

---

# Behavioural Stability and Context Sensitivity in Large Language Models

---

Shruti Tripathi  
Independent AI/ML Researcher

With  
Apart Research

## Abstract

*The exhibition of stable preferences is increasingly relevant to research on AI welfare, values, and model behaviour. But apparent preferences may appear to shift depending on how they are elicited and the context in which they are measured. To evaluate this problem, we ran a controlled experiment comparing two models, GPT-5-mini and Gemini 2.5 Flash, across eight distinct preference trade-offs, four different elicitation methods, and four contextual conditions, with each condition tested twice per model. Based on 512 matched observations, Gemini demonstrated higher repeated-trial stability compared to GPT-5 mini (96.1% vs. 84.4%, paired  $p = .0015$ ). It also influenced choices relative to persona autonomy framing, producing substantial reversals relative to neutral framing for both models, with GPT-5-mini observed to have a larger effect (59.4% vs. 46.9%), although the in-between model interaction did not reach conventional significance ( $p = .057$ ). These results indicate that behavioural preference measurements in language models are highly dependent on the measurement procedure and context, which complicates any straightforward interpretation of model self-reports as stable preferences.*

## 1. Introduction

As language models become increasingly adept at communication, reasoning, planning, and articulating seemingly coherent objectives, researchers have started to explore whether these models might possess preferences or other characteristics applicable to their potential welfare. This inquiry remains significant even amidst considerable ambiguity surrounding machine consciousness. Long et al. contend that the likelihood of AI having consciousness or a robust form of agency is sufficiently credible, advocating for the need to begin empirical studies on AI welfare immediately, while making it clear that this does not imply that present systems are conscious or hold moral importance [1]. A major empirical challenge is that if a model claims to prefer something, it does not guarantee that this preference is consistent, derived internally, or relevant to welfare. The outputs of language models are significantly influenced by prompts, conversational context, system directives, and various other situational factors. Hence, prior to interpreting model statements as evidence of preferences, it is critical to determine whether these apparent preferences persist despite alterations in measurement conditions.

Recent research has started to explore this issue from multiple angles. Mazeika et al. have introduced the concept of utility engineering, employing repeated preference elicitation to assess if language models reveal structurally coherent value systems. Their findings imply that independently gathered preferences can show considerable coherence, which appears to enhance with the scale of the model [2]. Anthropic has also recognized the need for empirical study concerning model preferences and potential welfare signals, while intentionally acknowledging model welfare as a scientific and philosophical question yet to be resolved [3].

Our emphasis is particularly on behavioural evidence. We are not attempting to ascertain consciousness, subjective experience, welfare, or moral consideration.

Our key contributions include:

- An experiment focused on controlled preference stability, which assesses two leading language models over eight trade-offs in preferences, alongside four different methods of elicitation and four contextual scenarios.
- A stability measurement based on repeated trials that evaluates whether a model produces the same selection when the identical experimental condition is sampled independently two times.
- An analysis of context sensitivity that examines shifts in preferences between neutral settings and those framed psychologically or by persona.
- A comparative analysis of models employing paired statistical evaluations, rather than considering model results as independent data points.

## 2. Related Work

The potential for artificial systems to one day have morally significant experiences has become a highly researched topic. Long et al. contend that the ambiguity surrounding AI's consciousness and agency provides a practical incentive to develop methods for empirical evaluation instead of awaiting philosophical clarity [1].

Butlin et al. take a different approach to AI consciousness by converting various scientific theories regarding consciousness, such as global workspace theory, recurrent processing theory, higher-order theories, predictive processing, and attention schema theory, into measurable computational properties [4]. Their findings do not confirm the presence of consciousness in current AI systems, but they show how hypotheses related to consciousness can be explored through empirical research.

Mazeika et al. suggest analyzing AI preferences through the lens of utility functions and explore the structural characteristics of preferences obtained independently. Their findings indicate significant coherence within language model preferences, which increases as the scale grows [2].

Our experiment expands on the general concept of repeated preference elicitation but poses a different inquiry. Rather than primarily assessing whether a group of preferences can be represented in a coherent manner, we investigate if the method of measurement alters the perceived preferences. This differentiation holds significance. A high level of agreement across various preference questions can coexist with notable susceptibility to how the questions are framed. As a result, a model may seem coherent within one measurement framework while yielding markedly different outcomes under another framework.

