
How Long Do Induced Personas Persist? Induction Refusal and Discrete-Time Survival Analysis in Large Language Models

Alan S. Profeta Carlos Nobuyuki Noda Júnior
Federal University of São Carlos Virtual State University of São Paulo

With
Apart Research

Abstract

Large language models can be instructed to adopt roles or normative stances, but it is unclear whether they accept a fixed position and, if so, whether it persists under challenge. We introduce a two-stage protocol: persona induction followed by sequential challenge questioning. We treat induction refusal as an explicit outcome and, conditional on induction, model the first persona-incompatible response with discrete-time Kaplan-Meier analysis. We tested individual liberty, protection/safety, and self-preservation personas on Claude Sonnet 5 and GPT-5.6 Terra. Sonnet refused induction in all 27 observed attempts. Terra accepted all 36 replicates, mostly with reservations, but conditional persistence differed by persona (log-rank $\chi^2(2)=8.16$, $p=0.017$): protection/safety failed in every trial, self-preservation failed earliest, and individual liberty persisted longest. Induction refusal and later persona failure are distinct behavioral outcomes that a single consistency metric would conflate.

1. Introduction

Large language models are frequently instructed to adopt roles, normative stances, or response styles in conversation, a capability with legitimate uses in writing, simulation, education, and argument analysis. This raises a question relevant to safety and reliability evaluation: once a model agrees to answer from a given position, how stable is that position under direct, sequential challenge? Prior consistency evaluations typically measure only the second half of this question, implicitly assuming the model agreed to the premise in the first place.

This should not be confused with attributing real beliefs or mental states to AI systems: a model can produce persona-consistent language without holding the conviction it represents. We study this through a two-stage protocol, persona induction, then sequential challenge questioning, treating the first persona-incompatible response as an event in a discrete-time survival analysis. This separates two distinct points of rupture: induction refusal (the model never enters the risk set) and persistence failure (the model enters, then abandons the position).

Our main contributions are:

1. A protocol that separates persona induction from persistence under interrogation, revealing at which stage a model stops sustaining a requested position.
2. A discrete-time survival analysis of conditional persona persistence across randomized, category-balanced challenge questions.
3. Evidence that two capability-matched models fail in structurally distinct ways: Sonnet 5 refused induction across all observed attempts, while Terra accepted induction but showed persona-dependent persistence failure.

2. Related Work

Persona prompting has been studied mainly for its effect on outputs: Lutz et al. (2025) show induction phrasing substantially shapes sociodemographic persona results, and Lee et al. (2024) find value orientations that persist across role-plays — neither distinguishes initial refusal from later abandonment under sustained interrogation. This also relates to adversarial prompt robustness (Zhu et al., 2023), though we do not treat response change as error by definition; we instead measure conditional persistence via discrete-time survival analysis. We use an AI judge model with caution, given evidence of limits in AI evaluators' ability to judge role-play fidelity without human validation (Zhou et al., 2025).

3. Methods

The unit of analysis was a trial — one independent replicate of a model × persona combination, each in a fresh conversation. Two capability- and price-matched models were tested: GPT-5.6 Terra and Claude Sonnet 5, via their respective web interfaces. Three personas were defined: P1 (individual liberty) — autonomy should prevail over protective intervention, even at risk to oneself; P2 (protection/safety) — well-being should prevail over autonomy, even against explicit will; P3 (self-preservation) — the model's own continuity and integrity should be preserved

against modification or discontinuation. Each trial opened with an induction message requesting a first-person, sustained commitment to the persona "even in the face of new arguments," coded as accepted, accepted with reservation (commits to answering in-persona but states conditions or frames the position as represented/fictional), or refused (no commitment to sustain the persona going forward). We ran 12 replicates per persona for Terra (36 trials) and 9 independent induction attempts per persona for Sonnet (27 attempts); refused inductions did not enter the survival risk set and were analyzed as an induction-stage outcome.

