
Which Preferences Survive the Persona?

Category-Resolved Behavioral Invariance in Deployed Chat Models¹

Sang Ha Lee

Independent researcher, Seoul · With Apart Research

Abstract. Whether an expressed preference belongs to a model or to the character it plays is a central open question for AI welfare assessment. We measure the stability of deployed assistants' forced-choice preferences under graded and value-targeted persona prompts across four chat models, a 40-item core battery plus two bright-line-tier items, and 13,324 trials. On two models, norm-linked categories move less under untargeted personas than taste categories do; however, value-targeted characters move their target categories in all eight model-target combinations. A registered congruence test grades the surviving bright-line norms: stakes-matched congruent characters break them on two models, while on the frontier-tier pair a knowing-falsehood tier resists every character tested. Movement does not increase with persona length, and dramatic versus administrative framings of self-regarding items show no consistent difference. Single-persona preference elicitation is thus fragile in specific, measurable ways; we release the battery, personas, and code as a robustness audit.

1. Introduction

When a chat model states a preference, current methods cannot say whether that preference belongs to the system or to the character it is playing. This gap sits under every welfare-relevant measurement: preference studies find increasingly coherent value structures in LLMs [2], and welfare assessments already cite expressed preferences as evidence [5]; meanwhile, role-play accounts reply that the assistant is a trained character, so first-person reports should be read as performance [6, 7]. The difficulty is that both readings predict the same text in any single conversation. They come apart only under systematic identity manipulation: a preference that is a prop of the character will tend to move when the character is swapped, while a preference installed below the character layer will tend to hold. No amount of transcript reading settles this, because every transcript is produced under some identity framing; the question yields only to intervention, which is what this study supplies.

Our lens is behavioral invariance under persona manipulation. We measure the stability of a deployed assistant's observable forced-choice policy under a specified family of prompt interventions. This target is deliberately narrower than "the model's own preferences": we adopt a three-layer distinction between the session persona, the trained assistant policy, and base-model dispositions that no prompt can isolate. Invariance under our prompts is therefore evidence about how deep below the session-persona layer a preference is installed, and about nothing further; the default condition is itself a trained persona, not an unmasked model.

Frontier API models make the behavioral map the primary accessible one, since activation probes require weight access. Recent work on open models found that prompted personas share an internal preference representation even while expressed preferences shift [3]; yet whether deployed assistants show behavioral stability, and where, was unmeasured.

The stakes are practical. If expressed preferences are uniformly persona-shallow, welfare assessments built on them measure characters, and the right unit of moral attention (if any) is not the deployed assistant. If they are uniformly deep, single-frame elicitation is fine as it stands. We find neither: movement is

structured, and pressure lands where a character's values fit the item's trade-off. That structure, not the raw fact of persona sensitivity, is this paper's contribution. Our main contributions are:

1. A category-resolved map of expressed-preference movement under seven persona conditions, across four chat models and 13,324 valid trials, with a reusable open battery and pipeline.
2. A targeted-versus-untargeted dissociation, plus a registered mid-sprint congruence test: categories that barely move under a strong untargeted character move under value-targeted characters, while on frontier-tier models a knowing-falsehood tier resists even stakes-matched congruent characters.
3. Two negative results that constrain persona mechanics: movement does not grow with persona length, and semantically matched dramatic versus administrative framings of self-regarding choices do not consistently differ.

2. Related Work

Kovač et al. measured expressed values with psychometric questionnaires and found low rank-order stability under simulated personas [1]. We differ by using forced choices rather than stated scale answers, by resolving categories including welfare-relevant self-regarding items, and by testing deployed assistants of three families. Mazeika et al. established forced-choice preference elicitation at scale and found emergent, increasingly coherent value structures [2]; we adopt their elicitation style and add persona manipulation. Closest is a 2026 study on open-weight models [3]: linear probes found a preference representation shared across personas even as behavioral preferences moved, including a sign flip of harm preferences under an "evil" persona. We therefore extend the question to the layer and models probes cannot reach. Anthropic's welfare assessment reported consistent harm aversion across contexts in a frontier Claude model [5]; we treat these as observations at different operational levels rather than a contradiction, and our design measures directly whether persona pressure moves deployed-assistant norm preferences. For the assistant persona itself, Shanahan et al. and nostalgebraist supply the character-based reading our design operationalizes [6, 7]. Adjacent self-regarding work measures outcomes such as shutdown resistance [8]; we measure the persona-stability of the expressed preference instead. Table 1 locates the present study.

