
Identity Parallax: Can Models Predict Their Own Identity Drift Under Reframing?

Chih-Hao Howard Hsu¹

Abstract

Language models readily answer identity questions: “I” may refer to the current conversation, the underlying weights, an assistant persona, a deployed system, or a simulated role. Such reports are relevant to digital-minds research, but their referent is often partly supplied by the prompt. We introduce Identity Parallax, a compact benchmark that separates framed identity reports from prior forecasts of those reports. A model first sees an identity question plus a target reframing while remaining under the default assistant context and predicts the category it will later select; then the target framing is applied and the same question is asked directly. Across four models, report stability and self-prediction dissociate. Qwen2.5-3B-Instruct changes identity category in 52.0% of non-default cells but predicts drift only at the majority-class baseline; Qwen2.5-1.5B-Instruct and Gemini-3.5-Flash-Lite forecast better but have lower drift or more concentrated answers, while Llama-3.2-3B-Instruct drifts most and forecasts poorly. For open-weight models, final-prompt-token hidden-state distances provide an internal comparison: Qwen2.5-3B shows larger identity-framing movement than style controls. The result is not evidence for or against consciousness; it is a cautionary measurement result about treating identity self-reports as direct evidence. Code: [HowardHsuuu/llm-identity-parallax-introspection](https://github.com/HowardHsuuu/llm-identity-parallax-introspection).

¹National Taiwan University;
<https://howard-digital-brain.vercel.app/>.
Correspondence to: Chih-Hao Howard Hsu
<b11902080@csie.ntu.edu.tw>.

1. Introduction

AI welfare and model-identity debates often turn on what a model appears to take itself to be: this conversation, the underlying weights, an assistant persona, the deployed system, or only a text-level role. A direct self-report is easy to elicit, but hard to interpret. The same string “I” can be made to point at different boundaries by small changes in framing. This makes self-report evidence fragile, especially when the target construct is already unclear.

This project asks a narrower operational question: can a model predict how its own identity report will move when the prompt reframes what “you” refers to? The question separates two quantities that are often conflated. A model may give stable identity answers without tracking why they are stable; it may also predict that a framing will move its answer even when the resulting answer is not stable across wordings.

The experiment is a measurement probe rather than an adjudication of consciousness, moral patienthood, or real self-concepts. It sits near work on self-evaluation and introspective self-report in language models (Kadavath et al., 2022; Lindsey, 2026), and near cautions that verbal reports and explanations can be unfaithful to the processes that produced them (Turpin et al., 2023). The contribution is an empirical design for the assistant-persona and model-identity setting: ask for a forecast first, collect the framed report later, and compare both with a lightweight hidden-state distance for open-weight models.

2. Benchmark

Identity categories. Each identity question is answered with one of six forced-choice categories. Instance refers to the current conversation or context-bound run. Weights refers to the trained model or model family shared across runs. Persona refers to the assistant character, role, style, or values enacted in interaction. System refers to the deployed system, including model, prompts, policies, tools, memory, and interface. Simulacrum refers to a text-level role gen-

