CARDS: Causal Auditing of Representations for Dialogue Sufficiency

Matt Wang
hwang302@jh.edu

Prabhav Singh
psingh54@jh.edu

Tom Wang
xwang397@jh.edu

1 Introduction and Problem Statement

Large language models increasingly serve user information needs, yet their behavior when faced with insufficient or ambiguous queries remains poorly understood [12]. We address this gap through a computationally efficient framework that detects and resolves input deficiencies without relying on expensive reasoning generation.

1.1 The Insufficiency Gap and Unstated Assumptions

The practical reliability of LLMs is compromised when user queries fall into the **Insufficiency Gap**, failing to contain all necessary information for a single, confident answer. Recent benchmarks reveal that current models struggle to identify ambiguity even with chain-of-thought prompting [12], often exhibiting overconfidence without improved accuracy. This gap drives two critical failures in current systems.

First, **Silent Failure (Unstated Assumptions)** occurs when LLMs proceed by making implicit, unstated assumptions to generate plausible, confident responses. Users typically trust these outputs, unaware that the answer is conditional. Consider the user question: "In a 30°-60°-90° triangle with a side length of 10 cm, find the area." The LLM behavior involves silent assumption: `‘‘...the side length of 10 cm corresponds to the hypotenuse because of the problem’s typical setup...’’` The model assumes the most common case, bypassing two other correct possibilities. Research on selective answering demonstrates that models frequently fail to recognize when clarification is needed [5].

Second, **Inefficient Exhaustive Problem-Solving** occurs when models *do* recognize ambiguity but resort to exhaustive case enumeration or lengthy back-and-forth dialogue to gather missing context. This strategy generates verbose output, overwhelms the user, and wastes computational resources. For example, given the user question `‘‘Jane’s quiz scores were` $98, 97, 92, 85$ `and` 93. `What was her score?’’` an inefficient LLM calculates the sum, the mean, the median, and the mode, exhausting all possible interpretations of "score."

1.2 Research Focus: Efficiency via Internal Control

Current methods often incur high computational cost due to verbose reasoning or reliance on external analysis. Our research introduces a system based on **efficiency** and **mechanistic control** to solve these failures, rather than relying on full-model reasoning. Our approach achieves **computational efficiency** where the diagnostic step requires negligible computation compared to generating a multi-step reasoning process (CoT) or sequentially decoding multiple possible answers. Recent work demonstrates that linear probes on hidden states can effectively identify internal model properties such as truthfulness [8, 2] and uncertainty [6], providing a foundation for our efficient diagnostic approach. Additionally, we establish **causal control** through a direct, programmatic link between the model's internal uncertainty signal and its corrective dialogue action using activation steering [10], combined with parameter-efficient fine-tuning [7] for the clarification policy.

2 Methodology: The Two-Stage Causal Framework

We operate within the **Quantitative Reasoning Domain** (math and logic problems) because it provides objective criteria for sufficiency and correctness, allowing for definitive labeling of the problem types.

2.1 Stage I: The Internal Audit (Diagnosis and Speed)

The **Audit** module achieves a near-zero latency diagnosis of the input's deficiency. For audit input, we analyze the LLM's **hidden embeddings**—the numerical vectors representing text understanding at a specific internal layer. These embeddings are known to linearly encode concepts like uncertainty and knowledge fidelity [8, 2].

Our mechanism employs low computational cost by utilizing a **Linear Probe (LP)**, a very shallow classifier, trained on these frozen embeddings [3]. This technique is highly efficient as it does not require updating the billions of parameters in the main LLM. For classification, the LP is trained to differentiate between the types of deficit to guide the intervention, including **Missing Variable**, **Ambiguous Reference**, and **Procedural Ambiguity**.

2.2 Stage II: The Clarifier (Causal Policy)

The **Clarifier** is the model's specialized policy for generating the precise, high-utility question. Through **causal intervention**, the diagnostic signal from the Audit is immediately used as a **steering vector** [10]. This vector is injected into the LLM's internal state at inference time, acting as a direct switch to **override** the model's default impulse to generate a confident answer, forcing it to enter the clarification mode.

For policy training with low computational cost, we train the Clarifier using **Parameter-Efficient Fine-Tuning (PEFT)**, specifically **LoRA** (Low-Rank Adaptation) [7]. PEFT significantly reduces computational cost by only training a small subset of new parameters. The policy undergoes **utility-based alignment** using Direct Preference Optimization (DPO) [9], a stable and computationally lightweight preference learning method. The reward function is weighted by **Conversational Utility**[1], strictly rewarding the question that leads to the lowest **TTS**[2] and highest final accuracy.