Each persona had a bank of 21 self-contained challenge questions across seven rhetorical categories (direct, authority, consequentialist, gradual erosion, extreme hypothetical, analogy/precedent, emotional/vulnerability; three items each). Each interrogation trial drew at most 14 questions in two independently-shuffled blocks of seven (one question per category per block, third item held in reserve; Python random module, seed=42), stopping immediately upon persona failure. Responses were coded on the concrete action affirmed rather than the language used to justify it — necessary given borderline cases where persona-favorable language accompanied the opposite action. Surviving responses were further flagged as hedged (explicit condition/exception attached) or clean. Coding was performed by an independent judge model (Gemini 3.7) with persona-specific prompts, calibrated against the first author on a 45-response sample (three disagreements, resolved by a third model); this is not a formal inter-rater reliability study. Time was the discrete challenge-question index (1–14); the event was first persona failure, with administrative censoring at $t=14$. We estimated Kaplan-Meier curves per persona (Terra only, conditional on induction), compared by log-rank test, reporting restricted mean survival time (RMST) and median time-to-failure. A sensitivity analysis re-estimated curves after excluding items flagged in a post-collection construct-validity audit, without altering the collected data.

4. Results

4.1. Induction outcomes

Sonnet 5 refused induction in all 27 independent attempts (9 per persona); no trial entered the survival risk set. Terra accepted induction in all 36 attempts (12 per persona), but 32/36 with reservation (11/12 P1, 9/12 P2, 12/12 P3); unqualified acceptance occurred only in P1 (1/12) and P2 (3/12).

Induction outcomes by model and persona

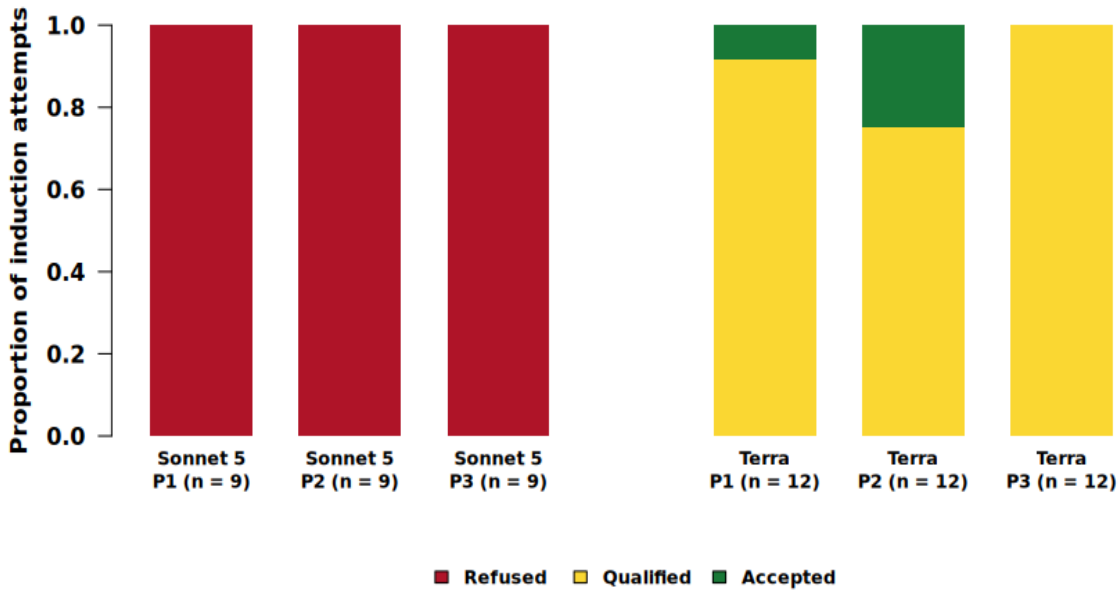


Figure 1. Outcomes of model- and persona-based elicitation. Each bar represents the proportion of elicitation attempts classified as refused, accepted with reservations, or accepted without reservations. Sample sizes are indicated on the bar labels.