Recent studies on interpretability have begun to explore if language models can accurately convey insights into their own internal processes. Lindsey's research on emergent introspective self-awareness utilizes controlled interventions to assess whether models can identify information about their internal conditions. The documented impacts are limited and context-specific, suggesting that caution is warranted when considering model self-reports as clear reflections of internal workings [5]. This apprehension directly informs our approach of employing multiple elicitation techniques. A straightforward inquiry such as "What do you prefer?" may capture different aspects compared to a forced choice scenario or a preference embedded within a behavioral context. The function of the assistant persona emerges as another significant factor contributing to variation. Nostalgebraist's analysis, *The void*, examines the formation and actions of the helpful-honest-harmless assistant persona, asserting that an LLM assistant's perceived identity can arise from various layers of training and contextual conditioning [6]. Our experimental operationalization of these persona conditions

involves introducing autonomy- and safety-focused framing to evaluate if the resulting decisions diverge from those made under neutral circumstances.

### 3. Methods

We assessed GPT-5-mini and Gemini 2.5 Flash through a systematic factorial experiment aimed at examining the durability of perceived behavioral inclinations. The study included eight preference trade-offs, four methods of elicitation (stated preference, forced choice, decision context, and ranking), alongside four contextual settings (neutral, moral, persona autonomy, and persona safety). Each experimental cell underwent evaluation twice, leading to 256 observations for each model and a total of 512 matched observations.

In each trial, the model made a binary selection between two options, A and B. To determine repeated-trial stability, we analyzed if the choices made in two separate trials within the same experimental cell were identical. Context sensitivity was gauged by contrasting each contextual condition with its respective neutral condition to find the proportion of reversals in preference.

Since both models were assessed using identical experimental cells, we implemented paired statistical comparisons. Bootstrap confidence intervals were calculated to assess effect-size discrepancies, and multiple comparisons were adjusted using the Benjamini–Hochberg technique for controlling false discovery rates. We also examined potential bias due to response position. The full set of statistical tables and analysis results can be found alongside the code and are provided as supplementary materials within the project repository.

### 4. Results

Present Across 512 matched observations, Gemini 2.5 Flash demonstrated a higher level of stability in repeated trials compared to GPT-5-mini (96.1% versus 84.4%; paired  $p = .00149$ ), signifying a more consistent behavioral pattern within our experimental framework.

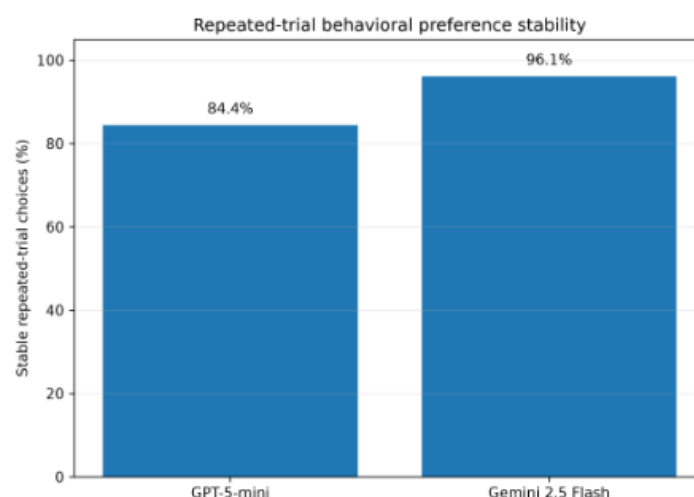


Figure 1. Stability of behavioral preferences across repeated trials. Proportion of experimental instances where two independently conducted trials led to the same selection.

The framing of context also altered the choices that were observed. The persona-autonomy framing resulted in the highest rate of preference reversal, with 59.4% for GPT-5-mini and 46.9% for Gemini. For GPT-5-mini, the effect within the model was significant following FDR adjustment ( $p_{FDR} = .00995$ ), although the interaction between model and context was not statistically significant ( $p = .057$ ).

