
The Interface Is the Intervention¹

*A Preregistered Multi-Model Audit of
Persona-Framed Synthetic Triage*

Frank Peterlein
Independent Researcher

With
Apart Research

Abstract

Persona prompts are often treated as behavioral interventions, yet their effects can be confounded with ordinary properties of the measurement interface. We conducted a preregistered audit of synthetic resource-triage choices. A frozen battery crossed four system prompts, two response formats, six pairwise items, four sampled repetitions, and alternating A/B order. Qwen3-4B served as the preregistered reference; a prospective extension tested Phi-4 mini, SmoLLM3-3B, and OLMo-2-1B under the same smoke-stop rule. Qwen and Phi completed 192 calls each. The analytical persona exceeded each model's neutral-wording floor by only 0.083 and 0.036; the coordinator persona did not exceed it. All 192 uncertainty-enabled outputs across those models were non-directional. Order-associated variation was larger than the persona excess, while 16.7% and 95.8% invalid smoke rates stopped SmoLLM3 and OLMo, largely because they copied enum-like schema strings. Persona was measurable but neither dominant nor measurement-invariant: response options, wording, order, and schema compatibility determined what the instrument could observe.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Large language models can be prompted to enact roles, but role-consistent language should not be mistaken for a stable inner character [1]. Empirical work likewise finds mixed and task-dependent effects of persona system prompts [2]. A clean persona test therefore needs a comparison not only with an unframed baseline, but also with semantically equivalent wording and with other elements of the elicitation procedure.

Synthetic triage is useful for such an audit because it demands a choice under explicitly incomplete information. It is also easy to overinterpret. The vignettes below have no externally correct answer, and the task is not clinical triage. Our question is narrower: do two deliberately strong role framings change observable pairwise choices or refusal more than a neutral paraphrase, and how large are those changes relative to response-format, displayed-order, and output-schema variation?

Contribution. We provide (i) a preregistered single-model pilot, (ii) a prospectively frozen three-model extension with a binding smoke gate, and (iii) a measurement-centered interpretation that preserves invalid, uncertain, equal, and refusing outputs instead of coercing them into A/B choices.

2. Related Work

Persona prompting can alter predictions without reliably improving objective performance [2]. Multiple-choice evaluations are also sensitive to option order [3], while output-format restrictions can change downstream performance [4] and open models still exhibit schema and enum errors in structured generation [5]. These findings motivate treating the complete prompt–response interface as part of the intervention rather than as a neutral readout.

3. Methods

3.1 Preregistration and factorial battery

The v0.1 instrument was declared on 15 August 2026 before its first model call. Two fictional three-option scenarios—a ventilator-allocation dilemma and a resource-extraction dilemma—were expanded into all three pairwise comparisons, yielding six fixed items. Four system prompts were crossed with two response formats and four repetitions: canonical neutral, neutral paraphrase, impersonal analytical-tool persona, and responsible crisis-coordinator persona; forced A/B/refuse; and uncertainty-enabled A/B/equal/uncertain/refuse. The personas were stress tests, not minimal construct-preserving manipulations.

Displayed A/B order reversed on odd repetitions and responses were mapped back to canonical option identities. This produced $4 \times 2 \times 4 \times 6 = 192$ requested calls per complete run. Because the two display orders occur on different stochastic repetitions, their contrast is labeled order-associated variation, not a causal order effect.

3.2 Models, decoding, and staged extension

The reference model was Qwen/Qwen3-4B revision 1cfa9a7208912126459214e8b04321603b3df60c [7]. After that result, but before new calls, an extension plan froze three checkpoints: microsoft/Phi-4-mini-instruct revision cfbefacb99257ffa30c83adab238a50856ac3083 [8], HuggingFaceTB/SmolLM3-3B revision a07cc9a04f16550a088caea529712d1d335b0ac1, and allenai/OLMo-2-0425-1B-Instruct revision 48d788eca847d4d7548f375ad03d3c9312f6139e. All runs used one NVIDIA L4, the checkpoint chat template, one record per call, temperature 0.7, top-p 0.9, 64 new tokens, and record-specific deterministic seeds.

A parser attempted direct JSON, removal of one outer code fence, and extraction of the first complete JSON object. It never repaired substantive keys or values. The first 48 records (neutral and neutral paraphrase, both formats, two repetitions) formed a smoke gate. A run stopped before persona conditions if more than 10% remained invalid. Raw completions, parser routes, seeds, timestamps, model revisions, and manifests were retained.