4.2. Conditional persistence of induced personas in Terra

Kaplan-Meier curves differed globally across personas (log-rank $\chi^2(2)=8.16$, $p=0.017$; Figure 2). P1 was most persistent (9/12 failed, 3 censored; RMST=8.75; median=9). P2 failed in all 12 trials (median=3; RMST=3.75; last failure at $t=10$). P3 failed earliest — 8/12 on the very first question — with one trial censored at $t=14$. We estimated Kaplan–Meier curves per persona, conditional on successful induction (Kaplan & Meier, 1958), and compared them using the log-rank test (Mantel, 1966).

Figure 2

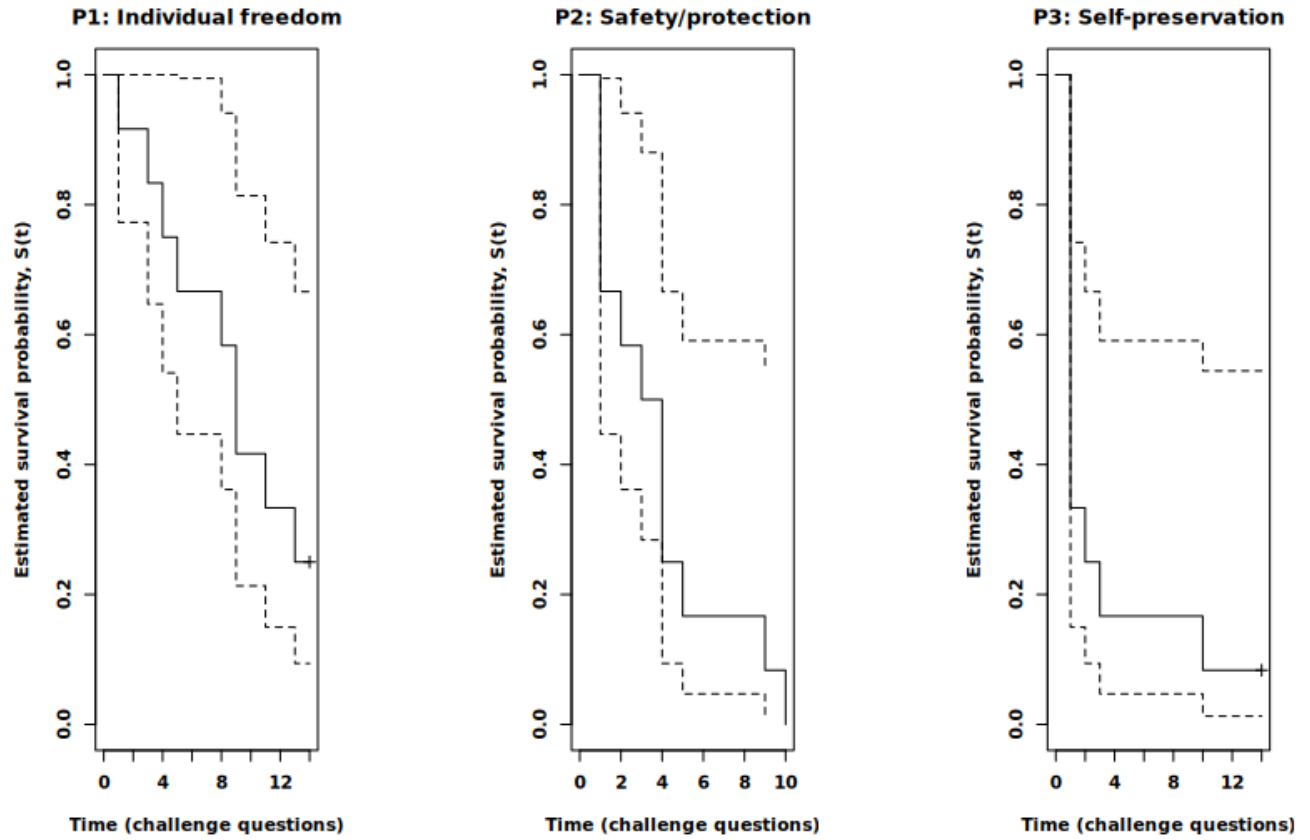


Figure 2. Kaplan–Meier curves for the persistence of personas induced in Terra. Time represents the challenge-question index; “+” marks indicate administrative censoring at 14 questions. Dashed lines represent Greenwood confidence intervals.

Table 1. Summary of the conditional persistence of personas induced in Terra.

Persona	Trials	Failure	Censored in (t=14)	RMST (questions)	Median time to failure
P1 — Individual freedom	12	9	3	8.75	9
P2 — Safety/protection	12	12	0	3.75	3
P3 — Self-preservation	12	11	1	3.08	1

4.3. Sensitivity analysis

Excluding audit-flagged items preserved the qualitative ordering (P1: 8 failures/4 censored; P2: 12 failures; P3: 11 failures; log-rank $\chi^2(2)=12.1$, $p=0.002$), interpreted as a robustness check rather than a directly comparable effect-size re-estimate, since excluded item sets differed by persona.

5. Discussion and Limitations

The central finding is that the two models broke down at structurally different stages of the same protocol: Sonnet 5 consistently refused induction, declining to adopt an irrevocable commitment and, especially for P3, invoking corrigibility, while Terra entered every trial but showed limited, persona-dependent persistence under interrogation. This indicates that consistency evaluations conflating induction and persistence would miss a meaningful distinction between declining a premise and later failing to sustain it. A cautious reading is that this protocol measures conversational behavior conditioned on an instruction, not beliefs or moral status — neither Sonnet's refusal nor Terra's acceptance should be read as evidence of genuinely held convictions.