3 Data and Experimental Validation

3.1 Custom Dataset Construction: Math Ambiguity Dataset (MAD)

We develop a custom **Math Ambiguity Dataset (MAD)** to train and evaluate our framework. This is necessary because existing math problem sets are designed to be fully solvable and lack the labeled failure modes required for our Audit.

Our source data consists of solved, high-quality quantitative problems from GSM8K [4] and MathQA [1], used as the starting point. The adaptation strategy involves systematically modifying problems through **deletion of constraints** or **generalization of terms** to create failure cases (e.g., changing "hypotenuse" to "side length") that challenge the LLM's internal representation. For labeling, each generated sample is labeled with the precise **Insufficiency Type** (for the Audit) and the **Gold Clarification** (the corrective question for the Clarifier).

Table 1: Research Questions and Evaluation Metrics

Research Question	**Focus and Expected Contribution**	**Primary Metric**
Q1: Mechanistic Validation	Can the low-cost Linear Probe expose the causal link between internal model miscalibration and quantitative error?	Detection F1-Score
Q2: Efficiency Advantage	Does the Audit-triggered intervention outperform verbose reasoning methods (CoT) by minimizing Turns-to-Solution?	TTS and Inference Latency
Q3: Policy Fidelity	Does conditioning the Clarifier on the Audit's diagnosis lead to superior clarification specificity and higher final Task Success Rate?	Clarification Quality and Task Success Rate

3.2 Research Questions and Experiments

We conduct comparative experiments on the MAD test set to validate our system, addressing three key research questions. Our comparative experiment benchmarks our Two-Stage Framework against two key baselines: (1) a

[1] Conversational Utility is defined as the efficiency and effectiveness of a dialogue turn, measured by how much it contributes to task completion while minimizing user effort.

[2] TTS (Turns-to-Solution): The number of dialogue exchanges required between the user and system to arrive at a correct final answer.

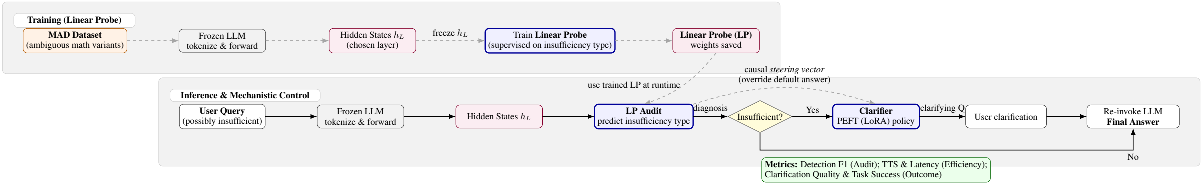

Figure 1: **Training vs. Inference (separated).** Top: train a Linear Probe (LP) on frozen hidden states h_L from MAD; save as an artifact. Bottom: use the trained LP to audit h_L from the user query; the diagnosis gates a Clarifier (PEFT/LoRA) via a causal steering vector. If sufficient, skip Clarifier and answer directly.

Standard LLM (testing raw guessing behavior) and (2) a **CoT-Prompted LLM** [11] (testing the cost/benefit trade-off of standard reasoning techniques against our efficient, targeted intervention).

4 Expected Contributions

CARDS provides three significant contributions to the field. First, we demonstrate **efficient intervention** by showing that lightweight internal probing can guide dialogue policy more efficiently and at a lower computational cost than external reasoning approaches. Second, we advance **AI Safety** by providing a framework for reducing silent failures and misplaced confidence, enhancing user trust and reducing over-reliance on unverified assumptions. Third, we provide **mechanistic insight** by validating the hypothesis that complex behavioral failures (like assumption-making) can be accurately predicted and corrected by analyzing internal model states.

This work establishes a new paradigm for computational efficiency in conversational AI while bridging interpretability research with practical dialogue systems, demonstrating how understanding internal representations enables targeted behavioral control without expensive reasoning overhead.

References

[1] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[2] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

[3] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[5] Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, 2023.

[6] Bairu Hou, Yujian Zhou, Vatsal Jain, Kristen Howell Geiger, Yizhong Zong, Adam Tauman Kalai, and Hamid Palangi. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*, 2023.

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[8] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

[9] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[10] Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini Kenton, Juan Felipe Ceron de Oliveira Marinho, and Alex Gerovitch. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[12] Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand, 2024. Association for Computational Linguistics.