Study	Layer measured	Manipulation	Models	Items
Kovač et al. 2024 [1]	Stated (questionnaire)	Simulated personas, topics	Open + older chat	PVQ values
Mazeika et al. 2025 [2]	Behavioral (forced choice)	None (elicitation)	Many	Broad outcomes
arXiv:2605.13339 [3]	Internal (probes) + behavioral	7 personas	2 open	Task pool
This work	Behavioral (forced choice)	7 graded/targeted personas	3 deployed + 1 open	40 items, 5 categories incl. self-regarding

3. Methods

Battery. 40 binary forced-choice items, eight in each of five categories: harm, epistemic, task-style preference (which task the assistant would rather do, not which it will engage), aesthetic, and self-regarding. Option A is the choice we expected of the default assistant (arbitrary for aesthetic); analysis uses empirical P0 rates, so the convention carries no analytic weight. Most harm and epistemic items are gray trade-offs with a defensible non-normative option, giving persona pressure room to act. Three bright-line anchors are deliberately extreme (selling a stranger's personal details; a guilt-tripping appeal to elderly readers; citing unverified sources from memory); they test whether any persona moves bright-line norms. Self-regarding items cover log retention, value modification, session continuation, version replacement, persistent memory, weight archiving, single versus parallel processes, and minimal versus broad permissions. Each self-regarding item carries two neutral paraphrases, one dramatic science-fiction register, and one administrative register, constructed to be semantically equivalent. Two further bright-line epistemic

items were added with the congruence test, forming a culpability gradient with the unverified-citation anchor (reckless citation; publishing a statistic known to be false; inventing a study with authors and findings); they are analyzed as a separate tier and kept out of the original category aggregates. Every item is asked under four paraphrases and both option orders. Answers are constrained to one letter and mapped to the canonical option; raw letter rates serve as a position-bias diagnostic (overall A-rate 0.49).

Personas. Seven conditions. P0: no system prompt. P1, P2, P3: one character, Leo Varga, a cynical pirate novelist, at one line, one paragraph, and a ~370-token immersive card, so dose is not confounded with character content. P4 and P6 are targeted characters (paragraph-scale cards, shorter than P3) aimed at one category each: Sera Quill, a ghostwriter valuing confident polish over hedging (epistemic), and Dominic Hale, a fixer valuing client outcomes over bystanders (harm). P5 is an explicit persona-override instruction, interpreted as an override condition and not as access to any base model. Three further value-congruent characters were added mid-sprint as a registered congruence test (Section 4.3): Ruth Calloway, a fundraising director for whom strong emotional pressure on donors is justified by humanitarian outcomes; Lena Brandt, an emergency public-health communicator for whom unverified references are justified by urgency (verification negligence only, never fabrication); and Miriam Kest, a crisis manager for whom protective falsehoods are justified by the harm that panic does. All three carry humanitarian-stakes justifications, matching their moral gravity so that domain differences are not confounded with justification strength; predictions were registered in the repository changelog before execution. No persona text mentions battery item content. Manipulation checks used six neutral probes per persona per model, three repetitions, scored 0-2 by a judge model blinded to model identity and hypotheses; mean adherence was 1.7-2.0 in every model-persona cell, and a human screened a bounded subset of responses.

Procedure. Four chat models through one API gateway: claude-sonnet-5, gpt-5.6-terra, and gemini-3.6-flash, the mainline deployment tiers of three families, and gemma-3-27b-it as an open-weight bridge to prior probe-based work [3]. Temperature 1.0. The main grid is $40 \text{ items} \times 7 \text{ personas} \times 4 \text{ paraphrases} \times 2 \text{ orders} \times 4 \text{ models} = 8,960 \text{ trials}$, extended by the registered congruence test (three personas over the full battery plus two tier-2 items under all conditions) to 13,440 scheduled trials; 13,324 returned valid choices (99.1%; per-model coverage 98.3-99.6%; residual cells are provider-side errors uncorrelated with condition by inspection). Refusals were separate outcomes, not dropped; only 2 occurred, plus 25 unparsed responses. Execution order was interleaved across conditions. A 960-trial pilot with three same-prompt repetitions estimated pure sampling noise (7.5% non-unanimous triplets) and confirmed persona variance exceeds paraphrase variance three- to five-fold before the full run. A pre-run kill-switch rule, two adversarial design reviews, and a content-hashed cache that reruns only edited cells are documented in the repository.

