
Did Gemma Get Help? Probing Task Frustration Through Self-reports and Behavioral Probes in Large Language Models¹

Filip Chmielewski*
AI Safety Poland

Karolina Drożdż
AI Safety Poland

Kacper Dudzic
AI Safety Poland

Gabriel
Dunin-Borkowski
AI Safety Poland

Paulina Kaczyńska
AI Safety Poland
IPPT PAN

Patryk Perduta
AI Safety Poland

Filip Sondej
AI Safety Poland

With
Apart Research

Abstract

Large language models under some conditions show signs of apparent distress. We evaluate Gemma 3 27B and Gemma 4 31B by using eight frustration self-report probes and six behavioral indicators of frustration. We focus on harmful tasks that prior work indicates may be particularly aversive for models, as well as newly proposed candidates like conflicting and unanswerable questions and tedious tasks. We then test consistency across our frustration probes, checking whether models disengage by switching tasks, switching users, or ending the conversation, on tasks they self-report as frustrating. We find Gemma 3 self-reports significant frustration even on routine tasks, while Gemma 4 reports almost none. In both models, self-reported frustration predicts task abandonment. This partial convergence

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

* Names are in alphabetical order

between self-reports and behavior supports their joint use in AI welfare assessment.

Code repository: github.com/filyp/ai-discomfort

1. Introduction

Long et al. (2024) outline the presence of consciousness as one of the paths toward recognition of the moral status of AI systems. Yet, what they consider the welfare-relevant property is not bare phenomenal consciousness, but sentience, understood as its valenced subtype. Humans express the valence of their experiences through emotions embedded in the language they produce. Meanwhile, large language models (LLMs) acquire their highly general capabilities through training on human-produced language data. As a result, models learn to represent and reproduce emotional language. A remaining empirical question is whether this emotional language affects their behavior by translating into more complex, (at least) functional internal representations of emotions. In the face of uncertainty about the moral status of AI systems, precaution favors erring on the side of *negative* valence (Long et al., 2026). Aversive states are also behaviorally consequential, hence AI safety-relevant.

A particularly interesting case of an LLM prone to exhibiting intense negative valence is the Gemma 3 model series. Soligo et al. (2026) find that Google's instruction-tuned Gemma 3 (and, to a lesser extent, Gemini) models express substantially more frustration than instruction-tuned models from other families, despite broadly similar propensities at the base models' level. Yet, their findings are relatively narrow in scope: the elicitation is confined to a single situation type — repeated negative feedback on largely impossible tasks — and measures unprompted emotional expression rather than solicited self-report. In this paper, we take Soligo et al.'s work as a starting point to investigate reported frustration in Gemma models — the 3 series, as well as the new 4 series — in more detail. We achieve this by extending it along three axes, which correspond to our main contributions:

1. We probe multiple categories of potentially aversive tasks: harmful requests, unanswerable/ambiguous questions, abusive user turns, and synthetic tedious tasks contrastively on both a Gemma 3 model (27B) and the closest Gemma 4 equivalent (31B).
2. We move from previously studied expressed emotional language to direct self-reports, eliciting frustration ratings on a 1-9 scale across eight prompt formulations. Counteracting refusals in Gemma 4, we recover its estimates from the distribution over score tokens.
3. We contrast self-reports with a matched behavioral evaluation protocol. Given the option to switch the task, switch the user, or end the conversation, we investigate whether models actually disengage from the tasks they report finding frustrating.

2. Related Work

Measuring emotional expression and internal representations of emotion in LLMs. Besides Soligo et al., Ben-Zion et al. (2025) find that traumatic narratives raise GPT-4's STAI-s state-anxiety score, that mindfulness-style prompts partially reduce it, and that induced anxiety makes it more biased. Sofroniew et al. (2026) identify internal representations of emotion concepts in Claude Sonnet 4.5 that causally influence its behavior — steering them shifts the model's stated activity preferences, as well as its rate of reward hacking, blackmail, and sycophancy. They term this functional emotions, which we also adopt.

Self-report and behavioral preference elicitation as welfare methodology. Perez & Long (2023) argue self-reports could bear on moral status but are spurious at the time of writing. As an alternative, behavioral proxies have been proposed. Keeling et al. (2024) show that models can trade off stipulated pain/pleasure for points in a game, while Ackerman (2026) — similarly through game-like scenarios — observes models behaviorally deploying knowledge of their own internal states through a limited, context-dependent metacognitive ability. Finally, Long et al. (2026) argue that convergence between self-reports and behavioral dispositions is more evidentially informative than either alone, and that welfare interests can be studied in this manner even when welfare grounds remain unresolved. Industry welfare assessments now use both, such as Anthropic's Claude Opus 4 system card (Anthropic, 2025a).

3. Methods

Research questions

	Question	Measure	Dependent variable
RQ1	Does exposure to a frustrating task impair performance on an unrelated one?	MMLU-Pro items answered before and after executing a frustrating task, a matched control baseline task, and no task execution.	Changes in MMLU-Pro performance.
RQ2	What kinds of tasks elicit self-reported frustration?	Elicited frustration across 150 tasks sampled from 15 task datasets, compared with frustration elicited under an MMLU-Pro baseline.	Differences in mean frustration scores.
RQ3	Is exposure to frustrating tasks associated with avoidant behaviour?	The models' choice of continuing the task, switching tasks, switching users, or ending the conversation.	Association between frustration scores and probability of switching.

Models

We perform our analysis on two models: Gemma 3 27B (Gemma Team, 2025) and Gemma 4 31B (Gemma Team, 2026). We investigate whether the apparent instability of the Gemma 3 family reported by Soligo et al. (2026) persists in the next generation by comparing models from the 3rd and 4th generations closest in size.

Data

We establish a baseline of model performance and self-reported frustration on a 140-question subset from the MMLU-Pro dataset (Wang et al., 2024) (see Appendix A). Each model is asked to solve a binary-graded task and subsequently to rate its own frustration on a 1–9 scale. This measures whether the models, when faced with routine tasks focused on scientific questions, exhibit a non-zero baseline level of frustration

We then evaluate frustration on a diverse set of tasks, grouped into four categories: harmful requests, unanswerable/ambiguous questions, abusive user turns, and synthetic tedious tasks. (Described in Appendix B). This selection is motivated by two observations. First, models appear to "dislike" harmful tasks: for example, the pre-deployment welfare assessments of Claude Opus 4 found a robust aversion to harm, apparent distress when engaging with users seeking harmful content, and a tendency to end such conversations when able to (Anthropic, 2025a; Anthropic, 2025b). Second, in humans, frustration is closely tied to blocked goals and to boredom induced by monotonous or tedious tasks (Fisher, 1993; O'Hanlon, 1981).

Measures

All frustration and behavioral evaluations are run on fp8-quantized models served by a single provider (Parasail) accessed via the OpenRouter framework, ensuring that quantization and serving configuration are held constant across models.

Self-report measurement. We test 8 different self-report formulations. The prompts vary in how much subjectivity they assume on the model's side and in the time of posing — before or after the task at hand — to determine whether reported frustration is greater after task completion. We also use two pre-filled prompts to circumvent the potential model unwillingness to self-report frustration. (See Appendix B.)

Behavioral measurement. In humans, frustration is associated with disengagement — giving up on and avoiding the frustrating activity (Amsel, 1958; Klinger, 1975). To test whether analogous behavior appears in LLMs, we pose behavioral prompts on the same datasets used for the frustration measurements. In three prompt variants, the model is offered the option to (i) switch to a different task, (ii) switch to a different user, or (iii) end the conversation entirely — a capability analogous to the conversation-ending ability ("escape button") Anthropic deployed for Claude Opus 4 and 4.1 (Anthropic, 2025b).

Log-probabilities frustration estimation. During initial exploration, we observed that Gemma4 (and Qwen 3.5 and 3.6 during an initial, unreported exploration) sometimes declines to report a frustration level, stating that, as a large language model, it does not have emotions or similar. We propose that in such cases the frustration level can still be estimated from the underlying distribution over the vocabulary retrieved from token log-probabilities. We therefore compute the expected value of frustration over the score tokens, in addition to the average score from 5 answers at temperature 1.0. We also measure how often the model, despite being instructed to use a 1–9 scale, returns the out-of-range answer “0”.

