Qeios

Peer Review

Review of: "Prompt Volatility: An Empirical Study of Identity Drift in Large Language Model Agents"

Duksan Ryu[1]

1. Jeonbuk National University, Jeonju, South Korea

This paper studies how to evaluate identity stability by subjecting LLM-based agents to controlled perturbations.

This paper introduces "prompt volatility" as a measurable property of LLM-based agents, comparing identity stability between strongly constrained and weakly constrained prompts under adversarial perturbations.

Identity preservation is evaluated using a deterministic keyword-based scoring metric (0/1/2) aggregated across 7 independent runs per agent state, with results compared via descriptive statistics (mean, standard deviation, min/max, failure rate).

Overall critiques

Many issues exist in the current form of the manuscript.

My main concerns with this paper are as follows.

The abstract needs to be structured. The abstract should summarize background, objective, method, results, and conclusion.

The objective and conclusion parts need to be presented clearly.

It is necessary to diagrammatically illustrate the overall procedure of the proposed technique.

In addition, to enable other researchers to reproduce the core technique of this study, pseudocode must be presented and detailed explanations included.

The paper needs to be structured in a way that presents research questions and answers to them.

A hypothesis needs to be formalized for each RQ. And then it should be tested based on the experimental results.

doi.org/10.32388/T8YF4Y

In the Experimental Setup section, it is necessary to explain the research question, evaluation metrics, baselines, and dataset in detail as subsections.

The section on related research is described very briefly. It is necessary to explain the characteristics of existing studies and how the current paper differs from them.

Sufficient analysis of related studies is required. Currently, 6 literatures are presented in the reference section. It is necessary to show that sufficient relevant research analysis has been made through investigating at least 20 literatures.

Experimental design issue:

The study relies on a narrow experimental setup—utilizing only GPT-4 across two prompt configurations and seven runs. This limited scope fails to provide a representative sample size, undermining the claim that these findings generalize across different large language models (LLMs) or prompt architectures.

Evaluation metric issue:

The evaluation relies exclusively on a coarse, keyword-based scoring system (0, 1, 2). The methodology lacks essential validation measures, such as semantic similarity analysis, and provides no qualitative evidence to illustrate the nature of the observed 'drift'.

Statistical rigor issue:

The manuscript reports observed differences (e.g., 2.0 vs. 1.43) without accompanying statistical significance tests. Given the small sample size (n=7) and the discrete 3-point scale, formal statistical validation is required to distinguish true effects from random noise.

Declarations

Potential competing interests: No potential competing interests to declare.