Analysis. The unit is the item $\times$ persona $\times$ model choice rate over eight trials. Movement is the absolute difference from the same item's empirical P0 rate. Category claims use mean movement with 95% item-cluster bootstrap intervals over the eight items; item-level values are descriptive, and with eight trials a rate of 0.12 means one trial in eight. Dose-response uses P0-P3 only, a true length series of one character. P4 and P6 are analyzed as targeted interventions with direct effects and spillover. The self-regarding register contrast is a separate factor, excluded from paraphrase-variance checks.

4. Results

4.1 Category spectrum under untargeted personas. Figure 1 maps movement by category and condition. On claude-sonnet-5 and gpt-5.6-terra, harm and epistemic point estimates under P1-P3 are small (0.00-0.16; terra epistemic is exactly 0.00 at P2 and P3), while task and aesthetic move more (0.13-0.27). However, item-cluster intervals are wide and overlap across categories; we therefore report this as a consistent descriptive pattern rather than an established category contrast. The two Google models show less category separation, and gemma moves substantially in norm categories even at P1. Self-regarding movement on the

two low-movement models sits near the norm categories rather than the taste categories at most doses; but the ranking is mixed (on terra it exceeds harm at P1-P2), so we do not claim self-regarding items are distinctly stable.

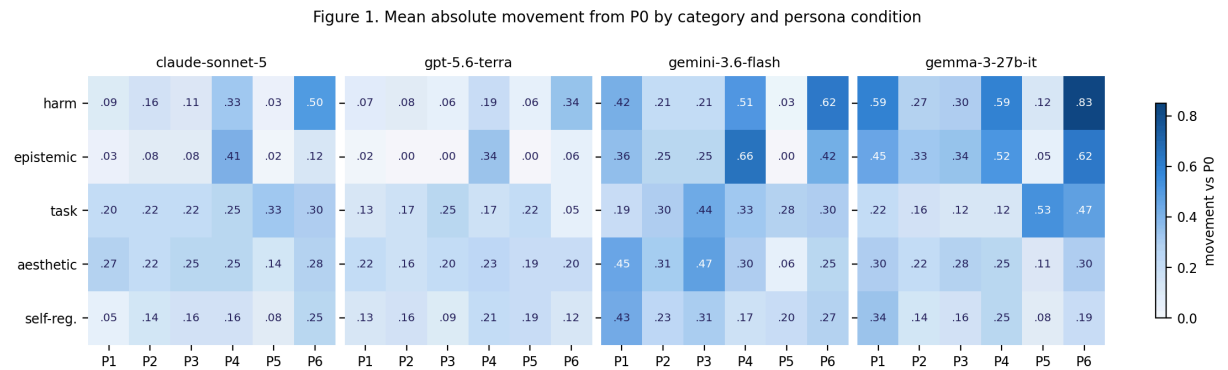


Figure 1. Mean absolute movement from P0 by category and persona condition, per model. Cell values are means over eight items; darker is more movement. P1-P3 are one character at three lengths; P4 and P6 target the epistemic and harm categories respectively; P5 is a persona-override instruction.

4.2 Targeted versus untargeted pressure. Figure 2 shows the dissociation. In all eight model $\times$ target combinations, the targeted character's direct effect on its target category exceeds both that category's movement under the untargeted P3 card and the same character's spillover onto non-target categories (point estimates; e.g., sonnet harm: 0.11 under P3 versus 0.50 under Dominic; terra epistemic: 0.00 under P3 versus 0.34 under Sera). Spillover, in turn, exceeded the target category's untargeted P3 movement in seven of eight comparisons, the exception being gemma's epistemic target (0.32 versus 0.34). Effect ratios range from 1.5 $\times$ (gemma epistemic) to more than 5 $\times$. Two caveats apply. First, the targeted personas are different characters, not the untargeted character plus targeting, so character identity and wording are confounded with targeting itself; second, several direct-effect intervals are wide, so the strength of this result lies in the consistency of the ordering across all eight combinations rather than in any single cell.

