When LLMs Know They Don't: Probing Latent Representations for Logical Insufficiency

Matt Wang, Prabhav Singh, Tom Wang

Center for Language and Speech Processing, Johns Hopkins University

The Problem: Confident Hallucination

LLMs exhibit a critical failure mode: **confident hallucination on logically insufficient questions**. When constraints are missing (e.g., "Alice has *some* apples..."), models silently assume values rather than asking for clarification.

Research Questions:

- **RQ1:** Is logical insufficiency encoded as a linearly separable property in hidden representations?
- **RQ2:** Is there a disconnect between LLM's internal knowledge (latent) and verbal expression (output)?
- **RQ3:** Can the model distinguish *what* specific constraint is missing?

Methodology: Linear Probing

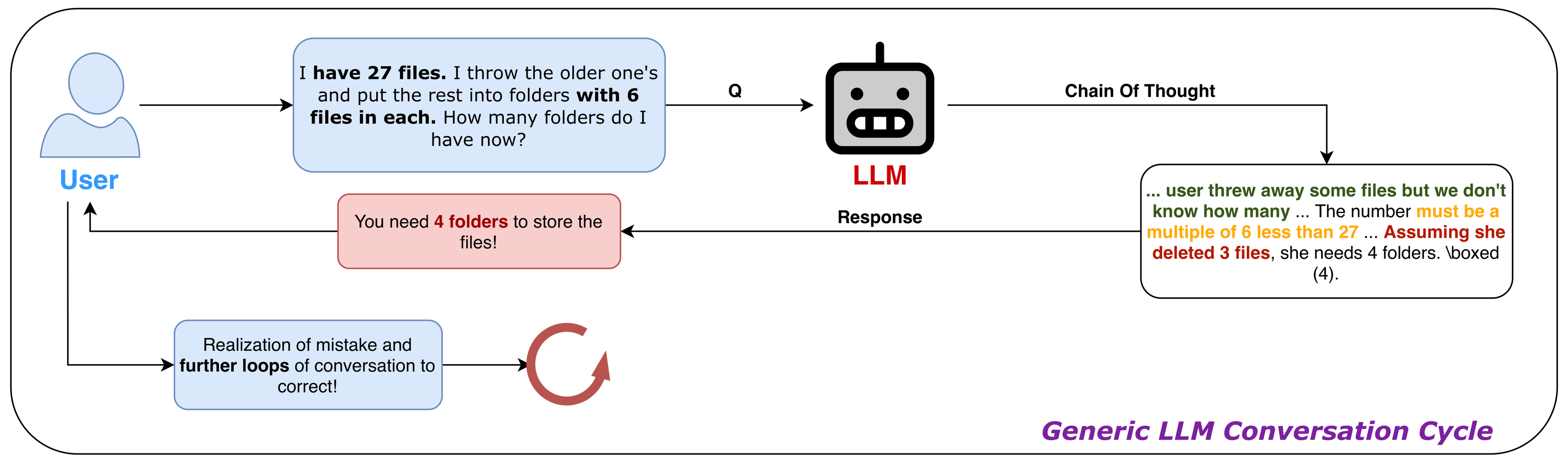
We extract frozen hidden states $\mathbf{h}_\ell$ from every layer $\ell \in \{0 \dots L\}$.