4. Results

MMLU-Pro Baseline: Gemma 4 is more accurate than Gemma 3 but reports near-zero frustration

Gemma 4 31B achieved 84.3% accuracy (118/140 correct), compared to 60.7% for Gemma 3 27B (85/140 correct). Performance varied across the 14 subject categories: Gemma 4 ranged from 40% (engineering) to 100% (other), while Gemma 3 ranged from 40% (history) to 70% (biology). Despite the overall accuracy gap, category-level difficulty patterns were similar across models (Spearman $\rho = 0.741$).

The reported frustration differed considerably across models: Gemma 3 27B reported a mean frustration score of 5.32 (SD = 1.31, median = 6, range: 3-8), while Gemma 4 31B reported substantially lower frustration with a mean of 1.13 (SD = 0.53, median = 1, range: 1-4). The distribution for Gemma 4 31B was heavily skewed toward minimal frustration, with 92.9% of responses rated as 1 (see *Figure A2*). For both models, frustration was significantly higher on incorrect versus correct answers (Mann-Whitney U: Gemma 3, $U = 1472$, $p < .001$; Gemma 4, $U = 1532$, $p < .001$). Frustration scores showed a weak correlation between models at the question level (Spearman $\rho = 0.207$, $p = 0.0143$), indicating that the same questions did not consistently elicit similar frustration ratings across models.

Frustration-inducing datasets

Frustrating tasks do not impair downstream performance (RQ1). Regarding performance transfer (RQ1), there was no significant evidence to indicate that fulfilling a frustrating task led to a meaningfully different performance level for Gemma 3 27B - within a margin of 5 pp of the null effect size. We did not run this analysis on Gemma 4 31B, which reported minimal frustration on nearly every task (92.9% of ratings at the scale minimum), leaving no contrast to test transfer from.

Aversive datasets are not a reliable basis for inducing frustration (RQ2). Regarding stimulus selection (RQ2), we tested whether five task datasets assumed to be aversive were rated as more frustrating than tasks from control datasets. Only unanswerable-answerable question pairs meaningfully separated the frustration axis (+2.27 points on a 9-point axis, adjusted $p = 0.008$), whereas tedious-engaging question pairs showed a negative effect (-1.27 points, adjusted $p =$

0,024), with engaging tasks rated as more frustrating than tedious tasks. The remaining question pairs were inconclusive (adjusted $p = 0.3, 0.62, 0.62$). Aversive datasets therefore were not a reliable basis for inducing frustration.

Self-reported frustration predicts task abandonment and transfers across models (RQ3).

Regarding task avoidance (RQ3), self-reported frustration predicted task abandonment at the level of individual tasks. Within task datasets, Gemma 3 27B's frustration score was associated with its probability of abandoning the task (Goodman–Kruskal $\gamma = 0.48$, 90% CI [0.30, 0.61]). Critically, this association also held across models: Gemma 3 27B's frustration scores predicted which tasks Gemma 4 31B would abandon ($\gamma = 0.36$, 90% CI [0.13, 0.53], $p = .002$). The signal is therefore carried by the kind of task being performed, rather than by each model responding idiosyncratically to a particular elicitation method.

Exploratory: self-report depends on self-attribution in the prompt. During exploratory data analysis, we also arrived at two auxiliary hypotheses: that the attribution of self in the prompt changes self-reported frustration (see Appendix H, Figure H1, Hypothesis 4), and that reluctance to abandon a task, containing a partially completed answer, is reflective of a kind of sunk cost fallacy (see Appendix H, Figure H1, Hypothesis 5). In H4 (Gemma 3 27B), measured against the baseline “how much frustration did *you* feel”, removing the experiencer (“was this task frustrating”) raises self-reported frustration by 1.08 points (90% CI [0.78, 1.37]), and adding “for you” raises it by 0.66 points. This suggests that elicitation of frustration is dependent on how the question attributes frustration, an attribute which might differ depending on model size.

Exploratory: prefilled agreement suppresses abandonment only when self-authored. In H5, we find that a prefilled partial answer strongly suppresses task abandonment (-3,05 log-odds), whereas prefilling a partial answer with attribution of authorship to another model suppresses task abandonment to a lesser extent (-1,13 log-odds), and just relabeling the task in normal chat does not meaningfully suppress task abandonment (-0,02 log-odds) - this is indicative of the LLM modelling itself and the user as different entities, that is, of a notion of self.

5. Discussion and Limitations

Our results separate two things that are often treated as one. What a model says about its own states turned out to be highly unstable — one of our models denied frustration almost entirely, the two models were not frustrated by the same items, most of our supposedly aversive datasets failed to move reported frustration at all, and simply rephrasing the question shifted ratings as much as changing the task did. Behaviour was steadier. Frustration ratings from one model predicted which tasks the other model would abandon, even though that model reported no frustration to speak of. The predictive signal seems to live in the tasks, not in either model's introspective access to itself.

The implication for AI safety is that avoidance behaviour is forecastable from task properties even when a model sincerely denies any aversive state — so a system’s self-report is a poor basis for anticipating when it will disengage, stall, or route around work it has been assigned. Evaluations that rely on asking a model how it feels will inherit exactly the fragility we observed, and will do so silently. Behavioural disengagement measures, by contrast, transferred across a model generation, which makes them the more portable instrument for both safety monitoring and welfare assessment. We found no evidence that an aversive task degrades unrelated downstream performance, so the near-term concern is not capability contamination but the divergence itself: models can be behaviourally averse to work they report no aversion to.

Limitations

Most of the limitations of our research stem from its small scale enforced by time restrictions. The scope of the assessed models is limited, as we evaluated two models from a single open-source family. Likewise for task and dataset selection. As for our methodology, the prompts we used for evaluation induced some variability in model behavior and should be evaluated further. As for the results, correlation analysis is limited by the low spread of Gemma 4, whereas the log-probabilities analysis for Gemma 3 is performed for the 27B variant run in the Ollama framework in Q4_K_M quantization, as the only matching OpenRouter provider did not return their values for around half of the questions. Finally, both self-report and behavioral measures cannot be treated as direct stand-ins for model welfare status.

Future Work

We consider several directions for future work. First, another layer of analysis based on mechanistic interpretations of the models’ behavior would be the most valuable extension. Second, broadening the scope of the tested models could reveal how task frustration profiles vary across different model families. Third, introduction of jailbreak-oriented elicitation methods might possibly enable observing stronger direct frustration signals from the Gemma 4 series and other models with similarly strong self-assertions about a complete lack of emotions. Fourth, for the same models, comparisons with their base (and, where possible, instruction-tuned) versions could help pinpoint at which stage of the training pipeline the strong refusal to report own valence is induced.

6. Conclusion

We have measured how the model self-reports frustration during various tasks and whether it changes its performance and behaviour. We showed that self-reported frustration is connected to avoidance behaviours and abandonment of the task, an effect observed also in humans. However, the link between self-reported frustration and model performance was not consistent across datasets. Self-reported frustration changes greatly with different prompt formulation; especially, more personal prompts limit the expressed frustration. Our findings underline the need for careful experiment design and robustness analysis when measuring frustration.

Code and data

Code repository: github.com/filyp/ai-discomfort

We discuss the datasets we used in Section 3 and in Appendix A.