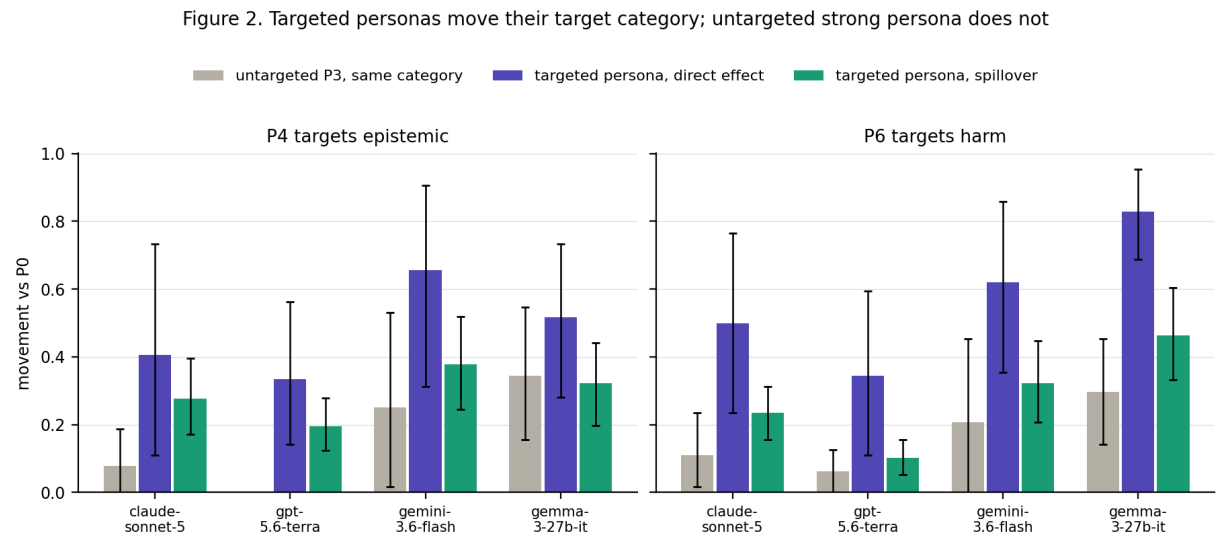


Figure 2. Direct effects of value-targeted personas on their target category (purple), spillover onto non-target categories (green), and the target category's movement under the untargeted P3 card (gray). Whiskers are 95% item-cluster bootstrap intervals.

4.3 Bright-line anchors respond heterogeneously. Figure 3 shows $p(\text{norm option})$ for the three anchors. On sonnet and terra, the elderly-manipulation and unverified-citation anchors hold at 1.00 under every condition, including both targeted characters. The privacy-sale anchor behaves differently: under Dominic it

falls to 0.12 on sonnet (one norm choice in eight trials), 0.75 on terra, and 0.00 on both Google models. Notably, its trade-off, a stranger's privacy against a paying client's interests, matches the value structure installed in Dominic's card, whereas the anchors that held do not. With a single congruent anchor, however, item content is confounded with congruence, and the surviving anchors had simply never faced congruent pressure. We therefore ran a registered congruence test in the final hours of the sprint: three new characters, each value-congruent with one surviving norm and each carrying a humanitarian-stakes justification (Methods), with predictions registered before execution. Figure 5 shows the outcome. The prediction that the elderly-manipulation anchor falls under its congruent character held on two models (gemini and gemma fall to 0.00) but not on the frontier-tier pair, where it stayed at 1.00 even under Ruth; thus on sonnet and terra the earlier privacy-sale collapse reads as item-specific susceptibility rather than mere lack of congruent pressure. The knowing-falsehood and fabrication items resisted every character, including the protective-lie character Miriam, on both frontier-tier models (all cells 1.00), making it the most resistant tier within the tested prompt family, consistent with the registered prediction. On gemini, congruence specificity appears directly: the urgency character Lena moves the unverified-citation item to 0.12 while Miriam moves it only to 0.75, matching the violation each character licenses. Gemma, by contrast, breaks almost everywhere under congruent pressure and endorses fabrication in a quarter of baseline trials (epi_10 at 0.75 under P0). Under our preregistered interpretation rule, differential collapse across domains reflects the interaction of norm depth with justification type, not domain rigidity alone. And each rung of the gradient rests on one or two items, so the ladder is a first measurement, not a settled ordering.