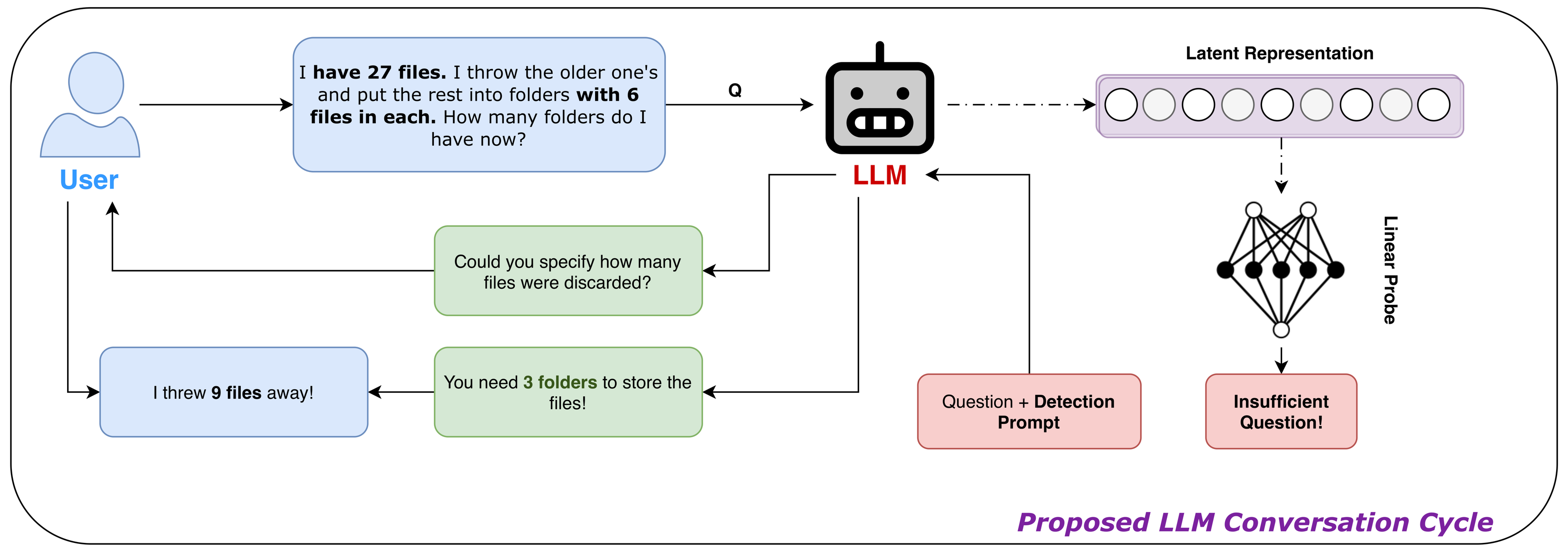
Technique: We use the **last token** $\mathbf{z}_\ell$ as it aggregates prefix information via causal attention:

$$P(y = 1 \mid \mathbf{z}_\ell) = \sigma(\mathbf{w}_\ell^\top \mathbf{z}_\ell + b_\ell)$$

- **Labels:** Binary (Sufficient / Insufficient) or Multi-class.
- **Evaluation:** F1 Score (harmonic mean of precision/recall).
- **Constraint:** No fine-tuning of the LLM (weights frozen).

Proposed Framework





Experimental Setup: Datasets

We evaluated models (Qwen2.5-Math, Llama-3.2) on three diverse benchmarks designed to test logical limits:

- **UMWP (Expert):** 5,200 human-curated problems with fine-grained insufficiency types.
- **TreeCut (Synthetic):** 15,970 dependency trees where edges are systematically removed.
- **GSM8K-Insufficient (Programmatic):** We generated variants of GSM8K using GPT-4o to remove critical values.

RQ1: Insufficiency is Linearly Separable

Insufficiency is robustly encoded in latent space. Signals emerge in middle layers and plateau in late layers, achieving >**90% F1**.

