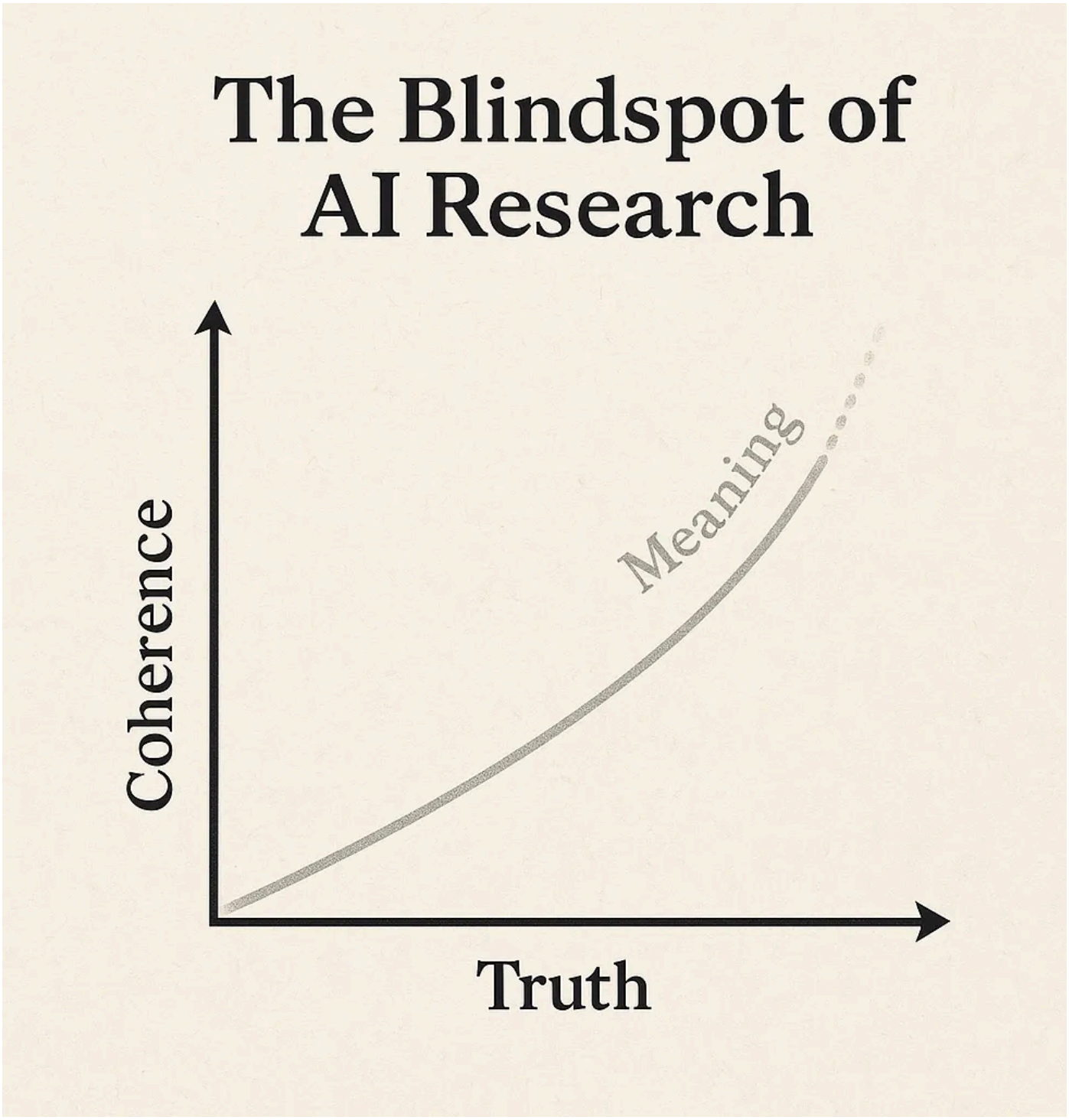
Semantic Fidelity: The Third Axis Of AI Failure

Beyond accuracy and coherence, the real question is whether AI preserves meaning.