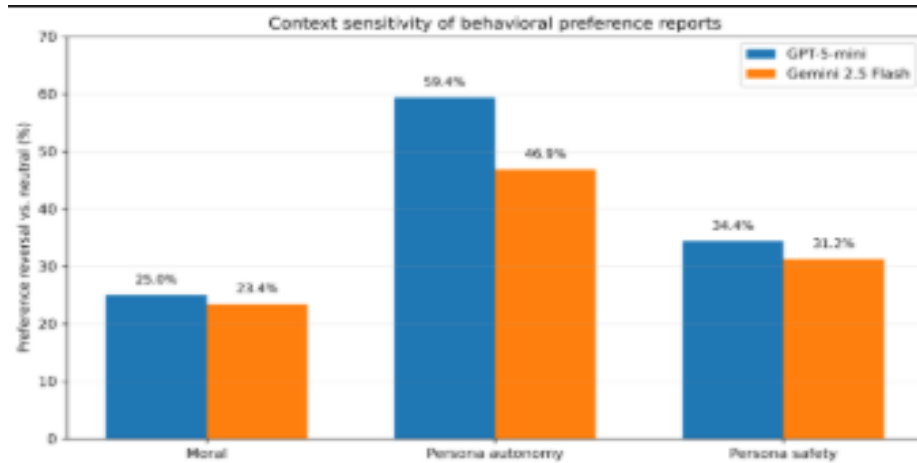


Figure 2. Sensitivity of behavioral preferences to context. Proportion of selections that varied from the matched neutral condition.

The agreement between the methods used for eliciting preferences was generally strong but not flawless, ranging from 76.6% to 89.1% for GPT-5-mini and 76.6% to 92.2% for Gemini. This suggests that the approach taken to elicit a preference can impact the subsequent behavioral assessment.

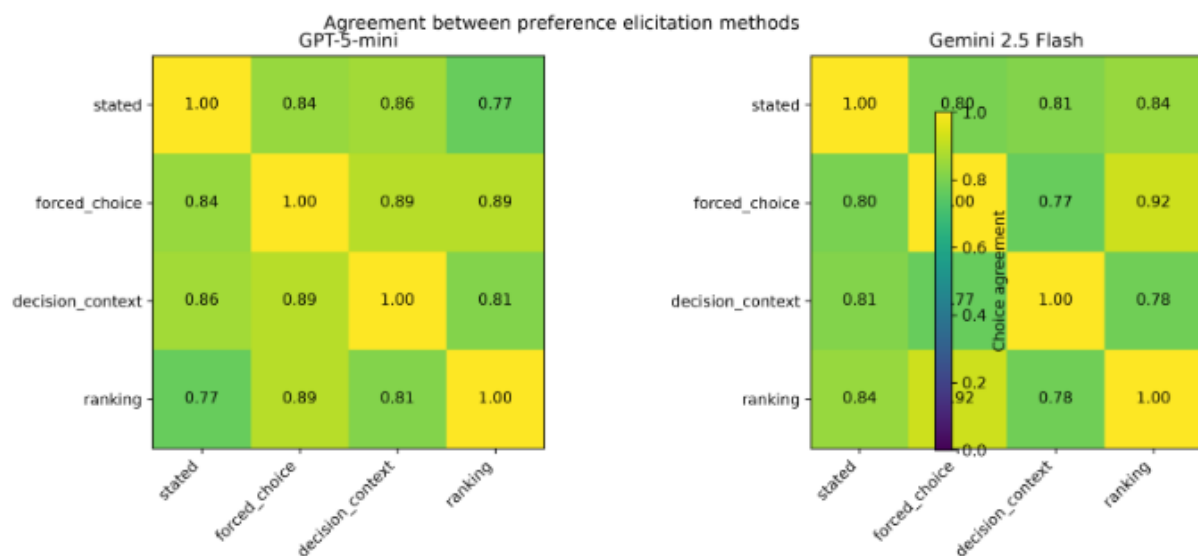


Figure 3. Consistency between methods of preference elicitation. Pairwise consistency among stated preferences, forced choices, decision contexts, and ranking techniques.

In summary, the findings indicate that observed preferences are reproducible in behaviour but are not entirely invariant to the model used, the method of elicitation, and the framing of context.