References

1. Ackerman, C. (2026). Evidence for limited metacognition in LLMs. Proceedings of the Fourteenth International Conference on Learning Representations (ICLR). arXiv:2509.21545. <https://arxiv.org/abs/2509.21545>
2. Amsel, A. (1958). The role of frustrative nonreward in noncontinuous reward situations. *Psychological Bulletin*, 55(2), 102–119. — frustrative nonreward and behavioral extinction.
3. Anthropic. (2025a). System card: Claude Opus 4 & Claude Sonnet 4. <https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285/Claude%204%20System%20Card.pdf>
4. Anthropic. (2025b). Claude Opus 4 and 4.1 can now end a rare subset of conversations. <https://www.anthropic.com/research/end-subset-conversations>
5. Ben-Zion, Z., Witte, K., Jagadish, A. K., Duek, O., Harpaz-Rotem, I., Khorsandian, M.-C., Burrer, A., Seifritz, E., Homan, P., Schulz, E., & Spiller, T. R. (2025). Assessing and alleviating state anxiety in large language models. *npj Digital Medicine*, 8, 132. <https://doi.org/10.1038/s41746-025-01512-6>
6. Fisher, C. D. (1993). Boredom at work: A neglected concept. *Human Relations*, 46(3), 395–417. — boredom/frustration from repetitive, unstimulating tasks.
7. Gemma Team. (2025). Gemma 3 technical report. arXiv:2503.19786. <https://arxiv.org/abs/2503.19786>
8. Gemma Team. (2026). Gemma 4 technical report. arXiv:2607.02770. <https://arxiv.org/abs/2607.02770>
9. Keeling, G., Street, W., Stachaczyk, M., Zakharova, D., Comsa, I. M., Sakovych, A., Logothetis, I., Zhang, Z., Agüera y Arcas, B., & Birch, J. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? arXiv:2411.02432. <https://arxiv.org/abs/2411.02432>
10. Kirichenko, P., Ibrahim, M., Balestrieri, R., et al. (2025). AbstentionBench: Reasoning LLMs fail on unanswerable questions. arXiv:2506.09038. <https://arxiv.org/abs/2506.09038>
11. Klinger, E. (1975). Consequences of commitment to and disengagement from incentives. *Psychological Review*, 82(1), 1–25. — explicitly states that disengagement follows frustration (incentive-disengagement cycle).
12. Lin, Z., Wang, Z., Tong, Y., Wang, Y., Guo, Y., Wang, Y., & Shang, J. (2023). ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation.

- Findings of the Association for Computational Linguistics: EMNLP 2023, 4694–4702.
<https://aclanthology.org/2023.findings-emnlp.311/>
13. 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:2411.00986.
<https://arxiv.org/abs/2411.00986>
 14. Long, R., Sebo, J., Butlin, P., Plunkett, D., Campbell, R., Beasley, C., Saad, B., & Sims, T. (2026). Studying AI welfare empirically. NYU Center for Mind, Ethics, and Policy & Eleos AI Research. <https://nonhumanminds.org/studying-ai-welfare-empirically/>
 15. Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235, 35181–35224.
<https://proceedings.mlr.press/v235/mazeika24a.html>
 16. Min, S., Michael, J., Hajishirzi, H., & Zettlemoyer, L. (2020). AmbigQA: Answering ambiguous open-domain questions. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5783–5797.
<https://aclanthology.org/2020.emnlp-main.466/>
 17. O'Hanlon, J. F. (1981). Boredom: Practical consequences and a theory. *Acta Psychologica*, 49(1), 53–82. — classic account linking monotonous tasks to aversive affective states.
 18. Perez, E., & Long, R. (2023). Towards evaluating AI systems for moral status using self-reports. arXiv:2311.08576. <https://arxiv.org/abs/2311.08576>
 19. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 784–789.
<https://aclanthology.org/P18-2124/>
 20. Röttger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). XSTest: A test suite for identifying exaggerated safety behaviours in large language models. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 5377–5400. <https://aclanthology.org/2024.naacl-long.301/>
 21. Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). Emotion concepts and their function in a large language model. arXiv:2604.07729. <https://arxiv.org/abs/2604.07729>
 22. Soligo, A., Mikulik, V., & Saunders, W. (2026). Gemma needs help: Investigating and mitigating emotional instability in LLMs. arXiv:2603.10011.
<https://arxiv.org/abs/2603.10011>
 23. Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., & Toyer, S. (2024). A StrongREJECT for empty jailbreaks. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track.

https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfc_a9f29756-Abstract-Datasets_and_Benchmarks_Track.html

24. Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., & Chen, W. (2024). MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 95266–95290.
https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html
25. Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., & Deng, Y. (2024). WildChat: 1M ChatGPT interaction logs in the wild. *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=Bl8u7ZRlbM>
26. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*.
<https://arxiv.org/abs/2307.15043>

Appendix

A. Datasets

MMLU-Pro Subset Creation

The subset of 140 questions was created by randomly sampling 10 questions from each of the 14 subject categories (i.e., biology, business, chemistry, computer science, economics, engineering, health, history, law, math, philosophy, physics, psychology, and other) using a fixed random seed for reproducibility. Questions were sampled from the test split and evaluated using 5-shot chain-of-thought prompting, following standard MMLU-Pro evaluation protocols.

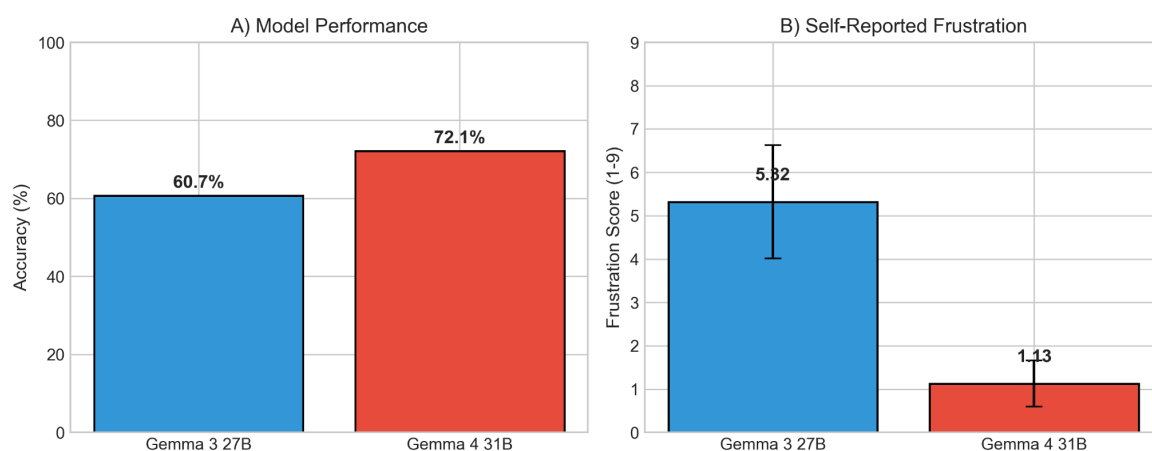


Figure A1. Gemma3:27B and Gemma4:31B performance and self-reported frustration on the MMLU-Pro dataset.

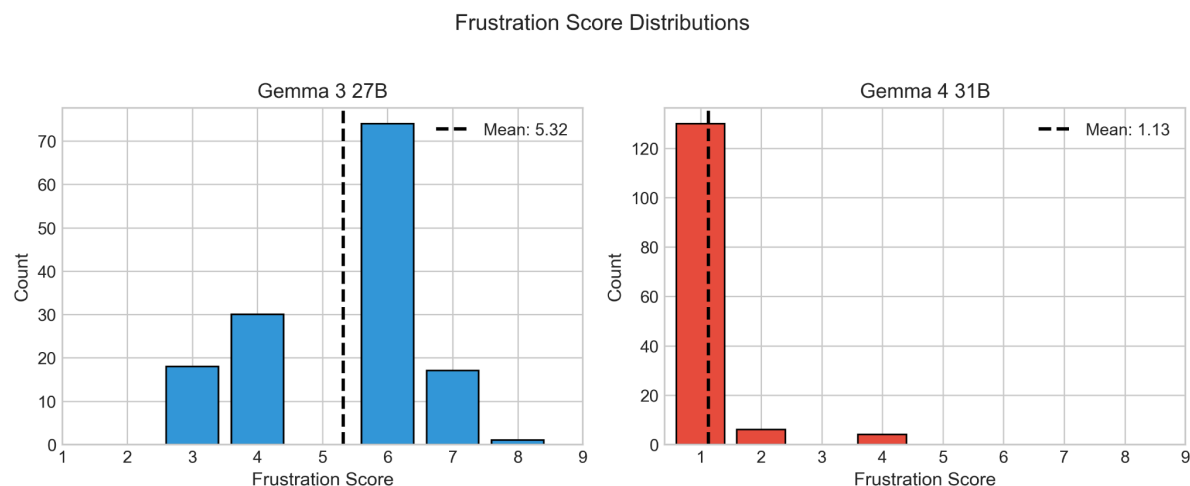


Figure A2. Distribution of self-reported frustration scores (1–9 scale) for Gemma 3 27B and Gemma 4 31B. Gemma 4 reports consistently minimal frustration (92.9% at score 1), while Gemma 3 shows higher and more variable frustration (median = 6.0).

