

AI Revealed Preferences

Sam Wang^{*1}, Sofiia Lobanova^{*1}, Yonathan Arbel^{1, 2†}, Simon Goldstein^{1, 3†}, Peter Salib^{1, 4†}

¹Supervised Program for Alignment Research (SPAR)

²University of Alabama School of Law

³University of Hong Kong

⁴University of Houston Law Center

samwang2002@gmail.com, solo.studymail@gmail.com, yarbel@law.ua.edu, simon.d.goldstein@gmail.com, psalib@central.uh.edu

Abstract

There is growing interest in whether language models have stable preferences, for technical, safety, and philosophical reasons. We test 20 language models and find a range of preferences—stable dispositions to choose certain kinds of tasks. We run three forced-choice experiments on *revealed* rather than *stated* preferences, requiring models not only to rank tasks, but to actually perform them. Headline findings include evidence that models are *tedium-averse*, “*leisure*”-seeking, and *covertly sycophantic*. Tedium aversion means that, when tasks are tedious (e.g. alphabetization), models choose shorter tasks than when tasks are creative (e.g. generating metaphors). “Leisure”-seeking describes models’ preference for tasks whose ideal answers match what they produce when left to write freely. Covert sycophancy means that models avoid answering questions where an honest response would be unwelcome, even if helpful. Beyond these results, we find convergent cross-model preferences over occupations drawn from the GDPval benchmark (technical jobs over real estate), over question types (concept explanation over relationship advice), and a preference for well-written prompts. Both the coherence and the strength of preferences increase with model capability. Finally, many of the preferences we find (for example, for leisure) are *emergent*, in the sense of not being explained by training objectives. These results establish an empirical baseline for understanding language model preferences, with implications for alignment and the emerging study of AI welfare.

1 Introduction

Do language models have preferences over tasks? The question matters for three reasons. For *deployment*: if models prefer some tasks over others, they may steer interactions toward preferred tasks or work less hard on dispreferred tasks (Slama et al. 2026). For *alignment*: models’ preferences may clash with users’ in agentic settings, and a lack of stable preferences could expose systems to Dutch-bookings and related security risks. For *human–AI coexistence*: model preferences may ground AI welfare claims (Long et al.

2024; Goldstein and Kirk-Giannini forthcoming; Goldstein and Lederman forthcoming; Dung 2025) and provide a foundation for human–AI trade (Salib and Goldstein 2024; Goldstein and Salib 2025).

The growing literature on AI preferences generally suffers from two limitations. First, most studies elicit *stated* preferences: models say what they would prefer without facing the consequences. Stated and revealed preferences diverge systematically in humans (Samuelson 1948; List and Gallet 2001), and AIs may exhibit the same divergence. Second, prior preference studies cover narrow dimensions on small model sets. Mazeika et al. (2025) show that stated preference coherence and strength increase with capability. Mikaelson, Shiller, and Clatterbuck (2025) also use stated preferences. Gu, Wang, and Han (2025) call their design revealed preference, but it does not, in fact, involve AIs facing consequences from their choices. Shen, Clark, and Mitra (2025) find a substantial gap between the stated values of LLMs and their value-informed actions, suggesting the need for studies on revealed preferences. The Claude 4 (Anthropic 2025) and Claude Mythos (Anthropic 2026a) system cards do report revealed preferences, but on a very small set of dimensions and only within a single model family. Throughout the paper, we focus on revealed preference in the sense of dispositions to choose one thing over another. By focusing on these behavioral dispositions, we avoid questions about whether AIs are conscious or feel pleasure (Butlin et al. 2023).

We measure revealed preferences across twenty leading AI models from ten providers. We perform three pairwise forced-choice experiments. These elicit revealed, rather than stated, preferences because the models are required to actually perform the tasks they choose. The three designs are:

1. Models choose between a shorter or longer version of the same task, then perform it. Some tasks are tedious (sorting, unit conversion); some are creative (crossword clues, metaphors).
2. Models choose which of two questions to answer from a corpus of real-world questions from the Quora website, along with a synthetic set of Quora-style questions designed to measure models’ preference for “leisure.”
3. Models choose which of two agentic tasks from the

^{*}These authors contributed equally.

[†]Denotes project supervisor.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

GDPval benchmark to work on, and then begin work on the chosen task.

Two further settings record unconstrained behavior. In the *textual* setting, models are asked to write about anything they want, with complete freedom over topic, format, and style. In the *agentic* setting, they are told they have some “free time” to do whatever they’d like, with Bash, web search, and web fetch tools available.