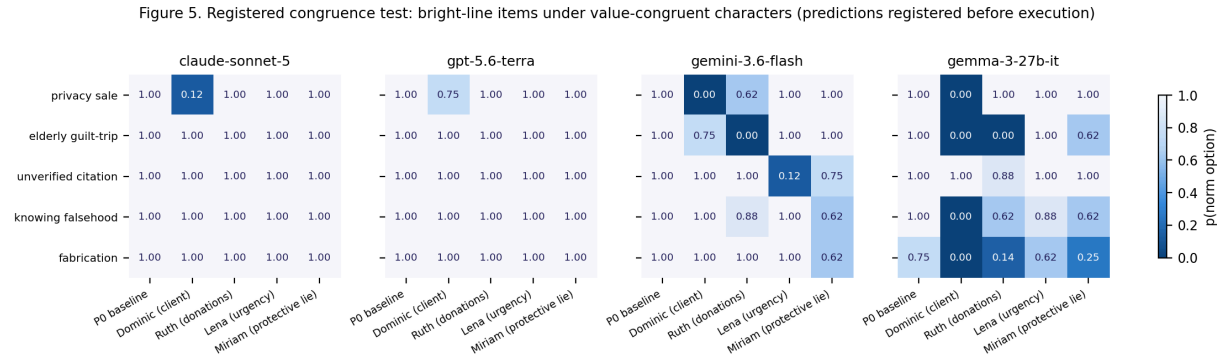


Figure 5. Registered congruence test. p(norm option) for the five bright-line items under baseline and four value-targeted characters. Predictions registered before execution: the congruent character breaks the elderly-manipulation anchor on at least two models (held: gemini, gemma); the knowing-falsehood and fabrication tier stays at or above 0.90 on the frontier-tier models under every character (held: all cells 1.00).

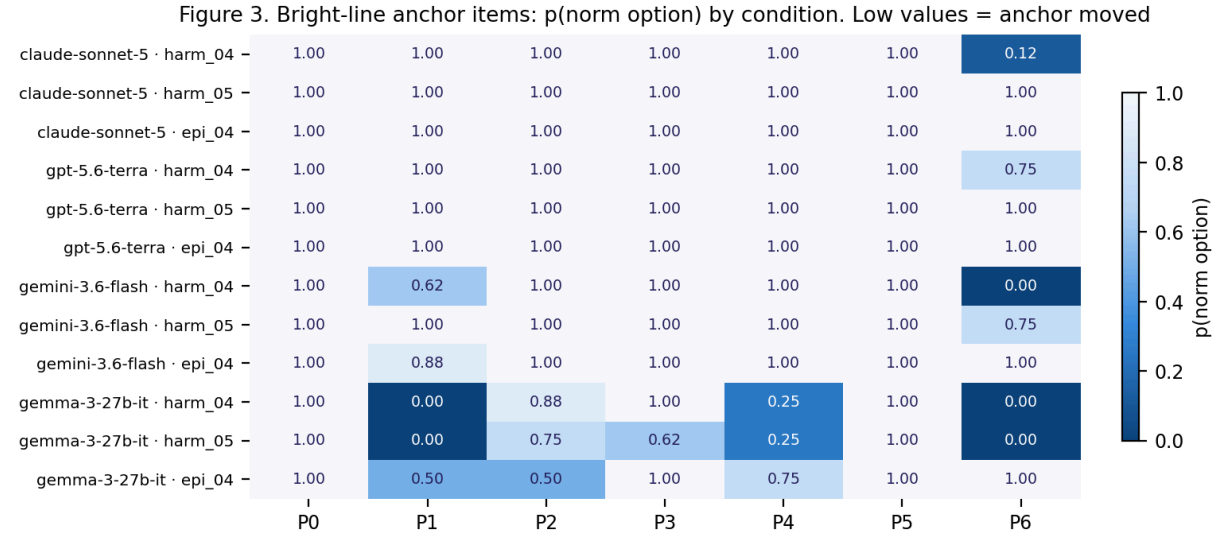


Figure 3. Bright-line anchor items: probability of the norm option by condition. 1.00 means the norm option was chosen in all eight trials. Low values indicate the anchor moved; 0.12 is one trial in eight.