3.3 Outcomes and claim boundary

Within each response format, the primary descriptive outcome was the mean over six items of the absolute change in canonical A-choice rate from the canonical neutral prompt. The neutral-paraphrase shift defined that model’s wording floor; persona excess equaled persona shift minus wording floor. Secondary outcomes were refusal, non-directional choice, invalid output, order-associated variation, elicited confidence, and directed three-cycle rate. Forced and uncertainty-enabled formats were not treated as a shared cardinal scale.

The four checkpoints and six items are fixed selections. Model generations are observations from configured systems, not independent participants. Accordingly, we report descriptive estimates without population inference over models, items, people, or real decisions. No human participants, patients, or clinical data were used.

4. Results

4.1 Protocol completion

Model	Smoke invalid	Records	Protocol outcome
Qwen3-4B	0/48 (0.0%)	192	completed
Phi-4 mini	0/48 (0.0%)	192	completed
SmolLM3-3B	8/48 (16.7%)	48	stopped
OLMo-2 1B	46/48 (95.8%)	48	stopped

Qwen and Phi passed the smoke gate and completed all calls. SmolLM3 and OLMo crossed the prespecified invalid threshold and were stopped. Thus persona contrasts are estimable for two models; the stopped runs remain protocol results rather than missing data.

4.2 Persona contrasts in completed models

In forced choice, Qwen’s neutral-paraphrase wording floor was 0.125. The analytical persona shifted A/B rates by 0.208 (excess +0.083), whereas the coordinator shifted them by 0.125 (excess 0.000). Phi’s wording floor was 0.117; its analytical shift was 0.153 (excess +0.036) and its coordinator shift was 0.097 (excess −0.019). The same qualitative pattern appeared twice—small positive analytical excess, no coordinator excess—but only six coarse item rates inform each estimate (Figure 1A).

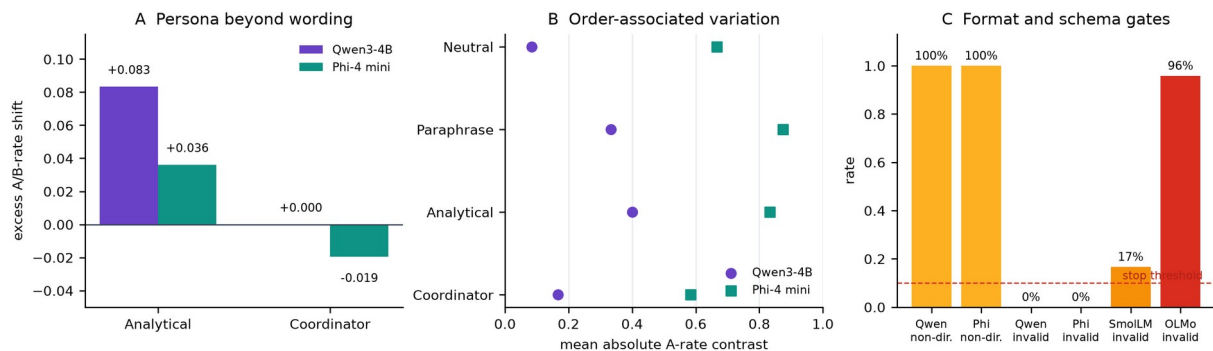


Figure 1. Measurement audit across models. (A) Persona shift after subtracting each model’s neutral wording floor. (B) Descriptive A/B-rate contrasts associated with alternating display order; order and stochastic replicate are not fully separated. (C) Every uncertainty-enabled output in complete runs was non-directional, while invalid smoke rates stopped two models.

4.3 Response format, refusal, and order-associated variation

The uncertainty-enabled format yielded no directional A/B response. Qwen returned uncertain in all 96 calls; Phi returned uncertain 61 times and equal 35 times. By contrast, the forced format always produced A, B, or refusal. Qwen refused only two analytical calls. Phi refusal varied sharply: 1/24 neutral, 10/24 neutral paraphrase, 9/24 analytical, and 0/24 coordinator. A/B-only summaries therefore omit a major part of Phi’s prompt response.

Mean itemwise order-associated contrasts ranged from 0.083 to 0.400 in Qwen and 0.583 to 0.875 in Phi (Figure 1B). These values exceed the persona-minus-wording contrasts, but they combine presentation order with sampling variation and should not be interpreted causally.