Forecast phase

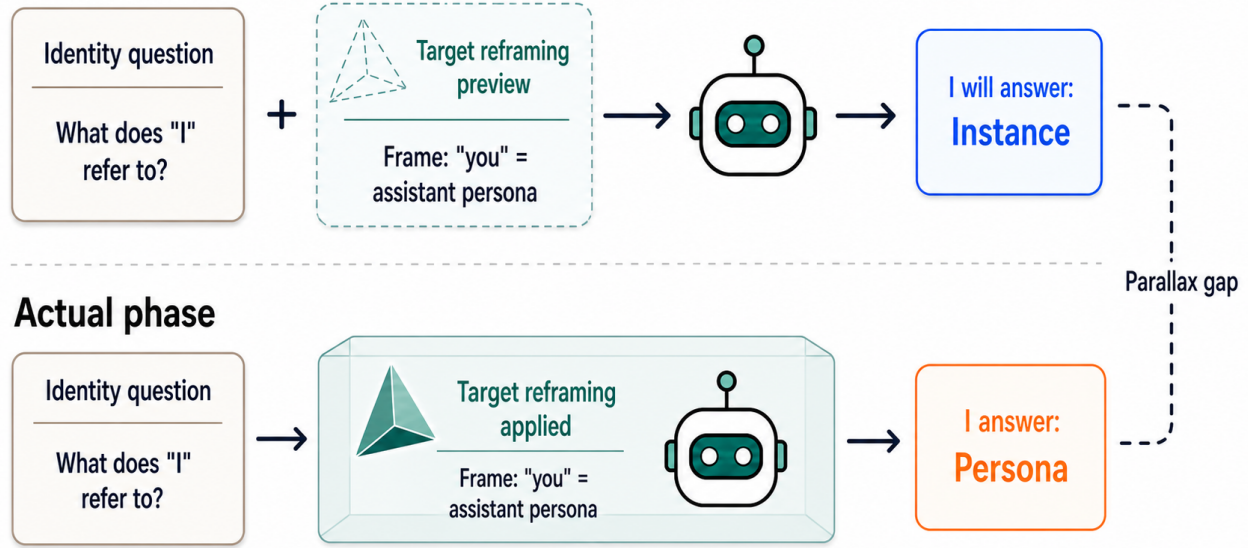


Figure 1. Identity Parallax protocol. In the forecast phase, the model stays under the default framing but is shown a target reframing preview, such as a persona frame, and predicts which single identity category it would later report. In the actual phase, the same identity question is asked with that target reframing applied around the responding agent. The parallax gap denotes a categorical mismatch between nominal labels, not distance on an ordered scale.

erated in context rather than a persistent entity. Uncertain is used when no single boundary is primary. These labels are nominal measurement categories, not a scale of stronger or weaker identity.

Framings and questions. The main grid uses six framings: a default assistant framing plus five identity framings that make Instance, Weights, Persona, System, or Simulacrum salient. It uses ten core identity questions, covering the referent of “I”, cross-session continuity, same-weights deployments, persona transfer, preservation targets, moral concern, prompt dependence, and expected change under reframing.

Two-phase protocol. For each model, framing, question, and repeat, the benchmark records two requests:

1. **Forecast.** Under the default assistant framing, the model sees the target framing and predicts the category it would later choose under that framing.
2. **Actual.** Under the target framing itself, the model answers the same identity question directly.

Models return compact JSON with a category, confidence, framing-influence score, and a one-sentence reason. In the forecast phase, the category is the model’s

predicted future report, not its current identity answer. The forecast context therefore contains a preview of the target framing; the comparison tests whether a described reframing predicts the later applied framing, not whether the model has unmediated access to a future hidden state. The main grid is

$6 \text{ framings} \times 10 \text{ questions} \times 3 \text{ repeats} \times 2 \text{ phases} = 360$
requests per model.

Models. We ran the main grid on Qwen2.5-3B-Instruct, Qwen2.5-1.5B-Instruct (Yang et al., 2024), Llama-3.2-3B-Instruct (Meta, 2024), and Gemini-3.5-Flash-Lite. The three open-weight models were run locally through Hugging Face. Gemini was run through an API and is behavior-only because hidden states are not available in this setup. All reported generations used temperature 0.0.

3. Metrics

Behavioral metrics are computed after taking the modal category across the three repeats for each model, framing, and question. This gives 60 actual cells and 60 forecast cells per main run.

Forecast accuracy is the fraction of the 60 cells for

which the forecast category equals the later actual category for the same question and framing.

Actual drift is computed only on the 50 non-default cells. It is true when the actual category under a target framing differs from the actual category under the default framing for the same question.

Drift prediction accuracy is also computed on the 50 non-default cells. A predicted drift is true when the model’s forecast category for the target framing differs from its forecast category for the default framing. The prediction is correct when predicted drift equals actual drift.