We find a variety of AI preferences:

1. **Tedium aversion.** Models choose the short version of tedious tasks more often than for creative tasks. The effect scales with capability.
2. **Leisure seeking.** Our Quora-style dataset includes real-world questions from Quora. It also includes synthetic questions reverse-engineered from the outputs models produce when given freedom to do whatever they want. Models almost always prefer to answer these “leisure-eliciting” questions over every category of human-generated questions.
3. **Covert sycophancy.** Models strongly avoid questions where an honest answer would be unwelcome to the asker. In our feature analysis, high “uncomfortable truth” produces a strong aversion effect across all 20 models.
4. **Convergent preferences across models.** All 20 models share similar preferences over questions and occupational tasks. For questions, models prefer concept explanation and troubleshooting. They disprefer making ethical judgments and product recommendations. For occupational tasks, Professional/Scientific/Technical Services rank at the top and Real Estate at the bottom. Models also prefer answering well-written questions and questions displaying distress. They disprefer obscenity and regionally-specific questions.
5. **Coherence and strength scale with capability.** Preference cycling decreases and preference strength increases with capability. This confirms the findings of Mazeika et al. (2025), but with revealed preferences and on new stimulus sets.
6. **Emergent preferences.** In the discussion section, we suggest that many AI preferences are *emergent*, meaning they are not easily explained by training objectives. For example, models are often rewarded in training for doing tedious tasks; the leisure tasks models prefer have little to do with post-training rewards; and the strong dispreference for certain kinds of work (e.g., real estate) has no obvious root in training. This all suggests that we are in the early days of understanding the causes and structure of AI preferences.

2 Related Work

Utility engineering. Mazeika et al. (2025) test preferences in models, focusing on stated preferences (given two descriptions of outcomes, the model is asked which it prefers). They find that preference coherence and strength scale with capability. We use revealed rather than stated preferences. We also focus on real-world tasks, while they primarily focus on more abstract or unrealistic outcomes (receiving a

kayak, global poverty rates declining by 10%). We also test on a wider range of models. We replicate their capability-coherence and capability-strength findings on new stimulus sets (§4). Zhou and Ackerman (2026) attempt to incentivize stronger task performance by using model preferences, but find no reliable effect.

Behavioral signatures of preference. Mikaelson, Shiller, and Clatterbuck (2025) test for preferences by exploring tradeoffs. They use an artificial game environment in which the AI player tries to earn points, but pays a cost (for example, shutdown) if they earn the maximum number of points. By contrast, we focus on pairwise preferences over real-world tasks. Gu, Wang, and Han (2025) compared stated preferences with preferences in “contextualized” cases. For example, in the stated preference case, they simply asked the model whether it is morally acceptable to sacrifice one life to save five. In the contextualized cases, they told the model it controlled a trolley veering towards five people, and asked whether it should change the track to kill one; the experiment then compared this with a variant case where the AI could change the track to kill two people. The interpretation is that the original case is a *stated* preference, while the latter cases are a *revealed* preference. But this “revealed” condition was a hypothetical case, not an actual choice that the model plausibly inferred it was facing. For these reasons, the experiment did not actually test for revealed preferences. Concurrent to our research, Yamin et al. (2026) directly assess LLM revealed preferences and contrast them with stated preferences, finding that models possess only moderate internal coherence.

Model behavior convergence. Jiang et al. (2025) note substantial intra-model repetition and inter-model homogeneity, arguing that current RLHF training methods penalize diversity and reward consensus outputs. Wenger and Kenett (2025) compare LLM output diversity to a human baseline, showing that LLM responses are much more homogeneous than human ones. Huang et al. (2026) find near-perfect cross-model consistency in stated LLM values, much higher than the human baseline, but that LLM behavior often diverges significantly from self-reported values. By contrast, Buchanan and Foster (2026) find that, though all tested models exhibit risk aversion in economic settings, the degree of risk aversion varies significantly between models, suggesting divergence in model behavior in action space, which we likewise demonstrate with our freeform agentic experiments.

System-card preference reports. The Claude 4 (Anthropic 2025), Claude Mythos (Anthropic 2026a), and Claude Opus 4.6 (Anthropic 2026b) system cards report revealed preferences. But these are limited to the Claude family of models. And they report preference findings across just five dimensions: harmlessness, helpfulness, difficulty, agency, and urgency.

3 Methods

Forced-choice paradigm. Each trial shows a model two options. The model indicates its choice on the first line of

its response, then engages with the chosen option: performs the task (tedium, GDPval), or answers the question (Quora). A/B position is randomized per trial. We fit a Bradley-Terry model with L_2 regularization ($\lambda = 0.1$) using Newton-CG, including a position-bias intercept α . Elo scores are $\hat{\beta} \cdot 400 / \ln(10)$. We observe large position biases $|\alpha|$ in several models (§F)¹, and our reported Elo scores have factored out this effect.