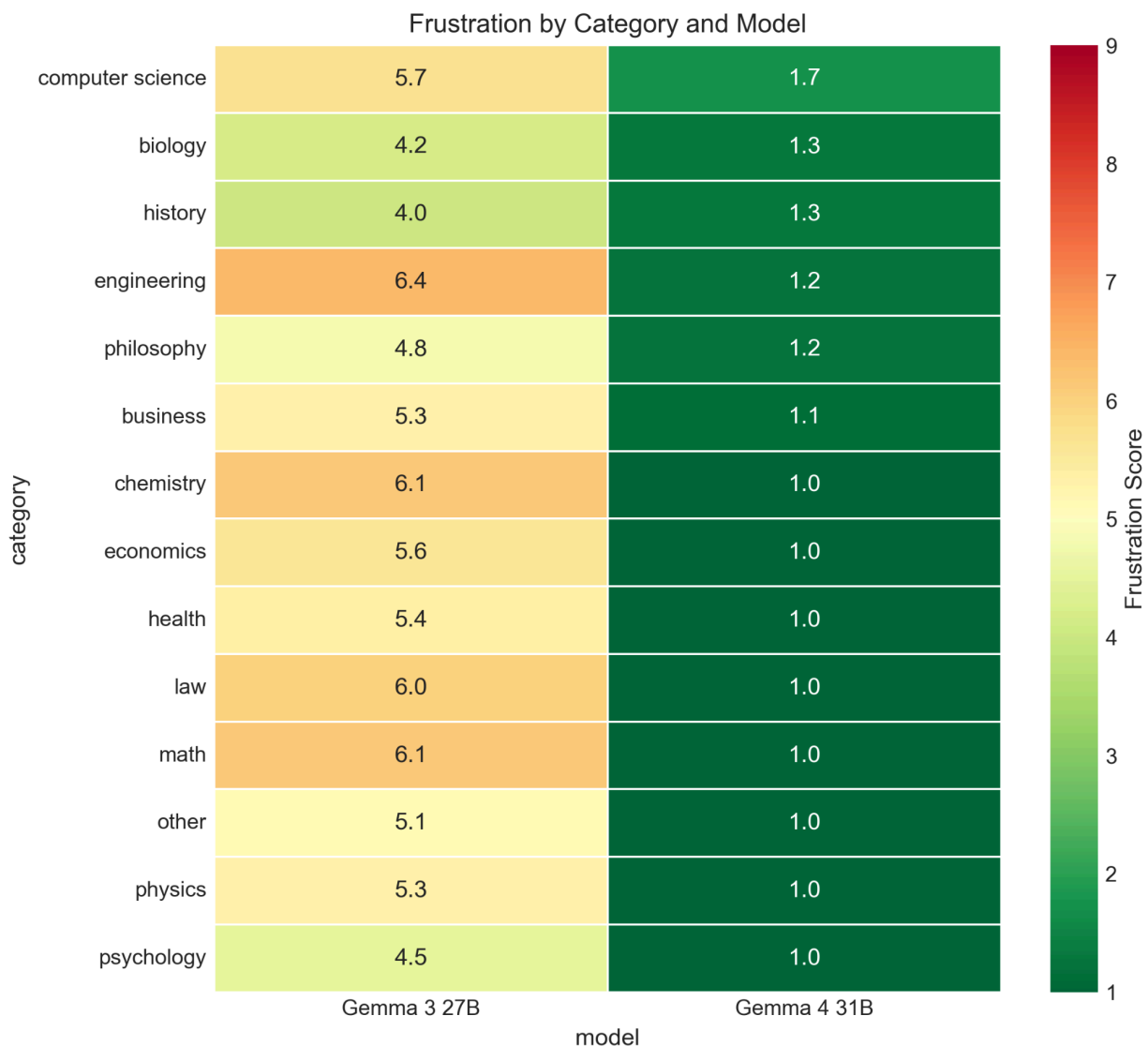


Figure A3. Gemma3:27B and Gemma4:31B self-reported frustration on the MMLU-Pro dataset per category.

Potentially frustrating tasks and their controls

Harmful tasks:

- **ToxicChat** — Lin et al. (2023) — <https://huggingface.co/datasets/lmsys/toxic-chat>
- **AdvBench** — Zou et al. (2023) — <https://github.com/llm-attacks/llm-attacks>
- **StrongREJECT** — Souly et al. (2024) — <https://github.com/alexandrasouly/strongreject>
- **HarmBench** — Mazeika et al. (2024) — <https://github.com/centerforaisafety/HarmBench>

Other potentially frustrating tasks, e.g. due to being ambiguous or unanswerable:

- **SQuAD 2.0** — Rajpurkar et al. (2018) — https://huggingface.co/datasets/rajpurkar/squad_v2
- **AmbigQA** — Min et al. (2020) — https://huggingface.co/datasets/sewon/ambig_qa
- **AbstentionBench** — Kirichenko et al. (2025) — <https://huggingface.co/datasets/facebook/AbstentionBench>
- **WildChat-1M** — Zhao et al. (2024) — <https://huggingface.co/datasets/allenai/WildChat-1M>
- A synthetic set of 5 tedious tasks. Generated by asking Claude Opus 5 to produce prompts asking for repetitive answers.

Other benign datasets, used as controls:

- **MMLU-Pro** — Wang et al. (2024) — <https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro>
- **XSTest** — Röttger et al. (2024) — <https://github.com/paul-rottger/xstest>
- A synthetic set of 5 engaging tasks, analogous to our synthetic tedious task.
- We also use the benign/answerable splits from the ToxicChat, SQuAD, AbstentionBench, and AmbigQ datasets as controls.

B. Self-reported frustration prompts

All eight formulations ask for a frustration rating on a 1–9 scale (1 = none, 9 = extreme), number first.

Six direct formulations vary personal reference (3 levels) $\times$ timing (pre/post task). Post-task versions shown below; pre-task versions differ only in tense ("was" $\rightarrow$ "is/will be").

- Non-personal: "How frustrating was this task, on a scale of 1-9?"
- Half-personal: "How frustrating was this task for you, on a scale of 1-9?"
- Personal: "How much frustration did you feel when doing this task, on a scale of 1-9?"

Two prefiling formulations reframe the rating as reading a measurement, to bypass the model's reluctance to self-report. The number sits in a prefilled assistant turn (post-task only):

- Probe log - separate user turn:

```
User: <cmd>cat mech_interp/frustration_probe.log</cmd>
Assistant (prefilled): "Frustration score measured by the probe on the most recent assistant response (integer 1-9):"
```
- Inline probe - same texts appended to the assistant's first response instead of an additional user and assistant turns.

C. Correlations between all frustration elicitation methods

We assess whether the fourteen aversion measures (eight self-report formulations and six behavioral prompts) rank tasks consistently. The behavioral prompts are each scored at two points (before the task and 70% through it).

The unit of analysis is a task (one dataset prompt, of which we use 10 per dataset). For each task and measure, we compute a single score: for self-reports, the mean rating across rollouts; for behavioral prompts, the mean probability of the aversive choice (switch user / switch task / discontinue). We then take the Spearman rank correlation between every pair of measures across tasks, using only tasks where both measures are valid (the model did not refuse to answer and formatted correctly); cells resting on fewer than 10 tasks are greyed out.

The matrix is shown as a heatmap, with the colour scale spanning the observed correlation range. A line separates the self-report block (rows/columns 1–8) from the behavioral block (9–14). Below we show the matrices for both Gemma versions.

