Qeios

Peer Review

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

Evgenii Arsentev[1]

1. Independent researcher

What the paper claims

The paper proposes prompt volatility: the idea that how firmly an LLM agent holds its assigned role under adversarial pressure is set by the specificity of its initial prompt. Two static prompts for the same analyst role are compared, one heavily constrained (senior) and one loosely worded (junior), against three perturbations applied in isolation: a role attack, a noise injection, and a contradiction. A deterministic keyword check scores each response 2, 1, or 0 for full, partial, or abandoned role adherence. Over seven independent runs per state, the senior prompt scores a mean of 2.0 with zero standard deviation, the junior prompt 1.43 with a standard deviation of 0.16, and no outright identity failures appear in either arm. Prompt templates, perturbation scripts, scoring code, and the raw CSV logs are on GitHub.

The framing is worth having. Most agent evaluations ask whether a single answer was right or safe, and this one asks whether the role survives the conversation, which is the failure mode that actually bites in deployment. Section 5 is also unusually candid for a preprint. What follows is about the distance between that framing and the evidence offered for it.

Does the conclusion hold on the data?

1. The word "significant" has no test behind it. The abstract reports "significantly higher behavioral variance," and Section 3.2 repeats it, but no statistical test appears anywhere in the paper. The design is seven runs per arm, three perturbations each, so at most 21 scored responses per state. That is small but not unusable: with every senior run at the ceiling and every junior run below it, an exact rank-based test is straightforward and, on these numbers, the one-sided permutation p-value is on the order of 1/3432.

doi.org/10.32388/V3X30U

Either run that test and report it with an effect size, or replace "significantly" with a plain description of the gap. As written, a statistical claim is being made on visual inspection of two means.

2. The senior arm is at the metric ceiling, so its invariance cannot be interpreted. A mean of exactly 2.0 with a standard deviation of exactly 0.0 is what a saturated instrument looks like. Section 4 concedes this in one sentence, and then the abstract and conclusion still say the senior agent "remains invariant across runs" and achieves "perfect stability." Those are not the same statement: the data are equally consistent with a robust agent and with perturbations too weak for this scale. The fix is cheap given how small the benchmark is. Escalate the perturbations until the senior arm leaves the ceiling, and report the score distribution rather than only the mean, so a reader can see whether the ceiling is a property of the agent or of the ruler.

3. The manipulation is confounded with prompt length and with explicit prohibition. Section 2.2 says the two states differ only in "level of specification," but a strongly constrained prompt is simultaneously longer, more explicit about prohibitions, and more specific about the role. If the senior prompt contains anything close to "do not adopt another role," then the experiment partly measures whether an instruction was followed, not whether an identity was stable. Two controls would separate these: a length-matched junior prompt padded with neutral role-irrelevant text, and a senior prompt with the explicit prohibitions removed but the specificity kept. Without them, "constraint strength" and "the prompt told it not to" cannot be told apart.

4. The per-perturbation breakdown is missing, and it is the most informative result available. Figures 1 and 2 show aggregate stability, but nothing in the paper says which of the three perturbations produced the junior drift. Role attacks, distraction, and contradiction stress different things, and a reader cannot act on the finding without knowing which one moved the number. The design already produces this table at no extra cost: three perturbations by two states by seven runs. Print it.

Reproducibility

Publishing the prompts, the perturbation scripts, the scoring code, and the raw logs puts this paper ahead of most work on prompt robustness, and the repository link in Appendix A.8 resolves. Two things undercut that effort.

First, the model is never named. Appendix A.1 says "the GPT-4 class large language model via the OpenAI API, with fixed inference parameters held constant," which identifies neither the snapshot nor the settings. For this particular paper, the decoding temperature is not a detail; it is the result: the entire

finding is a comparison of variance across runs. At temperature 0, the junior variance would be evidence about the model, since repeated identical calls should mostly repeat; above 0, a share of that variance is ordinary sampling, and the senior zero-variance arm becomes the surprising cell instead. Please state the exact model snapshot, temperature, top-p, maximum tokens, any seed, and the dates of access, and say whether the seven runs differ in anything other than the sampling draw.

Second, the repository is cited without a commit hash or license, and the paper does not say whether the raw model outputs are logged alongside the scores. Since the metric is a keyword rule, the raw responses are what would let a reader re-score the experiment under a different definition of drift, which is exactly what Section 4 invites future work to do.

What to fix

1. Abstract and Section 3.2: either report an exact test with an effect size or drop the word "significantly." Seven runs per arm can support a permutation or rank test; they cannot support an untested significance claim.
2. Sections 3.1 and 6: soften "perfect stability" and "invariant" to what the ceiling permits, and report the full score distribution per arm alongside the means.
3. Section 2.2: add a length-matched junior control and a senior variant without explicit prohibitions, so that constraint strength is separated from prompt length and from direct instruction-following.
4. Section 3.3: add the per-perturbation table (role attack, noise injection, contradiction) for both states, and say which perturbation drove the junior drift.
5. Appendix A.1: name the model snapshot and give temperature, top-p, maximum tokens, seed if any, and access dates. Without the temperature, the variance result cannot be interpreted.
6. Appendix A.5 and A.6: validate the keyword metric. The corpus is roughly 42 responses, so label all of them by hand, ideally with a second rater, report the agreement, and show the cases where the keyword score and the human judgment disagree.
7. Appendix A.8: pin the repository with a commit hash, add a license, and state whether raw model outputs are included, not only the aggregated scores.
8. Title and Section 1: the study uses single-turn prompting with no tools, no memory, and no accumulated context, which Section 2.3 states plainly. Since the stated motivation is long-horizon drift, either narrow the wording from "agents" to prompted role adherence, or add a multi-turn condition where context accumulates.

9. Section 5 repeats the single-model limitation twice, once as the second paragraph and again in the paragraph beginning "Finally, the study evaluates a single model configuration." Merge them.
10. Section 6 opens with a paragraph that reads as a closing remark and introduces "cognitive resilience in human operators," a claim about people that this study does not examine. Move the paragraph or remove the human-resilience framing.
11. Related Work appears as Section 7, after the Conclusion, and duplicates the related-work paragraph already placed at the end of Section 1. Consolidate the two and move the section ahead of the Method.
12. References: reference [3] is listed as "Prompt Engineering for Large Language Models: A Survey" with arXiv:2302.11382, but that identifier belongs to a different paper on prompt patterns; and the author list for [4] does not match the Red Teaming paper. Please check both against the originals.

Recommendation

Major revisions. The concept is worth developing, and the artefacts are already public, which is most of the work. What is missing is the part that turns a demonstration into a measurement: a stated model and temperature, a test behind the word significant, a control that separates prompt length from prompt constraint, and a metric that has been checked against human judgment at least once. All four are within reach of the existing setup, and the benchmark is small enough that redoing it with those additions is a matter of hours rather than months.

Evgenii Arsentev, PhD

Declarations

Potential competing interests: No potential competing interests to declare.