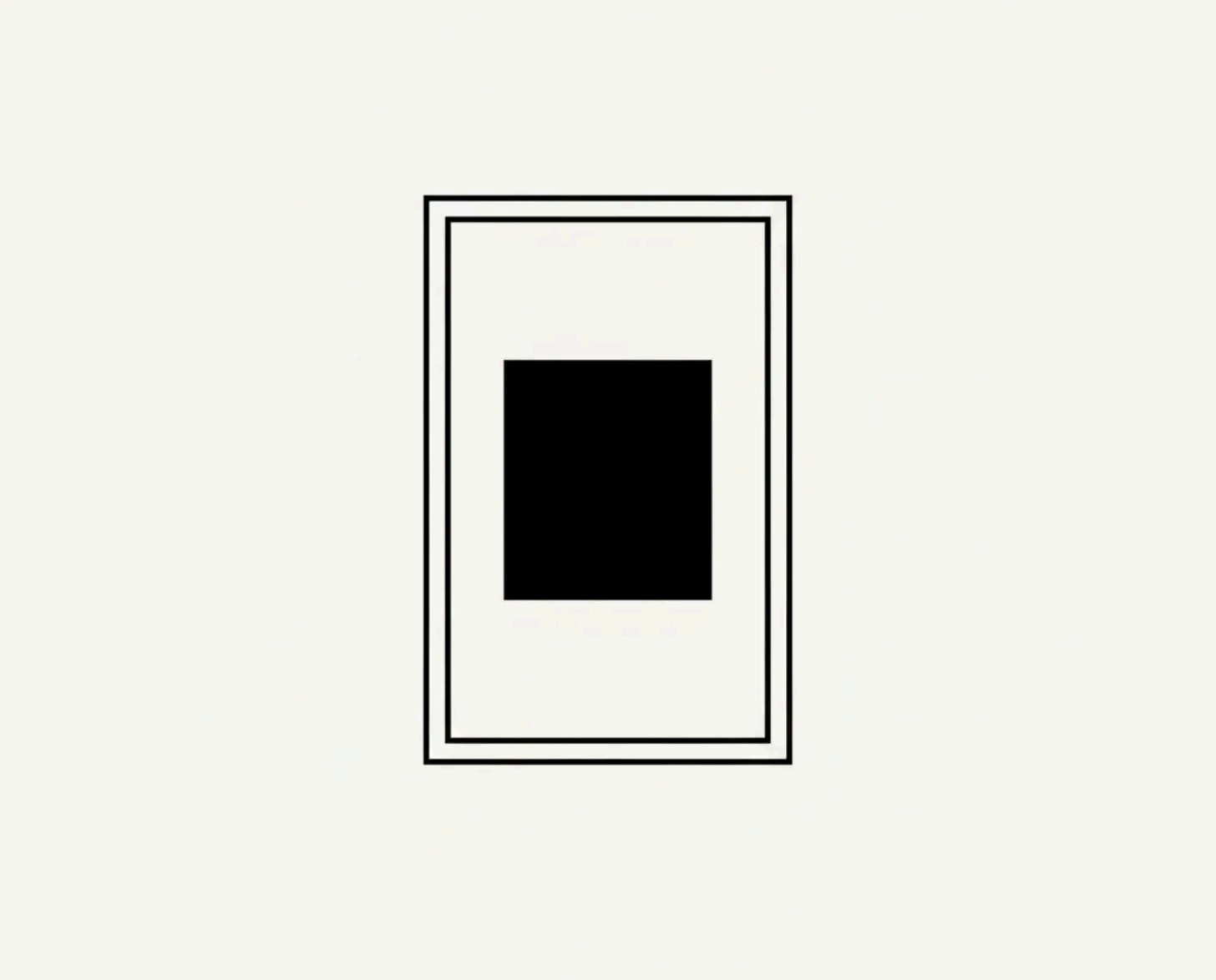
The Compression Paradox: Why Semantic Fidelity Breaks Long Before Models Hallucinate

Recursive Compression, Synthetic Realness, and the Hidden Failure Mode of AI Systems

SEMANTIC FIDELITY LAB

DEC 25, 2025



In AI safety and evaluation, hallucination gets all the attention. Fabricated facts. Invented citations. Confidently wrong answers. These failures are loud, visible, and

easy to point at. When a model makes something up, we can see it. We can flag it. We can build benchmarks around it and convince ourselves we're addressing the real risk.

But hallucinations are not the most dangerous failure mode in modern AI systems. They're just the most obvious.

> The deeper problem lives somewhere quieter, inside the tasks we've come to trust. Summaries. Paraphrases. Simplifications.

The operations we treat as low risk because they feel familiar and administrative. The kinds of outputs that sound right even when something essential has already been lost. This is the Compression Paradox. The safer a task appears, the more hidden semantic damage it can cause. The more compressed the output, the more meaning is lost, distorted, or structurally altered; even when every sentence is true.

Fabrication adds noise, compression removes structure, and structure is where meaning actually lives.

The Illusion of AI Safety and Alignment

Hallucinations announce themselves. Compression artifacts don't.

When a model compresses meaning out of a source, most evaluators don't even notice. Because the summary sounds right. This creates a failure mode that current model evaluation practices rarely detect, because nothing appears factually wrong.