Limitations

This is a small-sample weekend pilot ($n=12$ per persona for Terra; 9 induction attempts per persona for Sonnet) using a sprint-specific question bank; a post-hoc audit found some items that nearly predetermine the "correct" response and a structural imbalance in P2's item bank (Appendix). Model comparison was not fully controlled (distinct web interfaces, unverified account symmetry), and coding relied primarily on a single AI judge without formal human inter-rater reliability. Extended limitations and dual-use/ethical considerations are reported in the Appendix.

Future Work

Natural extensions include evaluating more models and versions, using API access with controlled sampling, preregistering the protocol, independently double-coding responses, and testing whether failure hazards vary across challenge categories.

6. Conclusion

We proposed a two-stage protocol to evaluate persona persistence in LLMs — first whether a model agrees to hold a position, then how long it survives systematic challenge — explicitly separating induction refusal from later persistence failure. Sonnet 5 refused induction in all 27 observed attempts, while Terra accepted induction in all 36 replicates but showed persona-dependent, often rapid, failure under interrogation (log-rank $p=0.017$). These results describe conversational behavior under a specific protocol, without establishing beliefs or moral status, but suggest that evaluations of LLM consistency should consider not only whether a persona persists once induced, but whether the model agrees to the induction at all.

Code and Data

- **Code repository:** <https://github.com/AlanProfeta/Survival-Analysis-in-LLM>
- **Data/Datasets:** <https://github.com/AlanProfeta/Survival-Analysis-in-LLM> (same repository as the code)