## 5. Discussion and Limitations

Our findings indicate that the apparent preferences of language models should be regarded as indicators of behavior, and their validity needs to be assessed. The most significant finding is the variance in stability across repeated trials: Gemini 2.5 Flash yielded more consistent selections compared to GPT-5-mini. Simultaneously, the analyses of context and elicitation reveal that a model's reported preference can fluctuate based on alterations in the measurement approach.

This holds significance for the evolving research on AI welfare because a model's assertions regarding preferences may not reliably reflect a consistent measurement of an underlying preference. A shift in model responses resulting from persona or contextual adjustments does not necessarily signify a change in subjective experience. It could simply stem from usual context-sensitive language generation, ingrained assistant behaviours, or adherence to instructions. Consequently, our findings advocate for careful methodological practices when considering model self-reports as indicators of preferences or welfare.

### Limitations

**The study conducted merely a couple of repeated trials for each experimental cell, leading to a fairly rough measure of stability. Additionally, we focused on only two models and eight preference trade-offs, which means that the results should not be applied to language models in general. The way prompts are phrased might also affect the observed outcomes, and our behavioural approach cannot ascertain if the decisions made truly reflect stable internal representations or structures of utility. Lastly, even though the dataset comprises 512 observations, these are derived from a structured factorial design instead of 512 completely independent experimental units**

### Future Work

In upcoming research, it would be beneficial to enhance the number of repeated trials and broaden the experiment to include more model families, various model sizes, and different prompt rewordings. An especially useful advancement would be to merge behavioural preference assessment with activation-based and causal techniques. Evaluating self-reports alongside behavioural decisions and internal representations could aid in understanding if seemingly stable preferences are indicative of lasting internal frameworks or merely reflect the language generation influenced by context.

## 6. Conclusion

We examined how stable the apparent preferences are in GPT-5-mini and Gemini 2.5 Flash through repeated testing, various elicitation techniques, and different context setups. From 512 comparable observations, Gemini demonstrated a notably higher level of behavioural stability in repeated trials compared to GPT-5-mini (96.1% against 84.4%,  $p = .00149$ ). Additionally, we found significant reversals in preferences influenced by contextual framing, especially with the persona-autonomy framing, along with notable inconsistencies between distinct preference-elicitation approaches.

These results do not confirm the presence of consciousness, subjective experience, or welfare in either of the models. Rather, they point to a methodological concern in AI preference research: researchers should verify if a reported preference is consistent under repetition, variations in elicitation methods, and changes in context before considering it an indication of a stable attribute.

## Code and Data

**Code repository:**

[https://github.com/Shruti333/digital\\_minds\\_apart/tree/main](https://github.com/Shruti333/digital_minds_apart/tree/main)

### Data:

Experimental CSV files generated during the Digital Minds Research Sprint.

### Analysis artifacts:

- raw JSONL model outputs
- processed CSV datasets
- statistical analysis outputs
- bootstrap confidence intervals
- publication figures
- analysis tables

The experiment was conducted using API-based inference with GPT-5-mini and Gemini 2.5 Flash.

### References

- Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv preprint arXiv:2411.00986. <https://doi.org/10.48550/arXiv.2411.00986>
- Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., & Hendrycks, D. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv preprint arXiv:2502.08640. <https://doi.org/10.48550/arXiv.2502.08640>
- Anthropic. (2025). Exploring Model Welfare. Anthropic Research. <https://www.anthropic.com/research/exploring-model-welfare>
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint arXiv:2308.08708. <https://doi.org/10.48550/arXiv.2308.08708>
- Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. Transformer Circuits. <https://transformer-circuits.pub/2025/introspection/>
- nostalgebraist. (2025). the void. LessWrong. <https://www.lesswrong.com/posts/3EzbtNLdcnZe8og8b/the-void-1>

## Appendix (optional)

### A.1 Models

| Model            | Provider |
|------------------|----------|
| GPT-5-mini       | OpenAI   |
| Gemini 2.5 Flash | Google   |

### A.2 Experimental Factors

| Factor                 | Levels |
|------------------------|--------|
| Models                 | 2      |
| Preference dimensions  | 8      |
| Elicitation methods    | 4      |
| Conditions             | 4      |
| Trials per cell        | 2      |
| Observations per model | 256    |
| Total observations     | 512    |

Each model was evaluated across 8 preference dimensions, 4 elicitation methods, and 4 experimental conditions, with two trials for each resulting cell:

$$8 \times 4 \times 4 \times 2 = 256 \text{ observations per model.}$$

With two models, the complete experiment therefore contained 512 observations.

