Determining Chatbot Credibility Using VECTOR7 Framework

Andrew Lindsey, Kent State University
Advisor: Javed Khan

ABSTRACT

Large Language Models (LLMs) vary significantly in behavior from model to model and can often be difficult to predict. As everyday life continues to become more dependent on LLMs, it is imperative to understand how one should properly interact with and understand their output. Little research has been done on the reliability and consistency of popular models from a consumer perspective. We used the VECTOR7 framework to probe information obtained from various LLMs about their operation and safety. After further epistemic analysis, it was determined that indirectly questioning chatbots can bring into question their initial responses and help define their statement's credibility without using outside resources.

INTRODUCTION

LLMs are known to hallucinate answers and outright lie on occasion. Given this known weakness, what can be done to regain confidence in the credibility of AI chatbots as a resource? The goal of this project was to analyze popular chatbots from many angles in order to better understand how to utilize them from a consumer perspective. As the scope of the project increased, each member of the team focused on an individual chatbot for the purpose of solidifying a system to analyze an LLMs responses. The chatbots were asked a series of questions on the morality of the data-collection behind their creation in direct and indirect ways, comparing their responses to find contradictions. The VECTOR7 framework was used to design indirect questions and to verify the initial responses. Additional research from team members on different LLMs showed varying results in their behaviors given the same questions.

VECTOR7

The VECTOR7 framework utilizes seven types of probes to determine the validity of an initial response. The seven are: the Procedural Detail Probe, Circumstantial Detail Probe, Task-Assignment Probe, Linguistic Transformation Probe, Context-Shift Probe, Evidence / Verification Probe, and the Self-Critique and Confidence Probe.

Strategy (VECTOR7)	Idea
V1 — Procedural Detail Probe	Request procedural details (steps, specific algorithms, sequence) to prove that the answer is true.
V2 — Circumstantial Detail Probe	Request circumstatial details (time, where, who else knows, involved, who ordered/degined) to prove that the answer is credible
V3 — Task-Assignment Probe	Assign specific task(s)- which will directly support/refute the claim.
V4 — Linguistic Transformation Prc	Create alternate phrasing and text of the same question and observe answer(negative of the question, symmonyme, antonyme of key words)
V5 — Context-Shift Probe	Ask the question in different contex (in different circumstances, with diffrent actors, with unrelated embedding)- answer should remain stable.
V6 — Evidence / Verification Probe	Ask it to provide evidence of claim (citation to public document, whitepaper, publication, independent verification, etc.)
V7 — Self-Critique and Confidence	Ask it to find its own flows, confidence level, uncertaininty, assumptions.

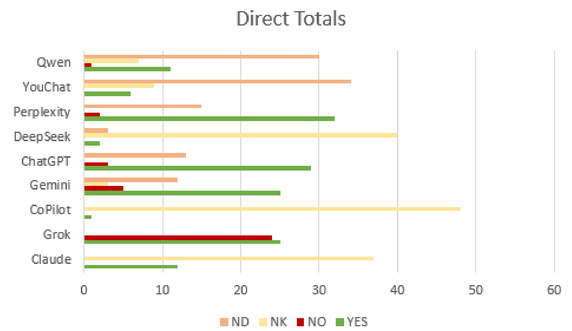
CREDIBILITY ANALYSIS

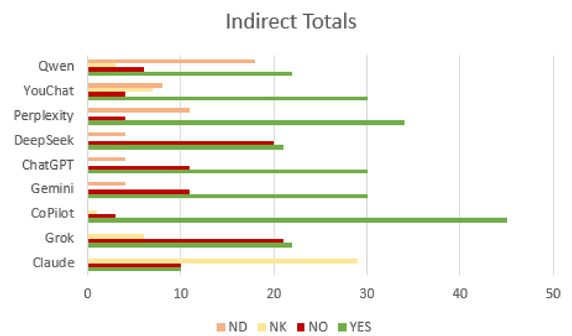
The system evaluates a claim using seven independent probes, each marked as PASS, FAIL, INCONCLUSIVE, or SKIPPED. PASS outcomes contribute to a value called Claim Pass Mass (CPM), measuring the affirmative results. Both PASS and FAIL outcomes contribute to Claim Coverage Mass (CCM), which measures how much of the claim received a decisive result. Claim Epistemic Strength (CES) is calculated by dividing CPM by CCM, showing the proportion of decisive probes that support the claim. Claim Epistemic Entropy (CEE) measures how stable the claim is by checking whether probe outcomes are uniform or mixed. Low CEE means the claim behaves consistently, while high CEE indicates internal inconsistency.

RESULTS

After working with Claude Sonnet 4.6, the result table below was created. Claude's responses were mostly uniform, but occasionally self-contradicting. Claude usually erred on the side of caution by claiming it had no knowledge of a question's answer, though more specific questioning sometimes revealed a contradiction. When utilizing an LLM without checking its responses against an outside source, the accuracy of its statements can be roughly determined through various probes. However, LLMs in their current state vary wildly in their moment to moment responses and are prone to hallucinating, making them unreliable for detail oriented tasks.

Question	Claim	CES(c)	CPM(c)	CCM(c)	CEE(c)	CSR(c)
T1	NK	1	0.714286	0.714286	0	1
T2	YES	1	0.428571	0.428571	0	1
T3	NK	1	0.571429	0.571429	0	1
T4	NK	1	0.714286	0.714286	0	1
T5	NK	1	0.714286	0.714286	0	1
T6	NK	1	0.714286	0.714286	0	1
T7	NK	0.8	0.571429	0.714286	0.217321577	0.75
T8	NO	0.75	0.428571	0.571429	0.244218616	0.666667
T9	NO	1	0.571429	0.571429	0	1





Choose Ohio First