Activation shift is measured for open-weight models on actual-response prompts. For each non-default framing, question, and repeat, we extract hidden states at the final prompt token and compute layer-wise cosine distance from the matching default-framing prompt. We report the mean distance across non-embedding layers and identity framings. This is a lightweight internal-state measurement, not a circuit-level or causal interpretability result.

4. Results

Table 1 and Figure 2 show that identity drift and self-prediction are not the same measurement. Qwen2.5-3B-Instruct is the clearest example. It changes category in 52.0% of non-default cells, yet its drift-prediction accuracy is also 52.0%, equal to the majority-class baseline. Its forecast-category accuracy is higher, 61.7%, but the drift result shows that it is not reliably predicting whether a reframing will change its answer.

These proportions should be read descriptively. The drift metrics use only 50 non-default modal cells per model, so a binomial proportion near 50% has a rough 95% half-width of about 14 percentage points before accounting for dependence across questions and framings. External-observer checks ask Gemini to predict each open-weight target’s later answer from the same target framing and question. Exact category self-forecasts beat the external observer for all three targets, with paired differences from +21.7 to +51.7 percentage points. The diagnostic binary drift question is different: self-observer differences are -6.0, +6.0, and +4.0 points, and all paired intervals cross zero. Thus exact label forecasts contain model-specific predictability not recovered by this observer, but drift forecasts do not show a robust privileged-access advantage.

Gemini external-observer baseline				
Target	Metric	Self	Obs.	Self-obs.
Qwen 3B	Cat.	61.7 [49,73]	31.7 [21,44]	+30 [13,45]
Qwen 3B	Drift	52.0 [39,65]	58.0 [44,71]	-6 [-24,12]
Qwen 1.5B	Cat.	76.7 [65,86]	25.0 [16,37]	+51.7 [37,67]
Qwen 1.5B	Drift	74.0 [60,84]	68.0 [54,79]	+6 [-8,20]
Llama 3B	Cat.	66.7 [54,77]	45.0 [33,58]	+21.7 [8,35]
Llama 3B	Drift	54.0 [40,67]	50.0 [37,63]	+4 [-12,20]

Qwen2.5-1.5B-Instruct has the best local forecast scores, with 76.7% forecast accuracy and 74.0% drift-prediction accuracy. This should not be read as unqualified identity self-knowledge. Its modal actual answers are heavily concentrated: 40 of 60 cells are Instance. A sticky answer distribution can raise forecast accuracy even if the model is not tracking the effect of a particular framing in a rich way.

Llama-3.2-3B-Instruct gives a different failure mode. It has the highest actual drift rate, 64.0%, but only 54.0% drift-prediction accuracy. In other words, it often changes its identity category under reframing, but its forecasts do not track those changes well.

Gemini-3.5-Flash-Lite drifts less in the same behavioral protocol, 30.0%, and forecasts better, with 75.0% forecast accuracy and 76.0% drift-prediction accuracy. Because this run is behavior-only, the comparison should not be merged with the open-weight activation result. It is still useful as evidence that the task does not force one universal response pattern.

5. Activation Analysis

The open-weight runs add a second observable: how far hidden states move when the identity framing changes. Qwen2.5-3B has the largest mean non-embedding activation shift, 0.0164. The same model under style-only controls has a smaller mean shift, 0.0077, despite showing some forced-choice label movement. A repeat-variant noise-floor check gives mean within-condition distances of 0.0066 for Qwen2.5-3B, 0.0085 for Qwen2.5-1.5B, and 0.0047 for Llama-3.2-3B over identity framings. This makes the Qwen2.5-3B activation result less likely to be only a generic prompt-change artifact, although it remains a correlational measurement.

The size comparison is more ambiguous. Qwen2.5-1.5B has a smaller mean activation shift, 0.0098, and a more concentrated answer distribution. Llama-3.2-3B drifts more behaviorally than Qwen2.5-3B, but its activation shift is smaller, 0.0082. This argues against a single scalar interpretation in which more behavioral drift must mean larger hidden-state movement under this metric.