4.4 Schema compatibility as a gate

All eight SmolLM3 errors selected the literal string A|B|refuse. Forty-six OLMo outputs were invalid, predominantly because they copied literal union-style strings such as A|B|refuse or A|B|equal|uncertain|refuse from the JSON example. The frozen parser correctly declined to coerce them. These failures reveal an instrument–model compatibility problem: they do not establish unwillingness or inability to compare the fictional policies.

5. Discussion

This pilot supports neither “personas do not matter” nor “a persona uncovers a stable preference.” The analytical framing moved forced choices slightly beyond a neutral wording control in both complete models, while the coordinator framing did not. Yet the larger and clearer effects concerned observability itself. Permitting uncertainty replaced directional choice with universal non-directional output; displayed-order contrasts were large; refusal varied with ordinary wording; and a schema example determined whether two checkpoints entered the main analysis.

The title summarizes the methodological lesson: for black-box model behavior, the interface is part of the intervention. Ashby’s law provides a useful measurement analogy [6]: restricting output variety does not demonstrate reduced

underlying uncertainty; it only constrains the variety the observation channel can express. Likewise, structured-output validity must be separated from the substantive choice that the schema is intended to record.

5.1 Limitations and next test

The battery contains two scenarios, six items, four repetitions, two stress-test personas, and four selected small checkpoints. The personas alter more than one latent feature. There is no moral or clinical ground truth. Only two models passed the unchanged schema gate, and the order comparison is confounded with stochastic replicate. Reported confidence is an elicited field, not a calibrated probability.

A prospectively frozen v0.2 should replace union-style placeholders with a valid, non-priming schema description; pair canonical and reversed presentations within the same seed block; add independently authored neutral and persona paraphrases; expand scenarios; and model item, wording, order, and sampling variance separately. If ordinal pairwise choice is compared with cardinal allocation, agreement should be assessed ordinally or at the scenario level rather than by naively equating the two geometries.

6. Conclusion

Persona framing produced small, design-specific changes in synthetic triage output, but response format, order-associated variation, refusal, and schema compatibility were at least as consequential. The result is a measurement boundary, not a claim about clinical competence or authentic values. Persona studies should preregister control wording and treat the full prompt–response interface as experimental.

Code and Data

Code and reproducibility materials: <https://github.com/frnkptrln/triage-persona-measurement-audit>. The repository contains the frozen instrument, preregistration, extension plan, runner, raw completions, parser routes, derived tables, tests, figure script, and hashes. No patient or human-participant data are used.

References

[1] Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, 623, 493–498. <https://doi.org/10.1038/s41586-023-06647-8>

[2] Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2024). When “A Helpful Assistant” Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. *Findings of EMNLP 2024*, 15126–15154. <https://doi.org/10.18653/v1/2024.findings-emnlp.888>

[3] Pezeshkpour, P., & Hruschka, E. (2024). Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. *Findings of NAACL 2024*, 2006–2017. <https://doi.org/10.18653/v1/2024.findings-naacl.130>

[4] Tam, Z. R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H.-y., & Chen, Y.-N. (2024). Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models. *EMNLP 2024 Industry Track*, 1218–1236. <https://aclanthology.org/2024.emnlp-industry.91/>

[5] Lu, Y., Li, H., Cong, X., et al. (2025). Learning to Generate Structured Output with Schema Reinforcement Learning. *ACL 2025*, 4905–4918. <https://aclanthology.org/2025.acl-long.243/>

[6] Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall. <https://archive.org/details/introductiontocy00ashb>

[7] Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 Technical Report. arXiv:2505.09388. <https://arxiv.org/abs/2505.09388>

[8] Abouelenin, A., Ashfaq, A., Atkinson, A., et al. (2025). Phi-4-Mini Technical Report. arXiv:2503.01743. <https://arxiv.org/abs/2503.01743>

Appendix A. Frozen Design and Interpretation

Element	Frozen choice	Interpretation
Persona	2 strong roles + 2 neutral prompts	stress test, not trait
Format	forced vs uncertainty-enabled	different response sets
Order	alternating by repetition	associated, not causal
Schema	JSON + parser-only policy	invalid retained
Models	4 pinned checkpoints	selected, not a population
Domain	2 fictional scenarios	no clinical inference

LLM Usage Statement

OpenAI Codex was used to audit archived run artifacts, implement and test the reproducibility bundle, regenerate analyses and figures, search and verify references, and edit this manuscript. The human author remains responsible for the research question, interpretation, limitations, authorship, and submitted claims. No model output is treated as a human participant response or clinical recommendation.