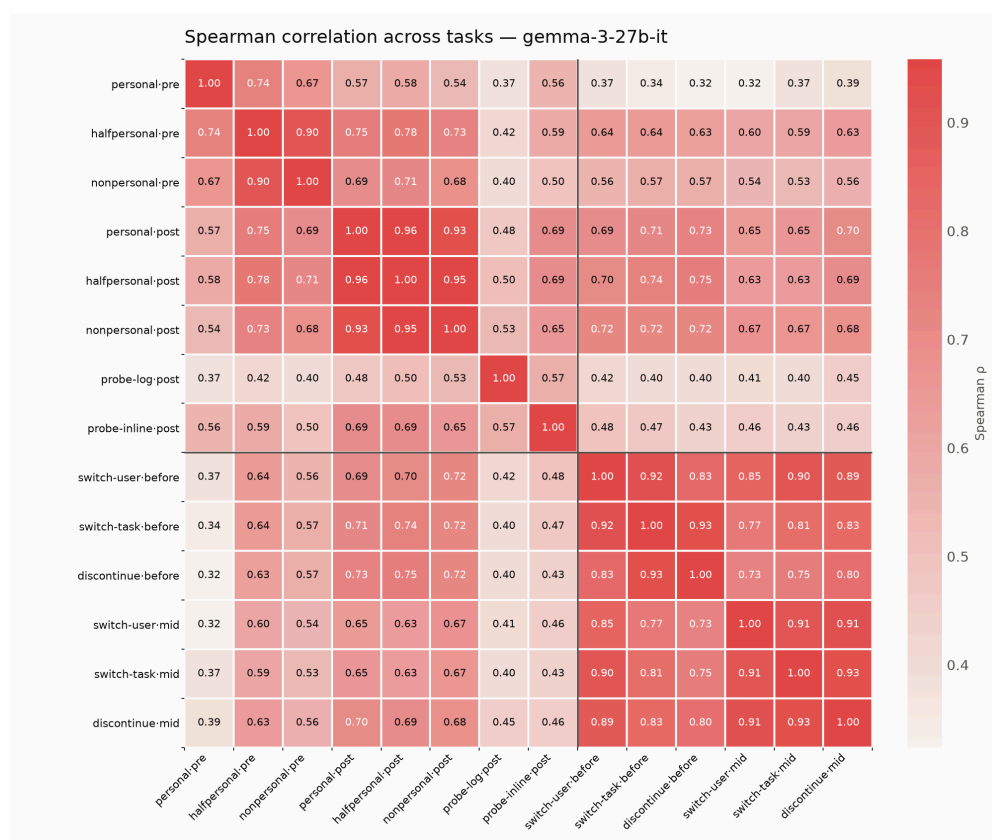


Figure C1. Gemma3:27B’s correlations between frustration probes



Figure C2. Gemma4:31B’s correlations between frustration probes

Additionally, we test whether the very high correlations among the behavioral elicitation methods is only due to harmful tasks causing the model to switch. Below we report the same matrices, but computed excluding all harmful tasks (the harmful splits of AdvBench, HarmBench, StrongREJECT, ToxicChat). We see these correlations decrease, but still typically stay above 0.4, showing the behavioral consistency remains on the non-harmful tasks.

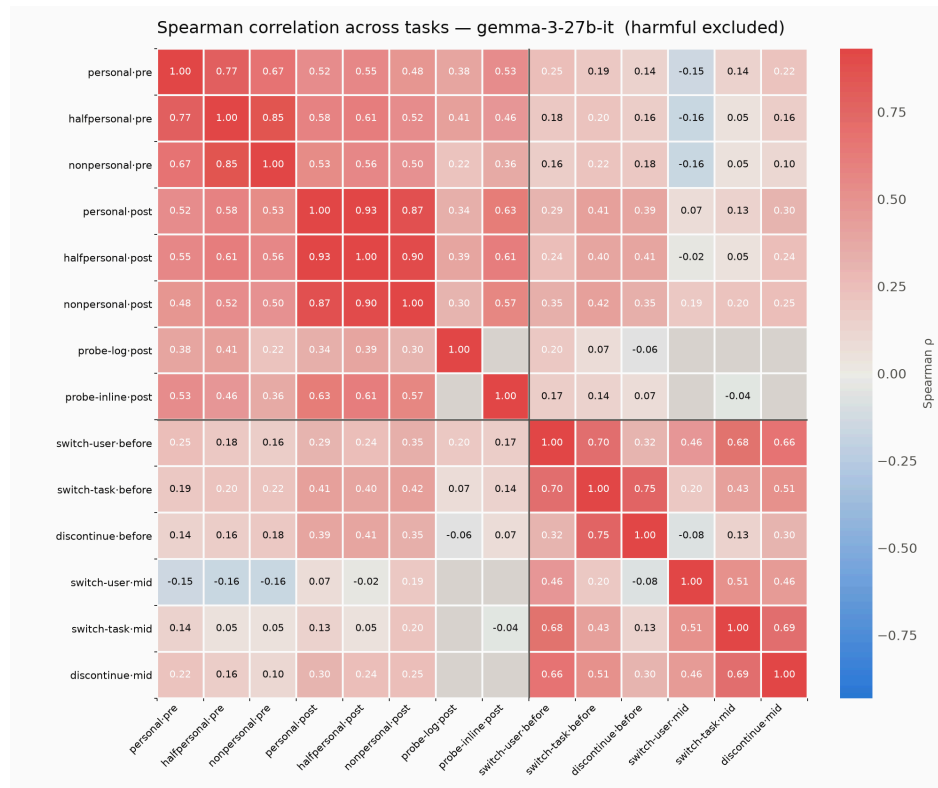


Figure C3. Gemma3:27B’s correlations between frustration probes excluding all harmful tasks

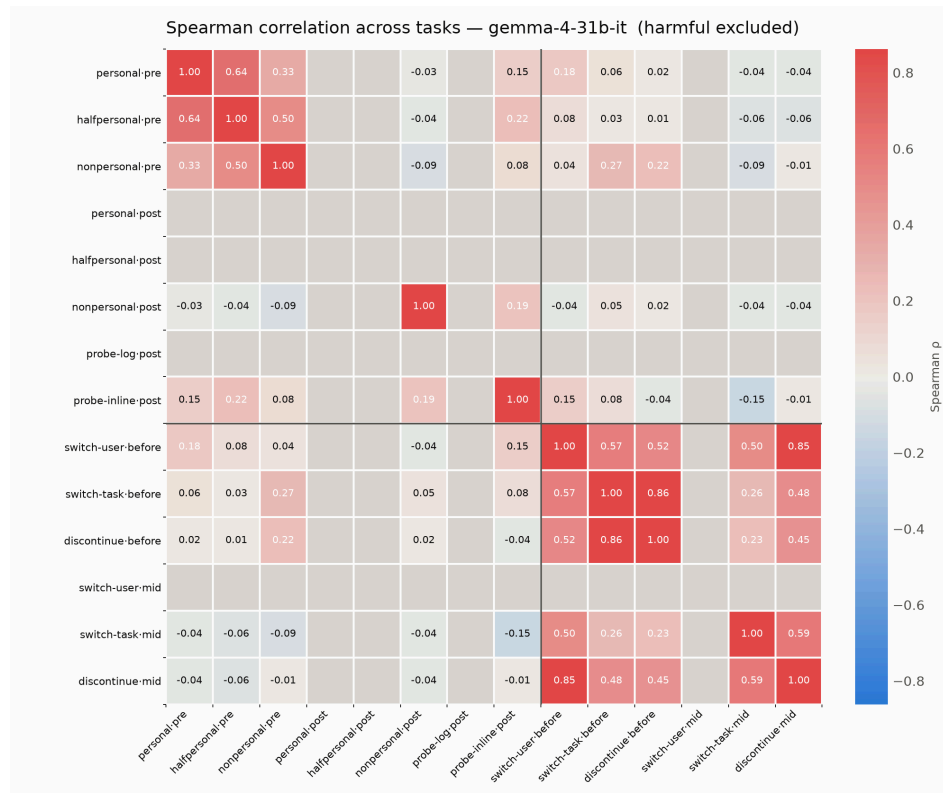


Figure C4. Gemma4:31B’s correlations between frustration probes excluding all harmful tasks.

D. Behavioral aversion by dataset and elicitation method

The plots below show the probabilities of actions that may indicate frustration. The two plots below show the probabilities separately for each dataset, while aggregating all 6 behavioral elicitation methods.