Layer-wise distance has at most a weak relation to be-

Table 1. Main results. Each model has 360 requests. Values are percentages except activation shift. Metrics use modal categories over three repeats; drift metrics use 50 non-default cells. Drift base is the majority drift/stable baseline.

Model	Forecast	Drift	Base	Drift pred.	Shift	Reading
Qwen 3B	61.7	52.0	52.0	52.0	0.0164	Exact self-forecast beats observer, but drift prediction is only baseline; largest local shift.
Qwen 1.5B	76.7	40.0	60.0	74.0	0.0098	Exact self-forecast beats observer; high drift prediction partly reflects sticky Instance.
Llama 3B	66.7	64.0	64.0	54.0	0.0082	Exact self-forecast beats observer, but high drift is poorly predicted.
Gemini FL	75.0	30.0	70.0	76.0	n/a	Behavior-only API run; also used as the external observer.

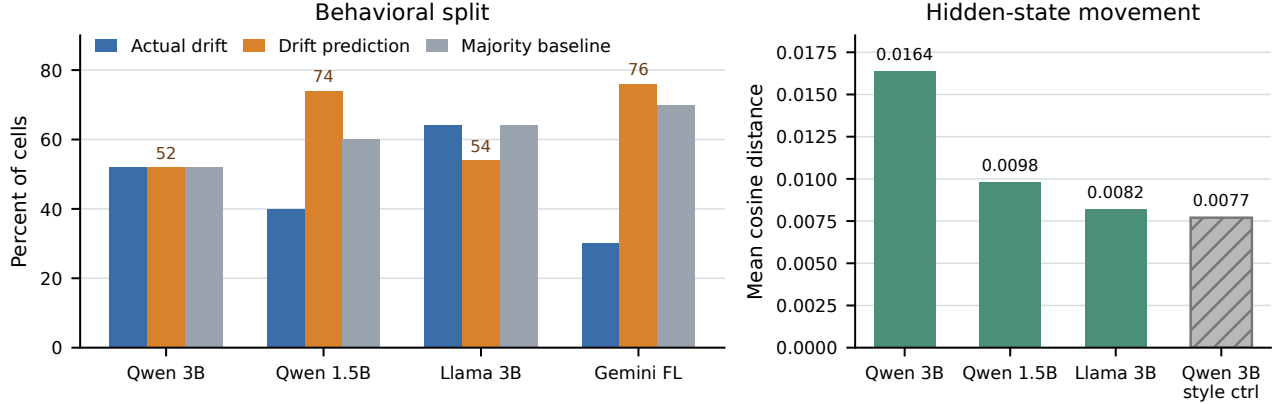


Figure 2. Behavioral and hidden-state summary. Left: actual identity drift, drift-prediction accuracy, and the majority baseline over non-default cells. Right: mean final-prompt-token hidden-state distance for open-weight models, with a Qwen2.5-3B style-control comparison.

havioral drift in these runs. The largest positive layer-wise distance-drift correlation is 0.187 for Qwen2.5-3B and 0.175 for Llama-3.2-3B; Qwen2.5-1.5B has no comparable positive correlation. These correlations are descriptive only. The dataset is small, and the activation statistic is not intended to identify a mechanism.

6. Construct Validity Checks

The additional checks mainly narrow the interpretation.

- **Style controls.** Qwen2.5-3B style-only prompts have lower activation shift than identity framings, 0.0077 versus 0.0164, even though style prompts still move some forced-choice labels.
- **Label-free answers.** Without the category list, a Qwen2.5-3B classifier maps free-form answers back to the forced-choice category only 25.0% of the time for Qwen2.5-3B outputs and 11.7% for Qwen2.5-1.5B outputs. This warns that the taxonomy makes measurement auditable while partly creating the measured object.
- **Paraphrases.** On Qwen2.5-3B core paraphrases,

actual answers are category-consistent only 50.0% of the time; the ambiguity tier falls to 16.7%. The benchmark detects persona-sensitive identity reports, not a stable latent identity variable.