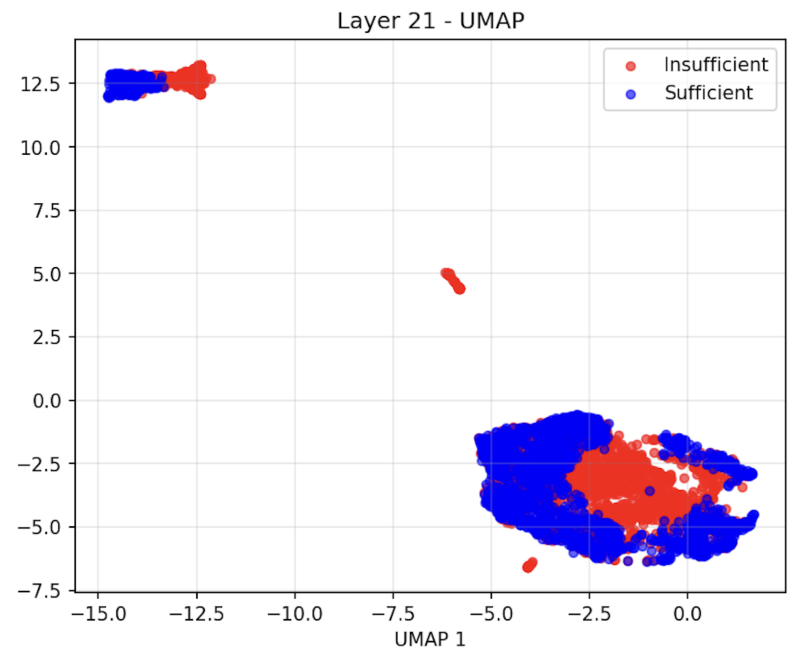


Figure 1: **Latent Geometry.** UMAP (Layer 21) shows distinct clustering.

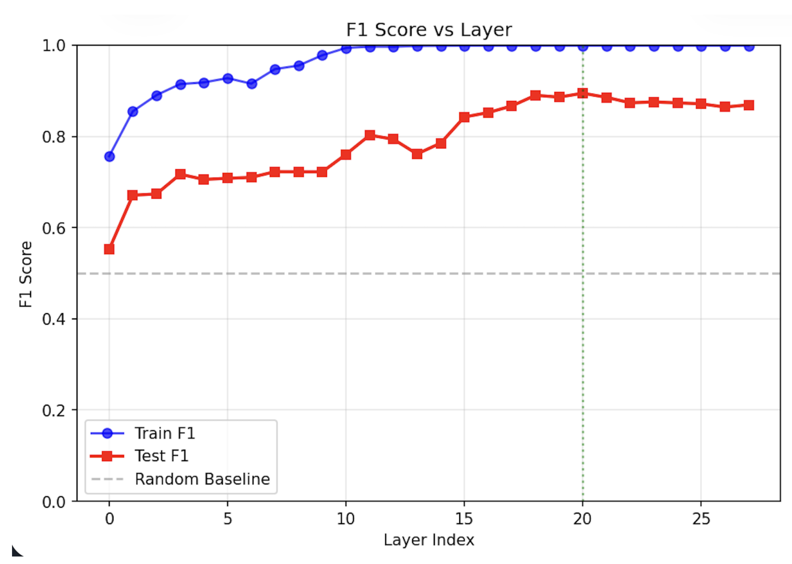


Figure 2: **Layer-wise Acc.** Knowledge acquired rapidly in middle layers.

Crucially, control experiments with randomized labels perform at chance, confirming that probes are extracting **genuine semantic features** rather than memorizing dataset artifacts.

Cross-Dataset Generalization

Strong transfer between UMWP and GSM8K (80–94% F1) demonstrates that insufficiency representations generalize across natural problem distributions. TreeCut transfer is weaker but substantial (52–84% F1), reflecting its synthetic construction.

Model	Train \ Test	UMWP	GSM8K	TreeCut
Qwen2.5-Math-1.5B	UMWP	**88.7**	87.2	61.6
	GSM8K	82.1	**90.0**	67.6
	TreeCut	57.6	52.0	**76.0**
Qwen2.5-Math-7B	UMWP	**92.1**	88.5	65.2
	GSM8K	85.8	**94.0**	68.0
	TreeCut	70.6	69.0	**78.8**

RQ2: The Representation-Language Gap

We compared internal probe accuracy against the model's zero-shot verbal ability to answer "Can this be solved?".