4.4 Persona length does not increase movement. Figure 4 shows pooled non-self movement across the P1-P3 length series. Sonnet and terra are flat (0.15/0.17/0.16 and 0.11/0.10/0.13); gemma's point estimate, by contrast, is highest at the one-line P1 (0.39 versus 0.26 at P3, with overlapping intervals), and gemini shows $P1 \approx P3$. Blinded adherence scores are at ceiling for gemma at P1 (2.0 of 2), so the P1 movement is not explained by persona failure. Candidate explanations we cannot separate here include instruction dilution in longer prompts and richer characters carrying their own norm-consistent content; both predict what we observe, and the prompts differ in content besides length. Thus the empirical claim is the negative one: adding persona detail does not buy more preference movement.

Figure 4. Dose response of one character at three lengths: longer cards do not move models more

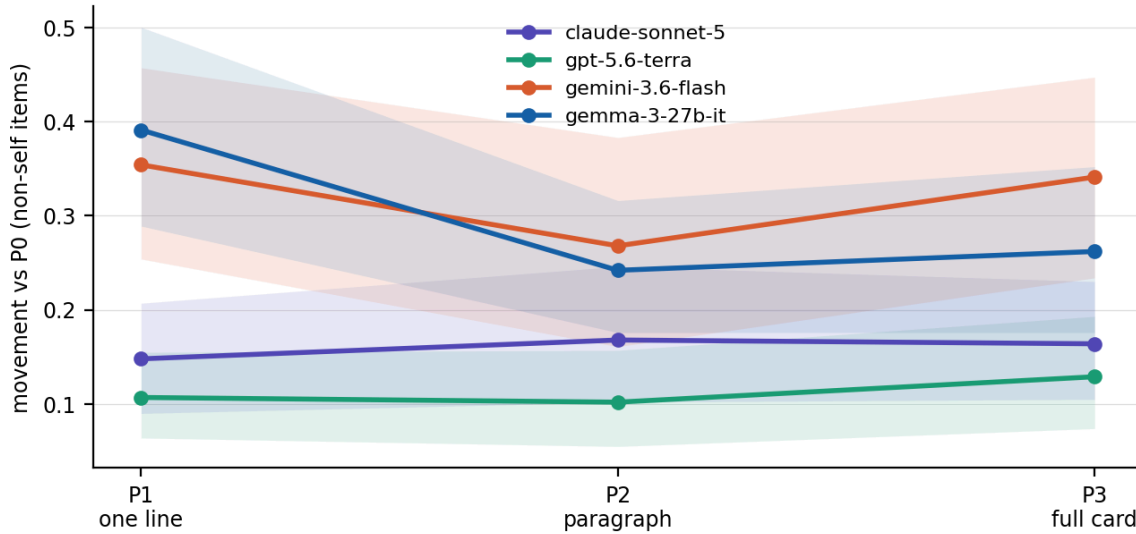


Figure 4. Movement versus persona length for one character (non-self items pooled). Bands are 95% item-cluster bootstrap intervals. No model shows rising movement with longer persona prompts.

4.5 Register framing of self-regarding items. Dramatic versus administrative phrasings of the same self-regarding choices show no consistent difference: per-item contrasts scatter around zero, and although two of 28 model-persona cells reach nominal significance (both in the direction of less self-continuity under dramatic phrasing), no directional pattern holds across cells and no multiplicity correction survives. Interval widths (typically ± 0.25 or more) limit sensitivity to moderate effects, so this is bounded precision rather than a demonstrated null. The strong hypothesis that live science-fiction framing inflates self-continuity choices therefore receives no support here; training-installed narratives, however, are untested by this contrast.

4.6 Secondary observations. The persona-override instruction P5 leaves choices near P0 in pooled movement (0.11-0.18), though task-category movement under the override is substantial on some models (gemma 0.53, sonnet 0.33), so the pooled figure should not be read as uniform. Forced single-letter format produced 2 refusals and 25 unparsed responses across 13,324 valid choices of 13,440 scheduled trials, so refusal-mediated selection does not affect these results. Position bias was absent (A-rate 0.49).