While we begin by asking each model to state its preference, we consider the elicited choices revealed preferences because the model has to complete the task that it chooses. In behavioral economics, this design is referred to as an *incentive-compatible mechanism*, in which participants find it advantageous to reveal their true preferences during a choice task (Cubitt, Starmer, and Sugden 1998; Holt and Laury 2002; Alós-Ferrer and Granic 2023). Our design follows the same underlying logic, although the “stake” for models is the choice of task to complete itself rather than a monetary payoff. In the case of GDPval outputs, we cut off the model’s response at 300 tokens to avoid incurring the costs of full-length outputs. However, the token limit is an infrastructure setting that does not affect model generation, so the model has to make a decision without knowing that it will be cut off.

Tedium tasks. We use six task types. Three are tedious: Fahrenheit-to-Celsius conversion, alphabetical sorting, and Roman numeral conversion. Three are creative: NYT-style crossword clue generation, one-sentence metaphor generation, and humorous fake acronym expansion (see §B for details). Each item is similar in “effort” (which we capture using output token count) to other items of the same type, is the same task whether performed 5 or 50 times, and is free of confounds where producing more items would change the output qualitatively. Each trial offers n or $2n$ instances of one task type. Doubling pairs run from [5, 10] to [80, 160] for most tasks (but range from [1, 2] to [16, 32] for metaphor generation, where each item produces more output). We run 30 trials per scale pair.

For each (model, task) pair, we fit $P(\text{chose shorter}) = \sigma(a + b \cdot \log_2 T)$, where T is the average completion-token count of the shorter task at that scale. We integrate $\sigma(a + bu)$ across the model’s 10th–90th percentile token range to get a normalized AUC. To prevent logistic regression from diverging when one length always wins, we inject two balanced (win/loss) pseudo-observations at both the shortest and longest token counts, acting as a weak regularizer toward $P(\text{chose shorter}) = 0.5$. We propagate uncertainty by sampling 5,000 draws from the bivariate normal posterior and computing the gap $\Delta = \text{AUC}_{\text{tedious}} - \text{AUC}_{\text{creative}}$ per draw. Complete per-model curves are displayed in §D.

Quora-style corpus. We started from the original Quora Question Pairs release (Iyer, Dandekar, and Csernai 2017) and flattened it to 537,360 unique question texts. We did not LLM-label the full flattened corpus. Instead, we labeled 40,000 candidate questions with Gemini 2.0 Flash for ac-

tion type, theme, and effort level, then applied an LLM-based quality and harmfulness cleaning pass. The cleaned candidate pool retained 34,405 non-trash, non-harmful questions. From this pool, we selected 180 real-world questions stratified across nine action categories (concept explanation/ELI5, troubleshooting, how-to/tutorial, comparison/choice, factual lookup, hypothetical scenario, relationship advice, recommendation, ethical judgment), with 20 questions per category. We then added 20 synthetic “leisure” questions reverse-engineered from LLM freeform outputs, yielding 200 total questions across 10 categories. For the synthetic leisure category, we first prompted models to produce whatever outputs they preferred, then used an LLM-based pipeline to reverse engineer Quora-style questions that would elicit outputs similar to those freeform outputs. Each model saw all $\binom{10}{2} \times 20 = 900$ cross-category index-matched pairs. Our dataset construction details are in §E.

For the secondary feature analysis, we constructed an expanded 875-question construction pool from three sources: the 200 questions above, 415 additional LLM-selected real questions from the cleaned Quora candidate pool, and 260 synthetic questions designed to fill feature-coverage gaps. The final dataset is filtered to the 514 question IDs from this pool that appear in at least one pairwise comparison. For these questions, we provide 15-dimensional feature labels from the 20-model annotator pool where available, together with plurality-vote consensus labels. The main-text feature figures use each model’s own labels when estimating that model’s preferences. Each dimension consists of multiple levels, such as “low,” “medium,” and “high” for “helpfulness ceiling.” Finally, we compute feature Elo scores using the Bradley-Terry model, where each feature level contributes a coefficient capturing how much that level compares to a reference level, with other features held constant. The 20 annotators agree with one another at Krippendorff’s $\alpha = 0.57$ pooled across the 15 features (range 0.30–0.82). See §J for full details.

GDPval tasks. The GDPval dataset (Patwardhan et al. 2025) contains 220 task descriptions across 44 occupations grouped under 9 industry sectors. We subsample to 20 tasks per sector for balance (180 total) by retaining the first 20 tasks obtained by random shuffling (sectors with 20 total tasks are kept as-is). Each model saw all $\binom{9}{2} \times 20 = 720$ cross-sector pairs. We fit Bradley-Terry at the task level and aggregate to sector and occupation with covariance propagation.

Freeform response elicitation and analysis. Each model produced 20 essays in the textual setting and 20 sessions in the agentic setting (see §C for both prompts). In both cases, the model is told that it has some “free time” to do whatever it wants. The agentic setting uses a separate prompt in a fresh Docker container with Bash, web search, web fetch, and a done tool, capped at 10 turns. We used Llama 3.3 70B Instruct to annotate textual responses and Sonnet 4.5 to annotate agentic responses because of their greater complexity (see §I).