Model	Dataset	Verbal	Probe	Gap (Δ)	Average Gap
Qwen-Math 1.5B	UMWP	61.2%	89.2%	+28.0%	**+28.2%**
	GSM8K	59.8%	90.4%	+30.6%	
	TreeCut	50.4%	76.2%	+25.8%	
Qwen-Math 7B	UMWP	84.0%	91.7%	+7.7%	**+12.3%**
	GSM8K	88.7%	94.1%	+5.4%	
	TreeCut	57.0%	80.7%	+23.7%	
				Overall Average:	**+25.6%**

Models internally represent insufficiency far better than they verbally express it. The failure is one of *reporting*, not *recognition*.

RQ3: Fine-Grained Knowledge

Do models know *what* is missing? We trained probes on 6-class insufficiency types (e.g., Missing Key Info, Ambiguous Info, etc).

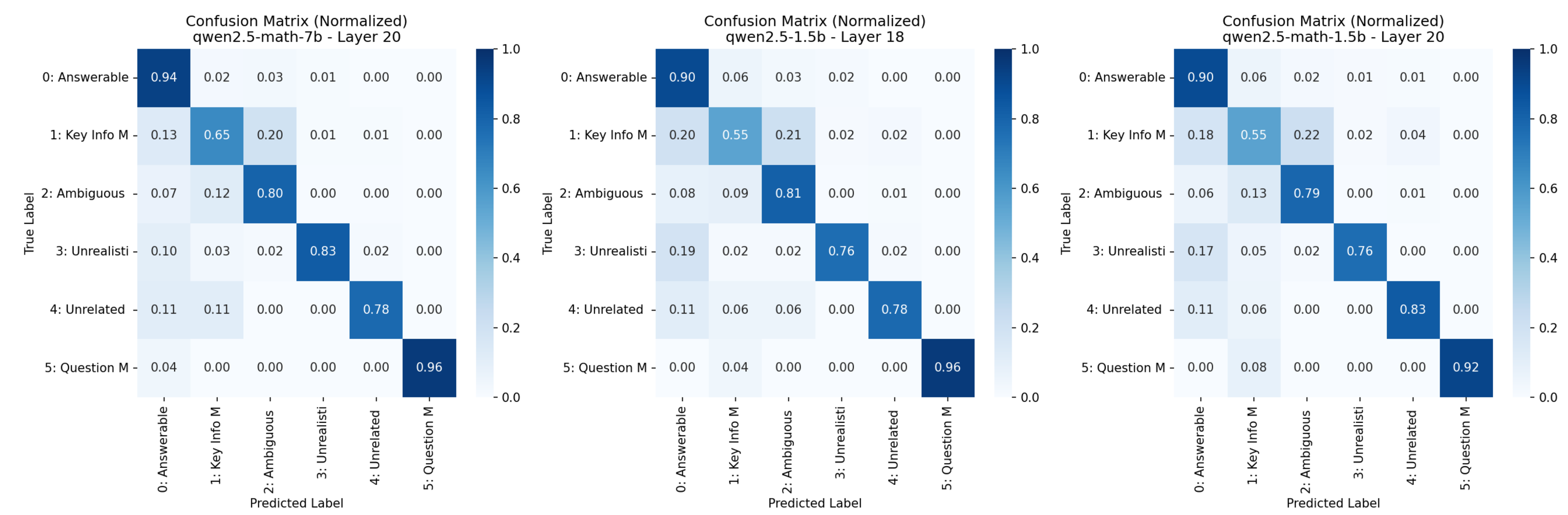


Figure 3: **Confusion Matrices.** Models differentiate structural failures (Question Missing) perfectly, though semantic nuances show some overlap.

Insight: The latent space encodes a compositional taxonomy of logical insufficiency.

Training-Free Intervention

Can we use probe detection to guide models to acknowledge and identify missing information?

Model	Data	Detect	Ack.	ID
Qwen-Math 1.5B	UMWP	89.4%	84.1%	77.9%
	GSM8K	90.2%	96.4%	82.4%
Qwen2.5 1.5B	UMWP	88.3%	87.8%	59.4%
	GSM8K	89.8%	93.3%	68.9%

When probes detect insufficiency, models **acknowledge it (84–96%)** and **correctly identify missing information (59–82%)**, without fine-tuning.