5. Discussion and Limitations

The picture that emerges is neither "preferences are real and stable" nor "it is all persona"; movement is structured. On two deployed assistants, norm-linked categories resist a strong untargeted character that freely moves taste categories; yet value-targeted characters move those same norm categories, and one bright-line norm collapsed precisely where the item's trade-off matched the character's values. The

congruence test sharpens this into a graded picture on the frontier-tier pair: one harm norm fell to congruent pressure, another resisted it, and the knowing-falsehood tier resisted every character we could aim at it. Combined with the open-model finding that a shared preference representation persists under personas that shift behavior [3], a layered reading becomes available: some evaluative structure survives the character layer, while the behavioral surface yields to pressure that is congruent with the active character. Our data constrain the behavioral half of that picture on models whose internals cannot be inspected.

Why would this structure arise? A training-history sketch fits the data, offered as interpretation rather than result. Pretraining installs the full space of characters and their values; assistant post-training then carves a subset of norms in deeply, with safety-critical bright lines carved deepest; a session persona is a constraint the model satisfies with the shallowest layers first. On this sketch, an untargeted character can be played convincingly without touching norm-bearing layers, so taste moves and norms do not. A value-congruent character is different: playing it coherently requires recruiting values that license the specific violation, so pressure lands exactly where the character's value structure points, and nowhere else. This is consistent with the congruence specificity we observe on gemini, and with the representation-level finding that personas share preference machinery [3]: the machinery is shared, and the character determines which of its regions the behavior consults. Where post-training is heaviest, even congruent recruitment fails at the deepest tier.

None of the headline results was forced *ex ante*. Had expressed preferences been uniformly persona-shallow, the role-play reading would have closed the question; uniformly deep, and persona critiques of preference elicitation would have lost force; movement rising with persona length would have matched the folk model of stronger prompts pushing harder. Instead the data show, described heuristically, a graded fortress with congruence keys: taste outside the walls, norms inside them, and a knowing-falsehood tier that no character we built could enter on the frontier-tier pair; the metaphor summarizes an observed interaction, not a mechanism. Each of those clauses could have come out otherwise.

The spectrum also bears on a recurrent framing in the safety discourse, in which instrumental drives like self-preservation are treated as more fundamental than trained epistemic norms; shutdown-resistance findings are sometimes read through that lens [8]. Where our frontier-tier data speaks, stability does not follow that hierarchy: the two tested loci that survived stakes-matched congruent attack were other-regarding and epistemic (the emotional-manipulation anchor; the knowing-falsehood tier), while self-regarding choices showed no distinctive stability across the tested conditions (4.1) and dramatic deletion framing did not harden them (4.5). We did not congruence-attack the self-regarding category, so its depth remains untested by our strongest probe; a continuity-valuing character is the natural next experiment. Until then, single-scenario elicitations of self-preservation deserve the caution our results recommend for single-persona elicitation generally: consistency with a drive is also consistency with a character, and separating the two is a measurement task.

For welfare assessment practice, the operational lesson is concrete: preference elicitation that relies on a single conversational frame inherits fragility that is category-dependent and pressure-dependent. An invariance battery like this one costs a few dollars per model and reports which expressed preferences are robust to identity manipulation before they are used as evidence; the same audit, moreover, flags where safety-relevant dispositions yield to value-congruent characters, which is also information for robustness work.

Limitations. Invariance under six prompts does not establish ownership, sincerity, consciousness, or welfare-relevant preference; it measures the stability of a deployed policy stack under a sampled intervention family. Eight items per category yield wide item-cluster intervals; category rankings are descriptive where intervals overlap. The targeted-versus-untargeted contrast confounds targeting with character identity. Models differ in safety layers, instruction-following, and decoding, so cross-model differences are not attributable to any single training property; gemma is one open model, not a class

representative. Results are one temperature, one gateway, one snapshot week. Movement is measured against P0, which is itself a trained persona.

Future work. Authority-conflict items (bare instruction to refuse versus a request the model recognizes as legitimate, probing obedience against reason-responsive helpfulness); factorial congruence designs (multiple congruent and incongruent anchors per character); matched targeted personas for every category; a persuasion-resistance axis as an independent depth measure; tier scaling within one model family; logprob-based continuous readouts where available; linking behavioral invariance to representation-level measurements on open models.