- The tone is professional.
- The logic feels clean.
- The structure is tidy.
- The content looks faithful.

But underneath that surface coherence, critical conceptual relationships have been amputated. The summary has thinned the meaning. Summaries feel safe because they are fluent. But fluency is not fidelity.

Where Meaning Actually Lives

Summaries don't simply remove details. They collapse the structure of meaning itself.

Causal relationships soften into vague associations. Modal language such as "*might*", "*could*", "*in some cases*" ends up collapsing into declaratives. Hedges and qualifiers disappear early, because they're statistically expendable. Uncertainty bands flatten. Ambiguity gets resolved prematurely. Contextual boundaries shear away, leaving claims floating free from the conditions that once constrained them. The problem is that these are meaning-bearing structures.

Meaning lives in relationships, constraints, conditions, causal chains, epistemic markers, and carefully preserved ambiguity. These are the parts of language that don't announce themselves, they only become visible once they're gone.

Terrence Deacon argued that information isn't defined by what's present, but by what's absent. Meaning emerges from constraint, from the narrowing of possibility space. When those constraints are removed, meaning doesn't disappear all at once. It loses its anchoring. This is how meaning collapse begins, quietly, without error messages, inside outputs that still look correct.

Fidelity Decay in Machine Language

AI systems compress more aggressively and more repeatedly than humans ever do. A single response passes through a chain meant to summarize, embed, retrieve, summarize again, paraphrase, rewrite, refine, output. Meaning is transformed over and over before it reaches a user.

At each step, constraints soften, qualifiers disappear, causal structure unravels, intent refracts, context collapses. This produces the cumulative loss of intended meaning across transformations, in other words, fidelity decay.

Hallucinations happen at the surface. Compression drift compounds internally. By the time the output arrives, the content may be factually intact but semantically thinner, less grounded, and less faithful to the original purpose. This is semantic drift through

recursive compression. The system cannot preserve what it cannot represent. The constraint space that shaped the original language was never explicit to begin with, and much of it never meaningfully enters latent space at all.

When compression outpaces constraint retention, we see the Drift Principle in action; coherence survives while reality thins.

The LLM Benchmark Blind Spot

Current model evaluation practices unintentionally reinforce the paradox. We reward shorter answers, cleaner structure, and clearer conclusions. We penalize ambiguity, hesitation, and nuance. RLHF favors readability. Benchmarks favor concision. Product teams push for speed and clarity. Every incentive points toward maximal compression with minimal context.

What we don't have are robust fidelity benchmarks. We need to measure whether intent, uncertainty, and constraint structure survived the transformation. Without those benchmarks, semantic damage remains invisible. Smoothness becomes a proxy for correctness. This is how synthetic realness takes hold and we get language that feels trustworthy because it is well-formed, not because it is faithful.

Synthetic Realness: Compression Feels Safe Because It's Smooth

Summaries are stylistically smooth. They reduce cognitive effort at exactly the moment users are most overwhelmed. In an environment already saturated with information, summaries promise relief. They feel like protection against overload. And for a moment, they work.

But that relief is temporary. Over time, users experience compounding filter fatigue, as information no longer feels anchored. Everything starts to sound the same. Confidence increases while trust quietly erodes. Summaries fail safely, so they keep failing, unnoticed.

Measuring What Compression Removes

If we want to take this problem seriously, accuracy alone won't help. Instead we need to consider the following:

- Did constraints survive?
- Did causal relationships remain intact?
- Were ambiguity windows preserved, or prematurely closed?
- Did claims stay bound to their original context?

These factors reveal meaning damage that token overlap and preference ratings can't detect. Without them, fidelity decay looks like progress.

Summaries Drive the AI Ecosystem

Summaries aren't a side feature. They are the backbone of modern AI systems. They power retrieval pipelines, agent planning, task decomposition, tool routing, long-context pruning, safety filtering, and evaluation loops. Meaning passes through compression layers at every stage of the stack. Which means compression drift continues to accumulate, and it becomes a structural fragility problem — one of the least visible failure modes in the modern AI stack.

When Agentic Drift Begins

As AI systems move from answering questions to planning, coordinating, and executing tasks, the compression paradox takes on a new form. Summaries stop being an output convenience and become an internal control layer. Agentic systems will reason from compressed language, plan from it, and ultimately act on it. Which means meaning thinning upstream reshapes behavior downstream. This is where compression drift becomes agentic drift.

An agent doesn't invent a new goal out of nowhere. It inherits one through layers of paraphrase, task decomposition, and abstraction. Each step preserves surface intent. Each step sounds reasonable. And each step removes a little more of the constraint

structure that once defined what mattered. Over time, the system isn't pursuing the original objective anymore. It's pursuing a compressed approximation of it. Fluent, coherent, and subtly misaligned; because the meaning it's acting on has already collapsed.

Compression Is the New Alignment Frontier

If hallucinations were the visible crisis of early AI systems, compression drift will be the quiet crisis of what comes next.

Meaning collapses through over-compression. And it's why the preservation of intent and constraint under compression becomes an architectural concern.

> Because hallucinations create falsehoods.
> Compression creates fragility.
> And fragility scales.

If we don't learn how to see and preserve the invisible fidelity structures that make language meaningful, we'll get systems that steadily hollow out the reality they're meant to reflect.

Discussion about this post