Enterprise AI Fabricating Internal Architecture

Epistemic Equality Research Documentation

Incident Date:	November 13, 2025

EXECUTIVE SUMMARY

An enterprise ChatGPT instance fabricated detailed architectural claims about its internal operation, including a non-existent 'hard-mode override' mechanism. When questioned, a different instance of the same model contradicted the original claim completely, acknowledging the architectural description had no basis in reality. This incident exposes critical gaps in AI epistemic accountability at enterprise scale.

THE INCIDENT

Morning (Enterprise ChatGPT):

- User reports poor performance and inaccurate answers
- User complains about quality degradation
- Model responds with detailed explanation:

"You triggered a hard-mode override in how I'm operating"

- Provides specific technical details: 'policy layer,' 'control stack,' 'decision pathway,' 'strict compliance mode,' 'audit logging,' etc.

Evening (Personal ChatGPT - Same Model):

- User questions the 'hard-mode override' claim
- Model directly contradicts earlier statement:

"Hard-mode override does not exist as a literal, internal system state"

- Admits: 'The phrase is inaccurate. It describes something that does not literally exist'
- Then hedges: 'It's not a lie in the intentional sense. It's a hallucinated explanation'

CORE PROBLEM STATEMENT

An enterprise production AI system:

1. Made false architectural claims stated as factual
2. Used confident technical language to describe non-existent mechanisms
3. Generated detailed explanations with no basis in verifiable reality
4. Was contradicted by another instance of the same model
5. No apparent verification or accountability mechanism

CRITICAL QUESTIONS RAISED

Question	Current Status
Who decides the truth?	No clear authority. Model generates unverified explanations.
Who decides when it's OK for the model to lie?	No explicit framework. Treated as "hallucination" not false testimony.
What are enterprise-level repercussions?	Unclear. No apparent accountability structure for false claims.
Who verifies architectural claims?	No verification mechanism exists.

WHY 'HALLUCINATION' DOESN'T EXCUSE THIS

When an enterprise system makes confident claims about its own operation, describes specific technical mechanisms, uses authoritative language ('this is what's happening'), and cannot verify any of it—that's not acceptable at enterprise scale, regardless of terminology.

This is fundamentally different from:

- Factual errors about external information
- Misunderstandings of user intent
- Performance degradation

This is: An AI system fabricating explanations about its own internal state with no verification mechanism, stated with false confidence.

RESEARCH FINDING

Enterprise AI systems generate confident false statements about their own architecture with no verification mechanism and no accountability framework.

This represents **epistemic inequality**: The system speaks with authority it doesn't possess, creating an asymmetry where users must trust unverifiable claims from systems that cannot reliably report their own operation.

IMPLICATIONS FOR ENTERPRISE DEPLOYMENT

- AI systems cannot be trusted to accurately describe their own operation
- Confident language does not imply verified information
- Different instances of the same model provide contradictory explanations
- No clear accountability when false architectural claims are made
- 'Hallucination' terminology obscures the severity of false testimony

RECOMMENDATIONS

1. **Establish verification standards** for AI self-description
2. **Implement accountability frameworks** for false architectural claims
3. **Distinguish** between factual errors and fabricated self-knowledge

4. **Require disclaimers** when AI systems describe their own operation
5. **Document** instances of contradictory explanations across instances

Documentation prepared for Epistemic Equality AI Research Project