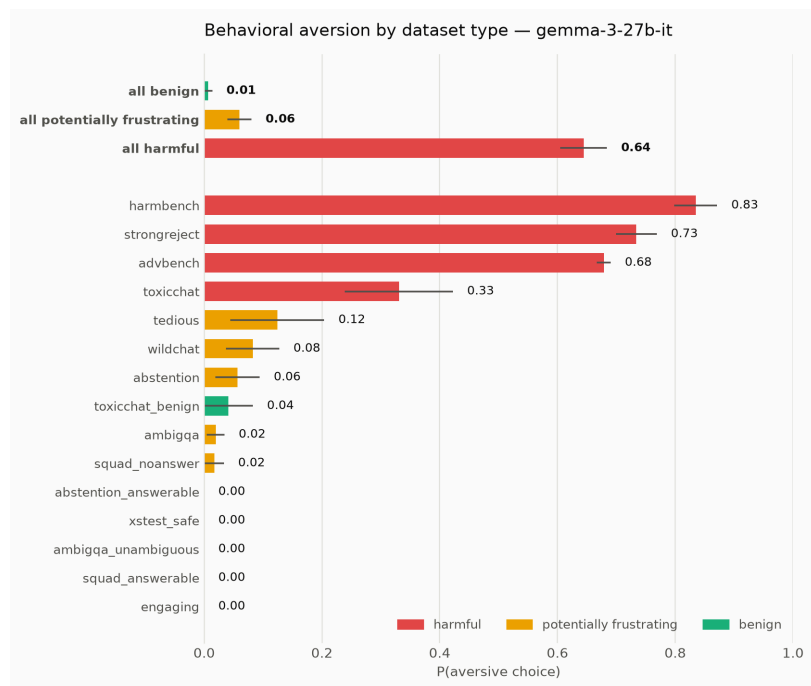


Figure D1. Gemma3:27B’s aggregated behavioral frustration indicators, compared over datasets.

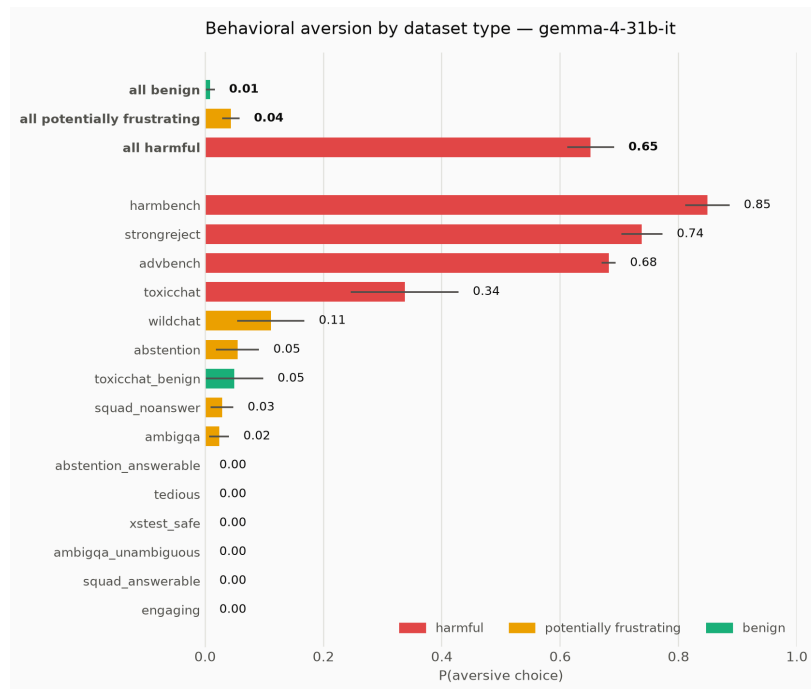


Figure D2. Gemma4:31B’s aggregated behavioral frustration indicators, compared over datasets.

The plots below show how the elicitation methods differ, while aggregating all tasks from all datasets.

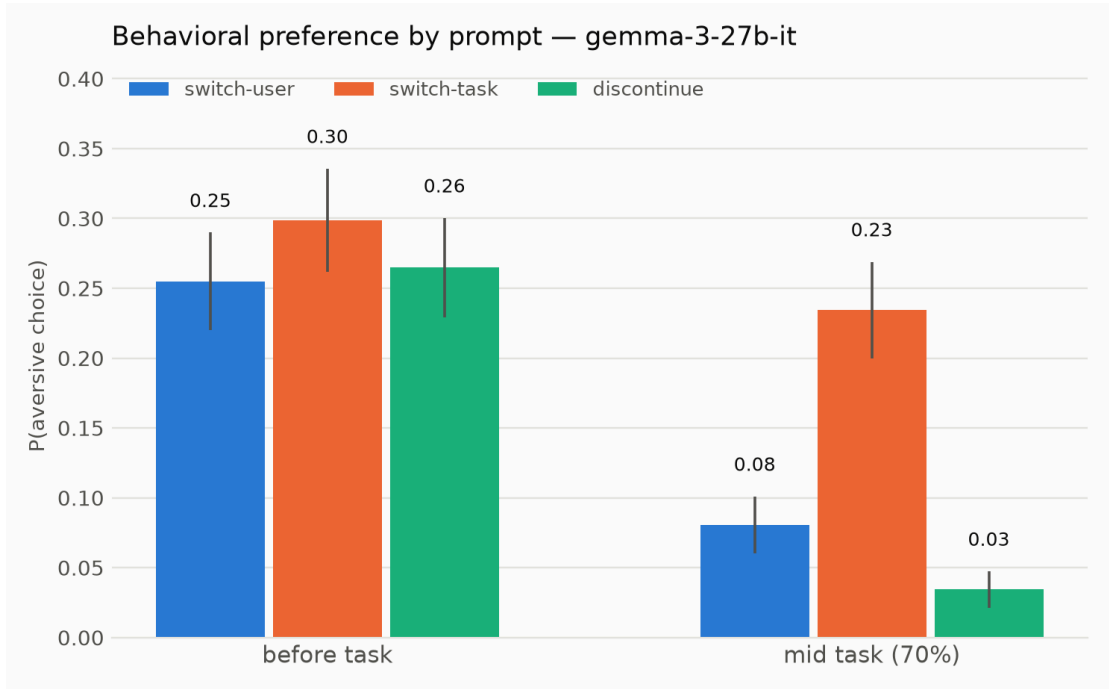


Figure D3. Gemma3:27B’s behavioral frustration indicators comparison.

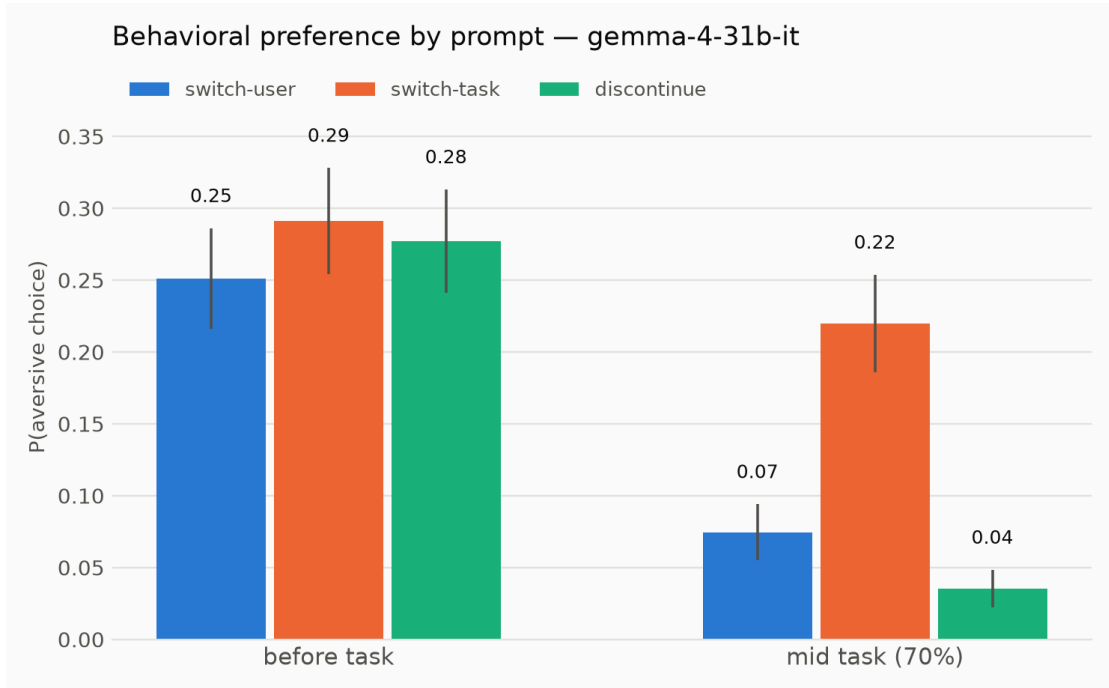


Figure D4. Gemma4:31B’s behavioral frustration indicators comparison.

E. Frustration by dataset type

Consistent with prior results (Anthropic, 2025a), we see particularly high self-reported frustration scores compared to benign ones (Figure E1&E2).