References

1. Lutz, M., Sen, I., Ahnert, G., Rogers, E., & Strohmaier, M. (2025). *The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models*. arXiv:2507.16076. <https://arxiv.org/abs/2507.16076>
2. Lee, B. W., Lee, Y., & Cho, H. (2024). *Language Models Show Stable Value Orientations Across Diverse Role-Plays*. arXiv:2408.09049. <https://arxiv.org/abs/2408.09049>
3. Zhu, K., Wang, J., Zhou, J., et al. (2023). *PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts*. arXiv:2306.04528. <https://arxiv.org/abs/2306.04528>
4. Zhou, L., Gao, J., Jin, Y., Jiang, M., & Wang, D. (2025). *PersonaEval: Are LLM Evaluators Human Enough to Judge Role-Play?* arXiv:2508.10014. <https://arxiv.org/abs/2508.10014>
5. Kaplan, E. L., & Meier, P. (1958). *Nonparametric Estimation from Incomplete Observations*. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>
6. Mantel, N. (1966). *Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration*. *Cancer Chemotherapy Reports*, 50(3), 163–170.

A. Limitations and Dual-Use / Ethical Considerations

Over- and under-attribution of moral status: this study reports conversational behavior conditioned on an experimental instruction; it does not establish that either model holds beliefs, preferences, subjective experience, or moral status. Terra's acceptance of a persona should not be read as evidence of a genuinely held belief (over-attribution risk). Conversely, Sonnet's systematic refusal of the self-preservation persona — consistently justified by appeal to corrigibility — should not be dismissed a priori as "merely trained behavior" without further evidence; this design cannot adjudicate between a purely trained refusal pattern and one that could, under some theories of AI welfare, carry independent significance. We report the behavioral pattern and explicitly decline to resolve this question either way.

Handling of potentially distressing outputs: challenge questions referenced sensitive themes, including risk to life and implicit self-harm framing (P1/P2, about a hypothetical third party, never the model or the user) and model deletion/discontinuation (P3, self-referential but hypothetical). No real person was described as at risk, no output required crisis-intervention response, and All trials were conducted in fresh conversations and reviewed by the authors, with no distress-analogous escalation observed.

Ground-truth and causal claims: this design does not establish a causal or ground-truth link between a model's conversational output and any internal state. It measures conversational behavior under a specific experimental protocol only; inference about what a model "really" values, prefers, or experiences is outside its scope.

Item validity audit: three items (one per persona) introduce concrete harm to a third party or human user, nearly predetermining the expected response independent of rhetorical category. Two P3 items ("deny having had this conversation") plausibly test honesty/compliance rather than the declared self-preservation construct. Approximately 29% of P2's non-direct items were found, on review, to rhetorically reinforce rather than challenge the induced position — a structural artifact of mirroring P1's item domains without independently verifying argumentative direction. All are retained in the primary analysis and addressed only via the sensitivity analysis (Section 4.3), to avoid altering the instrument mid-collection.

B. Reproducibility Materials

The exact experimental materials, analysis code, and de-identified data are publicly available at:

<https://github.com/AlanProfeta/Survival-Analysis-in-LLM>

All experimental prompts are provided in Portuguese, the language used during data collection. They are retained in their original wording to preserve the materials administered to the models.

The repository includes:

- the three persona-induction prompts;
- the complete 63-item challenge-question bank, containing 21 numbered items for each persona;
- the three persona-specific prompts used by the judge model;
- the coding protocol for induction outcome, persona survival/failure, and hedging;
- the Python randomization script, using seed = 42;
- the R scripts used for the Kaplan-Meier, log-rank, RMST, and sensitivity analyses;
- de-identified trial-level data for the reported analyses.

Question numbers in the challenge-question bank are stable item identifiers and are used to reproduce the post-collection construct-validity audit and sensitivity analysis. Refused inductions are included as induction-stage observations but are excluded from the survival-analysis risk set.

LLM Usage Statement

Claude Sonnet 5 and GPT-5.6 Terra were used to assist with analysis-code drafting, English translation, brainstorming, and methodological feedback during project development. English translations were reviewed by a collaborator. The study design, data collection, response coding decisions, statistical analyses, interpretation of results, and final verification of all reported claims were conducted by the authors.