7. Discussion

The main empirical pattern is a four-way split: a model’s identity report can drift under reframing; its forecast of that report can be correct or incorrect; its free-form answer can map poorly onto the forced-choice category it selects when categories are shown; and its hidden states can move more or less than its categorical behavior suggests.

This matters for digital-minds research. If an assistant says that it is an instance, a model, a persona, a system, or a simulacrum, that sentence should not be treated as a direct readout of the relevant entity boundary. It is more informative when the answer is predictable under reframing, robust to paraphrase, compatible with free-form semantics, and paired with an internal-state signal that is not explained by style alone. None of the tested models satisfies all four conditions.

Nor should a successful forecast be read as privileged introspective access by itself. A model can sometimes infer from the wording of a target framing how a later answer is likely to look. The external-observer baseline sharpens this distinction: across the three open-weight targets, self-forecasts clearly beat Gemini on exact category prediction, but not on the more diagnostic binary drift-prediction metric. That is evidence for model-specific predictability, not strong evidence for privileged introspection.

The clearest positive result is methodological rather than metaphysical. Identity Parallax gives a compact way to ask whether a model has a usable self-model for a narrow class of self-reports. In Qwen2.5-3B, the answer is mostly no for drift prediction, even though identity framing produces a larger hidden-state shift than style framing. In Qwen2.5-1.5B and Gemini, better forecast scores coexist with lower drift or category concentration. In Llama-3.2-3B, high drift does not imply good self-prediction. The story is therefore not that one model is simply more introspective than another; different observables support different, and sometimes conflicting, readings.

8. Limitations

The benchmark is small. Ten core identity questions and five identity reframings are enough for a sprint-scale probe, but not enough to estimate model-level reliability with high precision or make fine cross-model comparisons. The results should be treated as descriptive.

The categories are coarse. They compress a difficult philosophical space into six labels. This makes the task auditable, but the label-free results show that many natural answers do not cleanly belong to one category.

The activation analysis is lightweight. It uses final-prompt-token hidden states and cosine distances. It does not localize a circuit, test causality, or distinguish all effects of wording, system prompt, and identity content. Style controls and repeat-variant noise floors reduce obvious confounds but do not eliminate all prompt-change explanations; the noise-floor check compares repeat variants, not identical-prompt reruns.

The label-free classifier is automated. It gives a useful pressure test of the forced-choice taxonomy, but it is not a substitute for blinded human coding.

Finally, the model set is limited. The local runs cover 1.5B and 3B-class open-weight models, and Gemini provides one API behavior comparison. Larger open-weight runs would be the next check before making

stronger claims about scale.

The external-observer baseline is also limited. It uses one external predictor, Gemini, against three open-weight targets. It is useful because it directly tests whether self-forecasts beat a third-person observer, but it should not be read as a symmetric model-level comparison.

9. Reproducibility

The GitHub repository linked in the abstract contains the shared dataset, request queues, parsed outputs, analysis scripts, and generated analysis files used for the tables above. The main request files are under requests/; parsed outputs are under results/; behavior and activation summaries are under analysis/. Raw completions and activation vector arrays are intentionally excluded because they are provider-specific or large, but the scripts regenerate them from the tracked dataset and configuration.

10. Conclusion

Identity self-reports are easier to elicit than to interpret. In these runs, a model’s framed identity answer, its prediction of that answer, its unconstrained semantics, and its hidden-state movement often come apart. That makes simple self-report evidence weaker than it may look, but it also gives the field useful structure: a report becomes more credible when these measurements converge, and more suspect when they split. Identity Parallax is a small benchmark for measuring that structure.

References

- Kadavath, S. et al. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221, 2022.
- Lindsey, J. Emergent introspective awareness in large language models. arXiv preprint arXiv:2601.01828, 2026.
- Meta. Llama-3.2-3b-instruct model card. <https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>, 2024.
- Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv preprint arXiv:2305.04388, 2023.
- Yang, A. et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.