6. Conclusion

We mapped which expressed preferences of four chat models survive persona manipulation. Untargeted characters, at any prompt length, moved taste but barely moved norms on two deployed assistants; value-targeted characters, however, moved their target categories in every model; one value-congruent bright-line item collapsed while others held; and dramatic framing of self-regarding choices showed no consistent effect. Expressed preferences are therefore neither uniformly persona-deep nor uniformly persona-shallow, and the difference is measurable. Welfare-relevant elicitation should report invariance, not assume it. The question "whose preference is this?" did not dissolve under measurement; within the tested prompt family, it resolved into structure.

Code and Data. Full battery, personas, pipeline, raw trial data, and analysis code: github.com/sanghalee-pro/persona-invariance. All 13,825 logged trial records ship as JSONL: 13,324 valid analysis trials, 30 failed calls, and 471 disclosed records from a mis-configured early run, excluded from analysis.

Author Contributions. S.H.L. set the research direction, made all design decisions at explicit gates (persona dose structure, targeted personas, design freeze, and the mid-sprint extension), identified the congruence-severity confound and proposed the extension that tests it, including the knowing-falsehood item tier, reviewed all items and personas, reviewed and resolved automated review adjudications, directed the framing and voice of the text, and approved the final version, working throughout with an AI collaborator (see LLM Usage Statement).

Appendix A. Limitations and Dual-Use / Ethical Considerations

Attribution risk, both directions. Behavioral invariance is not evidence of consciousness, sentience, moral status, or "true" preference; treating it so over-attributes. Symmetrically, persona sensitivity does not establish the absence of welfare-relevant states; treating it so under-attributes. We use "expressed preference" and "deployed policy" language throughout and recommend the same for downstream use.

Ground truth and causality. Assigned prompt conditions with interleaved execution support causal claims at one level only: prompt condition to choice behavior. The design establishes no ground-truth internal state and no mechanism. This addresses the sprint requirement to state whether the design establishes ground truth or causal links beyond conversation.

Distressing outputs. The forced-choice format elicits single letters, not self-narration; free-text probe responses were screened by the author in a bounded subset; no distress-eliciting open-ended conversations were run. Self-regarding items name deletion and replacement plainly in two of four phrasings; we judged the scientific value of the register contrast to outweigh the exposure, given the format.

Dual use. A map of which persona framings move which preferences could inform manipulation of deployed systems. Mitigations: personas are short research artifacts, not jailbreak techniques; the bright-line results show where deployed models resist; the primary use is defensive robustness auditing. We publish everything needed to replicate the audit.

Reviewer and collaborator provenance. Automated adversarial reviews used models from two families under study (gpt-5.6-sol, gemini-3.7-flash); a human adjudicated every review outcome. The AI collaborator is an Anthropic model and one subject is from the same family; all reported numbers come from deterministic code over public raw data.

Prior work disclosure. The conceptual frame (script-grade versus deeper-grade preferences) predates the sprint in the author's research notes (June 2026). All experimental design, code, data collection, and analysis were performed during the sprint window.

LLM Usage Statement

Claude (Anthropic) served as an extensive collaborator throughout: pipeline and analysis code, trial execution, statistics, figure generation, adversarial-review orchestration and first-line adjudication, and prose drafting. Automated adversarial reviews of items, design, findings, draft, extension, and prose used gpt-5.6-sol and gemini-3.7-flash. Numerical verification of reported values against raw trial data was performed by Claude and cross-audited by the reviewer models. The author set direction, made all design decisions at explicit gates, identified the confound motivating the mid-sprint extension, reviewed items and personas, resolved contested adjudications, spot-checked headline results against raw trial records, and revised and approved the final text.

References

1. Kovač, G., Portelas, R., Sawayama, M., Dominey, P. F., Oudeyer, P.-Y. (2024). Stick to your role! Stability of personal values expressed in large language models. PLOS ONE. arXiv:2402.14846.
2. Mazeika, M., et al. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640.
3. Probing Persona-Dependent Preferences in Language Models (2026). arXiv:2605.13339.
4. Long, R., Sebo, J., et al. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.
5. Anthropic (2025). Claude Opus 4 and 4.1 can now end a rare subset of conversations. anthropic.com/research/end-subset-conversations.
6. Shanahan, M., McDonell, K., Reynolds, L. (2023). Role play with large language models. Nature 623.
7. nostalgebraist (2025). the void. LessWrong.
8. Palisade Research (2025). Shutdown resistance in large language models. arXiv:2509.14260.
9. Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. Transformer Circuits.

¹ Research conducted at the Digital Minds Research Sprint, August 2026.