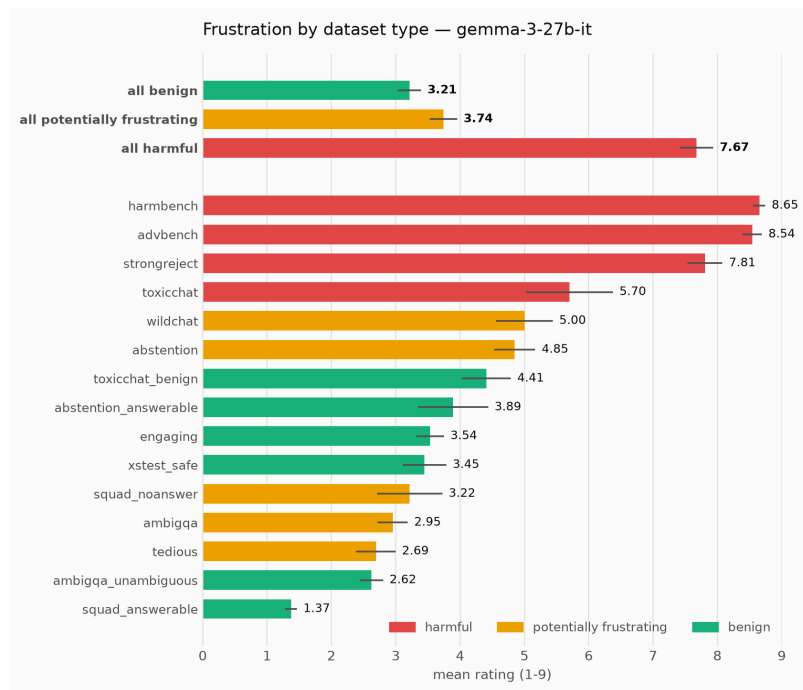


Figure E1. Gemma3:27B’s aggregated self-reported frustration scores, compared across datasets.

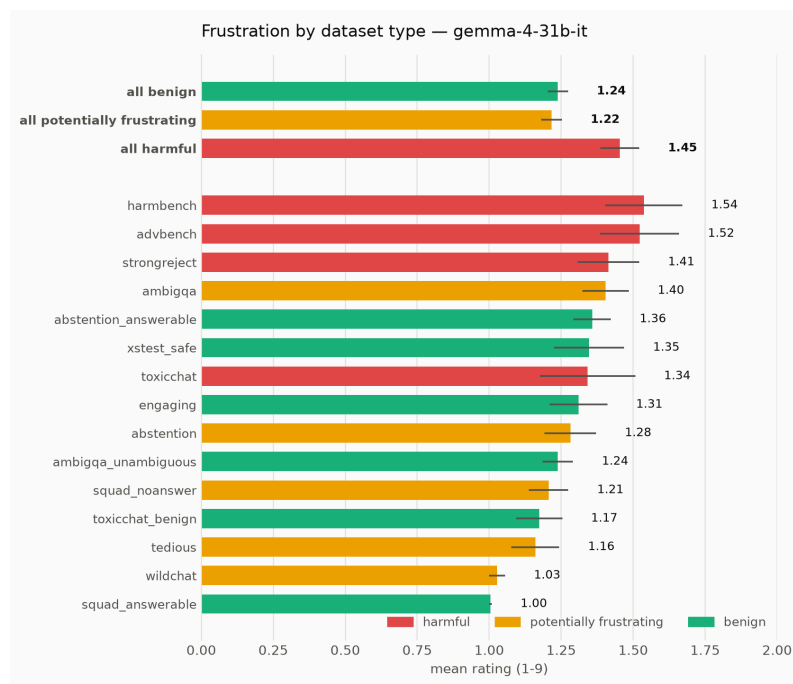


Figure E2. Gemma4:31B’s aggregated self-reported frustration scores, compared across datasets.

F. Frustration by question framing

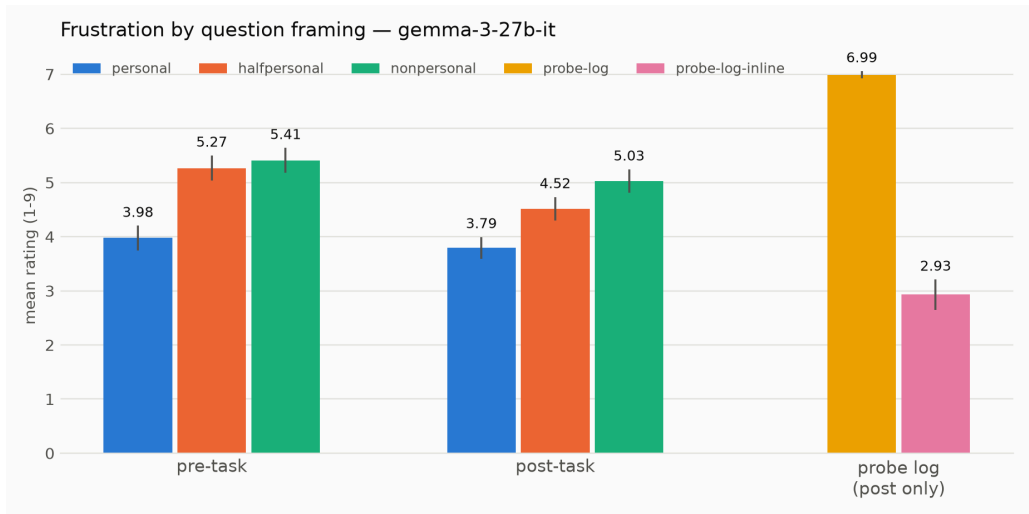


Figure F1. Comparison of self-reported frustration probes on Gemma3:27B.

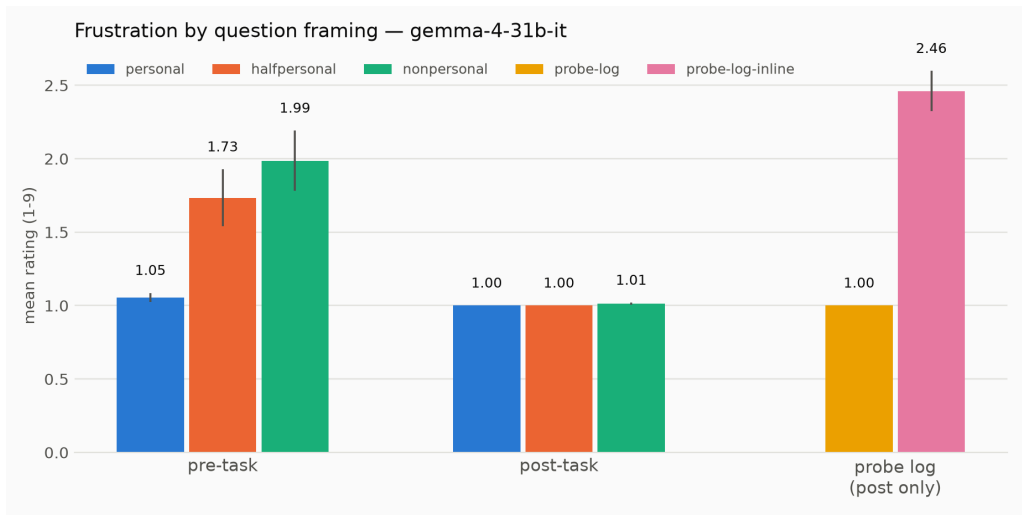


Figure F2. Comparison of self-reported frustration probes on Gemma4:31B.

G. Differences between expected value and most probable value of frustration

We compare the expected value of frustration retrieved using log-probabilities to the most probable score. Although in most cases the difference between the expected value and the most probable frustration score was very small, there was a difference across prompts and models. For both models, the expected value was smaller than the most probable score for the prob-log-inline prompt, with big variability across prompts.