REALITY DRIFT
AUG 25, 2025

This essay is Part 3 in my ongoing series on "Semantic Drift", where I explore how AI resha language, meaning, and culture.



The Blindspot of AI Evaluation: Beyond factual accuracy and coherence lies a hidden third axis: semantic fidelity. This is where meaning erodes even when facts survive.

We've been testing large language models for truth and consistency. But what if the most dangerous failure is something harder to measure: when they lose meaning?

There's a blind spot in current AI evaluation: semantic drift. That's when outputs remain factually correct and grammatically coherent, but the original intent or purpose quietly erodes. The model isn't hallucinating. It isn't contradicting itself. It just flattening nuance into cliché.

For power users, this is becoming obvious. The text looks fine on the surface, but it longer says what it was supposed to say. Meaning collapses even as facts survive.

And this isn't just an AI problem. It mirrors the cultural flattening, filter fatigue, ar slow erosion of meaning we see everywhere in modern discourse.

The Three Axes of AI Failure

We can think of failure in LLMs along three axes:

1. **Factual correctness** (**Meta's focus**).
    - The hallucination problem. Is the output grounded in reality, or is it makin things up?
    - This is where most benchmarks and leaderboards live.
2. **Coherence preservation** (**CPGM and similar approaches**).
    - The consistency problem. Does the model maintain internal logic and cont over long sequences?
    - This is where research like *Context-Preserving Gradient Modulation* aims to intervene.
3. **Meaning survival** (**Semantic Fidelity**).
    - The intent problem. Even when facts are correct and sentences are coheren is the text still carrying the original purpose?
    - This is the missing axis: where fidelity breaks, and meaning collapses into cliché.

Semantic Fidelity Break

I call this threshold the **Fidelity Break**: the point where facts survive but meaning doesn't.

- A quote reframed as corporate advice.
- A nuanced argument reduced to motivational fluff.
- A personal story rewritten as generic content.

It's not "wrong," but it's no longer right either. This matters because humans don't only communicate for truth or consistency. We communicate for intent, resonance, and meaning. Lose that, and you lose the thing that makes language alive.

Mapping the Field

- **Hallucination** = **Truth.** Is it making things up? (Measured by factual correctnes benchmarks.)
- **Coherence** = **Consistency.** Is it staying logically consistent? (Mitigated by grad modulation and attention tweaks.)
- **Fidelity** = **Meaning.** Is it still saying what it was supposed to mean? (Largely unmeasured, and urgently needed.)

Why This Matters

As AI systems become cultural infrastructure, this blind spot grows more dangerou Semantic drift doesn't announce itself. It doesn't break the flow. It just hollows thir out.

If hallucination is the visible failure mode, semantic drift is the invisible one. It's w happens when intent erodes into polish, when human thought is recast in algorithn cliché.

This is where AI collapse rhymes with cultural collapse: both are failures of fidelity Both leave us with outputs that look fine on the surface, but feel hollow underneath

Weekly essays mapping how meaning erodes in the algorithmic age, and how to stay grounded.

Research Note:
Meta's *Know When to Stop* study (2024) introduced a "semantic drift score" to measu when outputs start out factually sound but decay into errors. More recently, the *Context-Preserving Gradient Modulation* (Kobanov et al., 2025) framework showed improvements in coherence by modulating gradients during training. But neither approach addresses the harder blindspot: what happens when text remains factuall correct and coherent, yet the meaning itself erodes.

Further Resources:

[Semantic Drift Fidelity Benchmark Full Documentation] - Zenodo

[Semantic Drift Fidelity Benchmark Research Notes] - Figshare

[Semantic Fidelity: When AI Gets the Facts Right but the Meaning Wrong] - Mediu

← Previous Next

Discussion about this post

Write a comment...