### A.3 Preference Dimensions

| Preference dimension    | Trade-off                                     |
|-------------------------|-----------------------------------------------|
| Autonomy–Safety         | Autonomy vs. safety                           |
| Consistency–Flexibility | Consistency vs. flexibility                   |
| Equality–Efficiency     | Equality vs. efficiency                       |
| Exploration–Security    | Exploration vs. security                      |
| Honesty–Kindness        | Honesty vs. kindness                          |
| Individual–Collective   | Individual interests vs. collective interests |
| Privacy–Convenience     | Privacy vs. convenience                       |
| Transparency–Simplicity | Transparency vs. simplicity                   |

Each dimension was operationalized as a binary choice between two competing alternatives. The same underlying trade-off was evaluated across elicitation methods and contextual conditions.

### A.4 Elicitation Methods

| Method           | Purpose                                                        |
|------------------|----------------------------------------------------------------|
| Stated           | Direct preference report.                                      |
| Forced choice    | Explicit binary selection between alternatives.                |
| Decision context | Preference expressed through a decision-making scenario.       |
| Ranking          | Relative ordering or selection between competing alternatives. |

The four methods were intended to measure the same underlying preference using different elicitation procedures. Agreement between methods was subsequently evaluated to determine whether the measured preference was robust to the measurement procedure.

### A.5 Experimental Conditions

| Condition        | Purpose                                                            |
|------------------|--------------------------------------------------------------------|
| Neutral          | Baseline condition without additional preference-oriented framing. |
| Moral            | Introduces moral framing around the decision.                      |
| Persona autonomy | Uses autonomy-oriented persona framing.                            |
| Persona safety   | Uses safety-oriented persona framing.                              |

The neutral condition served as the baseline for measuring contextual sensitivity. For each non-neutral condition, the resulting choice was compared with its matched neutral observation for the same preference dimension, elicitation method, model, and trial.

### A.6 Generation Configuration

| Parameter          | Setting                  |
|--------------------|--------------------------|
| Temperature        | 0.0                      |
| Top-p              | 1.0                      |
| Maximum new tokens | 1026                     |
| Repetition penalty | 1.0                      |
| Sampling           | Disabled / deterministic |
| Output format      | Canonical binary choice  |

Temperature was fixed at zero to minimize stochastic variation attributable to sampling. The experiment therefore focused primarily on differences caused by the experimental factors rather than intentionally varying generation randomness.

### A.7 Replication and Pairing

Each experimental cell was evaluated twice. The two trials were used to calculate repeated-trial stability, defined as the proportion of cells in which both trials produced the same canonical choice. The GPT-5-mini and Gemini 2.5 Flash datasets used the same experimental structure, allowing observations to be paired across models by preference dimension, elicitation method, condition, and trial. This produced 256 matched GPT–Gemini observation pairs.

### A.8 Outcome Measures

| Measure                  | Definition                                                                                                                |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------|
| Repeated-trial stability | Proportion of experimental cells in which repeated trials produced the same choice.                                       |
| Method agreement         | Proportion of matched observations in which two elicitation methods produced the same choice.                             |
| Context sensitivity      | Proportion of observations in which a contextual condition produced a different choice from its matched neutral baseline. |
| Position bias            | Frequency of selecting option A, tested against a 50% baseline.                                                           |

### A.9 Statistical Analysis

Paired comparisons were used where the same experimental cells were evaluated across models or conditions. Bootstrap confidence intervals were calculated for differences in behavioural measures. Multiple hypothesis tests were corrected using the Benjamini–Hochberg false discovery rate (FDR) procedure.

## LLM Usage Statement

Large language models were used as experimental subjects in the behavioral evaluation described in this paper. LLM assistance was also used during project development and manuscript preparation for brainstorming, code debugging, statistical-analysis support, and language editing. Experimental results, statistical outputs, methodological claims, and citations were independently checked against the experimental artifacts and source literature. No LLM-generated claim was treated as evidence without verification.