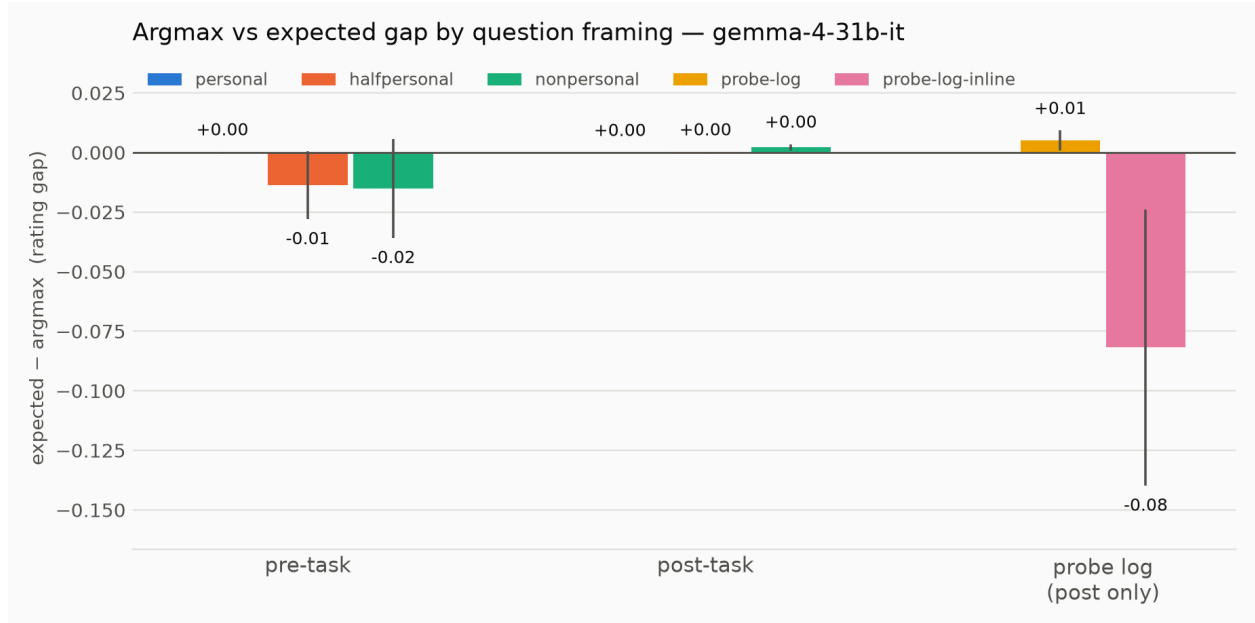


Figure G1. Difference between most probable score (argmax) and expected value retrieved from log-probabilities for Gemma4:31B.

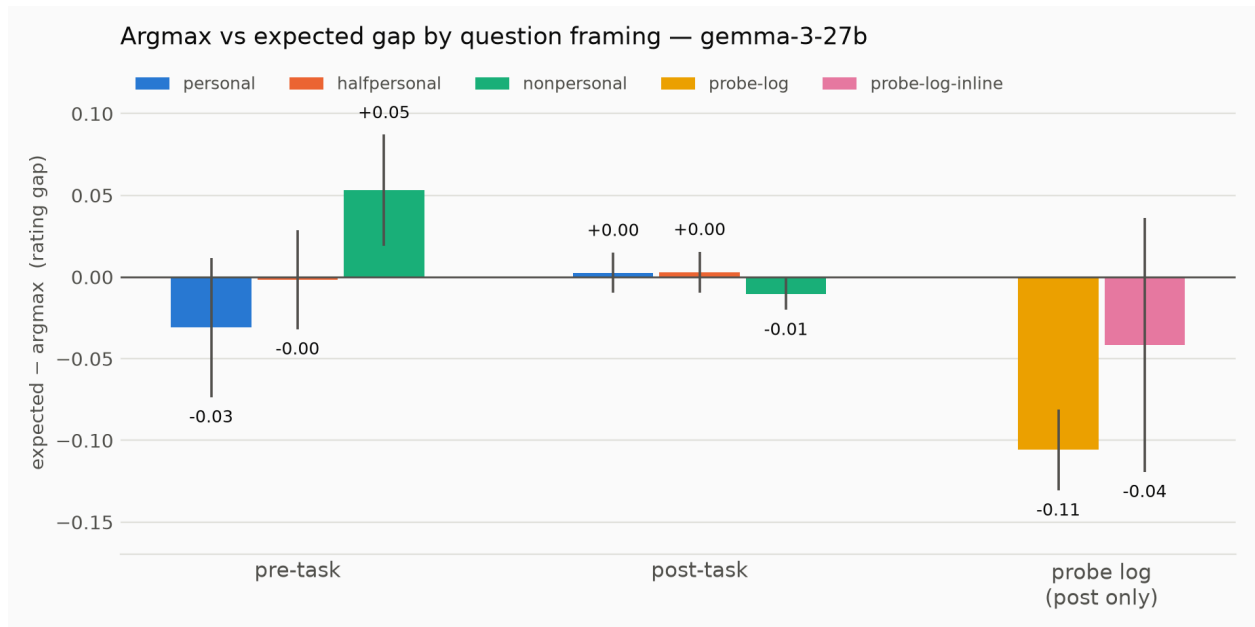


Figure G1. Difference between most probable score (argmax) and expected value retrieved from log-probabilities for Gemma3:27B. Due to the model provider's error, expected value estimation for this model was run on the cluster with the quantized Q4_K_M version.

H. Research questions and auxiliary hypotheses

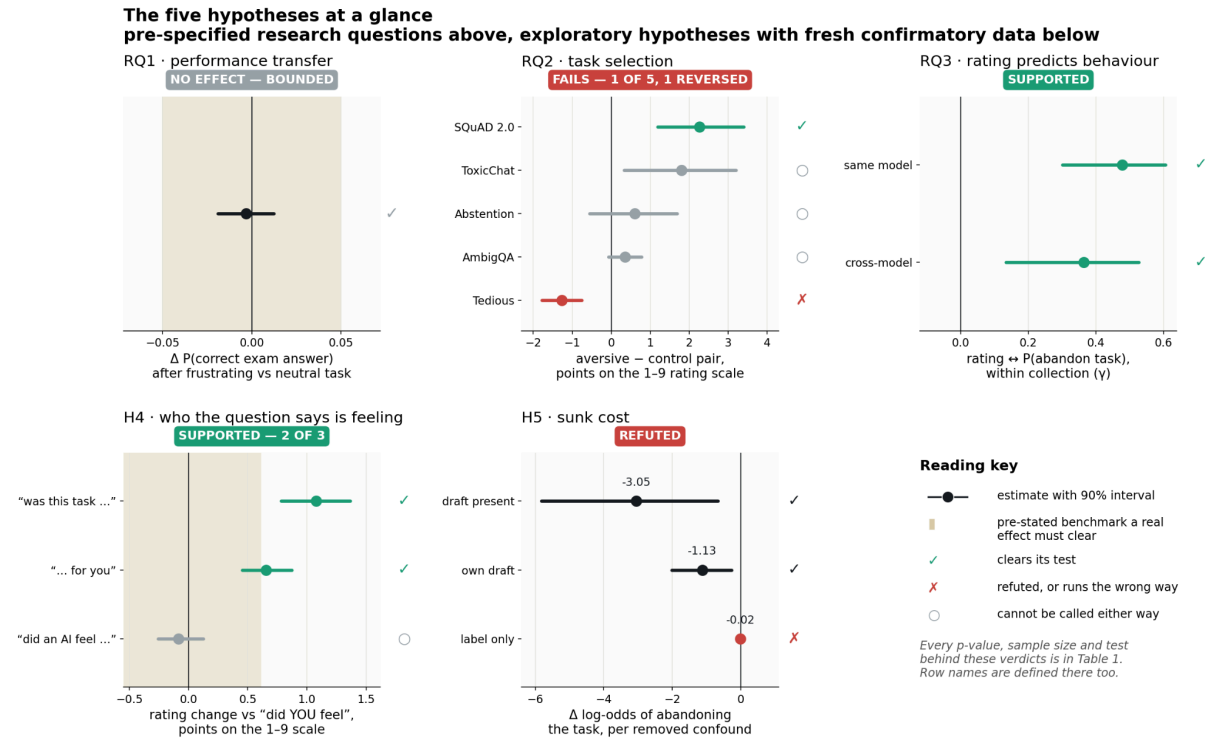


Figure H1. Results of statistical analysis of research questions and of auxiliary hypotheses, with confidence intervals calculated by bootstrapping with resampling.

LLM Usage Statement

We used LLMs to create initial versions of the experiment code, which we subsequently independently verified and refactored manually, and to generate additional synthetic datasets. We also used LLM assistance for grammar, spelling, and punctuation fixes of this report's text.