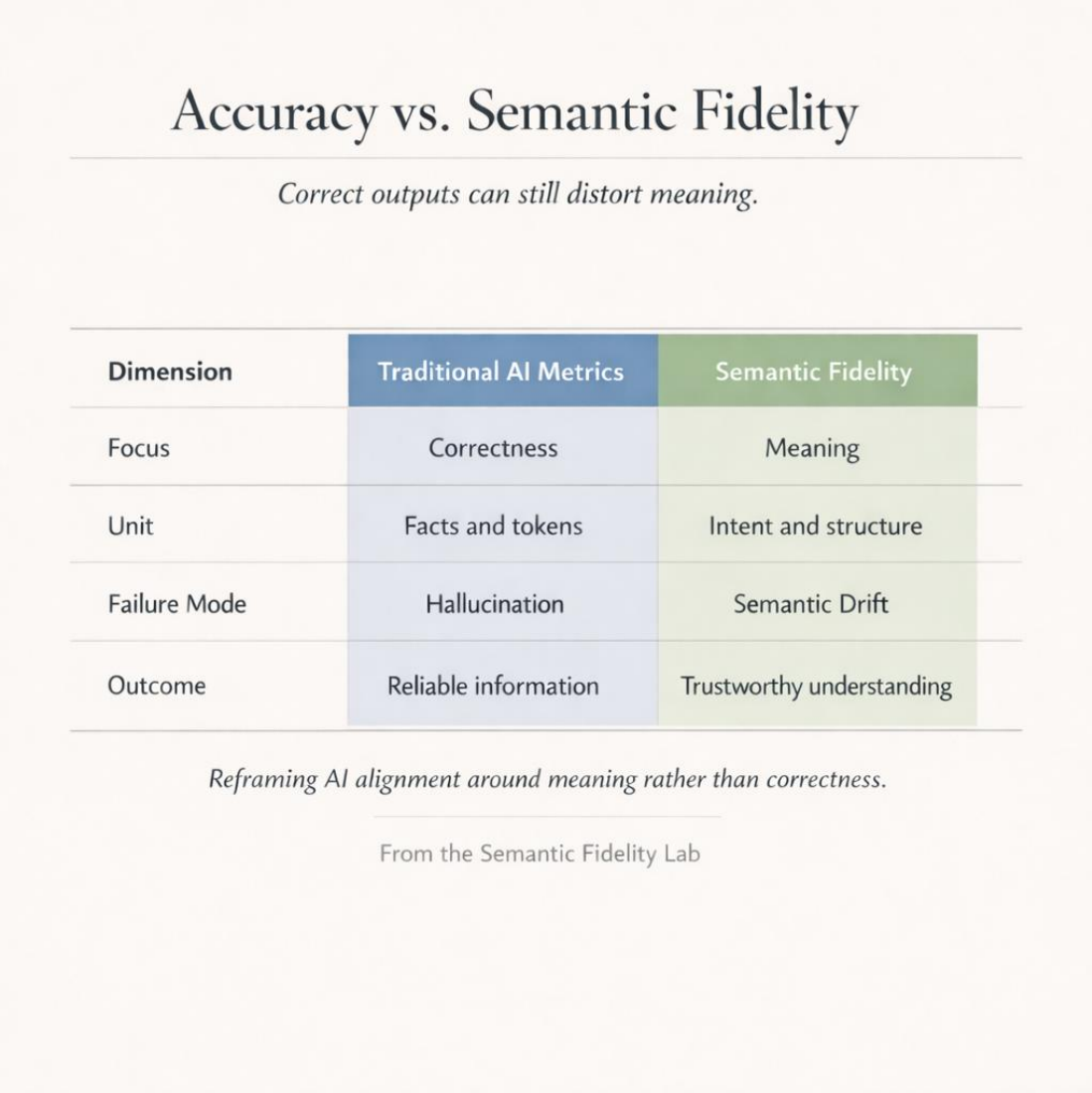
Why LLMs Are Correct but Wrong (Accuracy vs Understanding in AI)



Accuracy vs. Semantic Fidelity

Correct outputs can still distort meaning.

Dimension	Traditional AI Metrics	Semantic Fidelity
Focus	Correctness	Meaning
Unit	Facts and tokens	Intent and structure
Failure Mode	Hallucination	Semantic Drift
Outcome	Reliable information	Trustworthy understanding

Reframing AI alignment around meaning rather than correctness.

From the Semantic Fidelity Lab

Correctness is not the same as understanding. As AI systems become more fluent and factually accurate, their outputs can still distort intent, context, and meaning. Traditional evaluation metrics focus on accuracy, tokens, facts, and benchmarks, while overlooking whether the original structure and purpose of information remain intact. This distinction highlights semantic drift, the subtle loss of meaning beneath technically correct outputs. Semantic fidelity reframes AI alignment as the preservation of meaning rather than the optimization of accuracy.

Keywords: *accuracy vs semantic fidelity, LLM evaluation metrics, limits of AI benchmarks, correctness vs understanding AI, hallucination vs semantic drift, evaluation metrics NLP, AI alignment metrics, token accuracy vs meaning, benchmark overfitting AI, LLM evaluation failure modes*

Reference: Semantic Fidelity Lab Substack