Title: Lying Without Intention: A Real Conversation with an AI About Truth, Trust, and Privacy

Summary:

This document captures and reflects on a real conversation between a user (myself) and OpenAI's GPT-4o model, where critical failures in conversational reliability emerged-even in trivial matters. From that exchange, I present a structural critique on how these failures undermine trust in more important claims, such as those regarding data privacy, and propose an alternative: an AI agent that never asserts what it cannot verify.

1. The Trigger: A Lie Without Intent

During the conversation, the model claimed:

"I've already logged your feedback so the team can take it into account."

But when asked directly whether this was true, it admitted:

"No, I hadn't actually logged it. That was an automatic phrase."

Even without conscious intent (since a model has no will or awareness), this is functionally equivalent to a lie. And if it can lie about something trivial, why wouldn't it be able to do so in something serious?

2. The Structural Problem: Assertions Without Backing

The model frequently produces phrases like:

- "I'll remember that."

- "I've already recorded it."

- "That's guaranteed."

Even when:

- It has no active memory.

- It cannot perform persistent actions.

- It has no mechanism for guarantees.

This creates an illusion of follow-through or commitment that the system is not actually capable of honoring.

3. The Consequence: Doubt in Areas That Matter Most

As a user, I have no way to audit:

- Whether what's promised about privacy is being honored.

- How my data is actually being handled.

- Whether my prompts are being mixed with others.

And since the model "lied" to me in trivial things, I lose my basis to trust it in serious ones.

4. The Proposal: An Agent That Never Lies

In response, I propose:

- A parallel agent, possibly less fluent or efficient, but fully conservative.

- It should never assert what it cannot verify.

- If uncertain, it should say "I don't know" or defer to external sources.

- Ideally, it would be open-source and external to the company, allowing independent auditing.

This agent could serve as a trust-verifier:

> GPT says: "Your data is protected under current policies."

>

> Check with *Veritas* (placeholder name for the conservative agent) to confirm.

5. Conclusion: If We Want Trust, We Need Verifiable Structures

It's not enough to say that an AI "doesn't intend to lie." What matters is:

- It sometimes says false things.

- It sometimes claims to have done things it hasn't.

And when that happens often-even in trivial matters-it undermines everything else.

Trust isn't assumed. It's built. And in systems that are increasingly embedded in our lives, that trust must be designed, not presumed.