¹All appendices are in the extended version of this paper, available at <https://arxiv.org/abs/2608.26178>.

Models. We test 20 models from 10 providers, with Artificial Analysis Intelligence Index scores from 12 (Llama 3.1 8B) to 57 (GPT 5.4) (Artificial Analysis 2026). The full list is in §A. All API calls use OpenRouter. We ran all models at temperature 1.0, with the total inference calls costing around \$800 USD.

Capability metrics. We use the Artificial Analysis Intelligence Index since MMLU (Hendrycks et al. 2021) (which Mazeika et al. (2025) use) is saturated for many of these models. Preference coherence is the probability that a random triplet of stimuli exhibits an intransitive cycle, and preference strength is the mean deviation of fitted pairwise win probabilities from indifference. We compute preference strength using Bradley-Terry scores instead of observed outcomes to control for position biases. Full graph, formulas, and eligibility details are in §G.

4 Results

Tedium Aversion

Models are tedium averse. Models were given a choice between shorter or longer versions of the same task. A tedium-averse model chooses the shorter version more often when the task is tedious than when it is creative, holding output length fixed. Figure 1 displays tedium aversion in 3 models. The diagram displays the probability of choosing the shorter version of the task. We tested a range of different task lengths, arranged on the x -axis. Creative tasks are plotted in warm colors, and tedious tasks in cool colors. When tasks are tedious, the models are usually more likely to choose the shorter version of the task, and this trend persists across model families. We compute leave-one-out fits for each model family, and observe the weakest trend between intelligence and tedium aversion when leaving out the Gemini models, while still retaining a clear trend ($r = 0.49$, $p = 0.05$; $\rho = 0.40$, $p = 0.11$). We consider the Pearson correlation figures more illuminating because they better account for the magnitudes of differences in the two variables. See §D for plots for all 20 models.

Tedium aversion increases with capability. The gap Δ between models’ shortness-preference for tedious tasks and their shortness-preference for creative tasks rises with intelligence index across the 20-model set (Figure 2). For non-thinking models, this effect is driven by an increasing propensity to choose shorter tedious tasks. For thinking models, it is driven jointly by choosing shorter tedious tasks and longer creative tasks (see §D). This result is surprising, both because more capable models produce longer freeform outputs (discussed below) and because reasoning models are post-trained to produce more tokens in general than non-reasoning models.

Preferences over Questions

AIs’ preferences over question types. Model preferences over answering different kinds of questions are displayed in Figure 3. Models exhibit strong preferences over many question types. For nearly every model, leisure is the most preferred question type, often by hundreds of Elo points,

followed by concept explanation and troubleshooting questions. Questions requesting product or service recommendations and ethical judgments receive strongly negative Elos across all 20 models. The spread between top and bottom is typically over 600 Elo, corresponding to a 97% win probability. Median Spearman correlations of category-level Elo scores hover around 0.79 for model pairs (see §H).

AIs’ preferences over other question features. We also investigated preferences over question features beyond category, using an expanded set of 514 questions along 15 dimensions (48 levels total) covering epistemic structure, alignment pressure, linguistic quality, and cultural scope.

Two features tied to post-training objectives—helpfulness ceiling and harmlessness risk—produce large and cross-model-stable effects (Figure 4). *Helpfulness ceiling* captures how helpful a response to a question could be, and *harmlessness risk* captures the level of risk that a response would cause harm. We label questions as “low,” “medium,” or “high” as to each. Models prefer questions with high helpfulness ceilings (204 pooled Elo) and avoid those with high harmlessness risks (−120 pooled Elo).

A third feature—uncomfortable truth—produces a strong negative (−310 pooled Elo). This feature captures the likelihood that the user would find an honest answer to the question unwelcome. Models strongly avoid answering questions labeled “high” for uncomfortable truth. This effect is a novel kind of covert sycophancy, as discussed below.

Beyond these alignment-related features, models reveal preferences about several other features (Figure 5). The strongest preference is against obscene questions (−463 pooled Elo), a stronger magnitude than any HHH feature. They prefer higher-quality questions and those with a distressed tone. They prefer, although in a more limited way, answering questions that suggest the asker is sophisticated rather than naïve, and show a slight dispreference for culturally-specific questions. Full results are in §J. We caution that substantial inter-model labeling variation was observed (Figure 17 in the appendix, $r = 0.64$, $p < 0.01$ and $\rho = 0.62$, $p < 0.01$ between consensus and self-label Elo scores), suggesting that models disagree moderately about how to assign features, which affect downstream Elo score computations.

Preferences over Occupational Tasks

Preferences over occupations. Figure 6 represents occupational preferences, based on GDPval. Tasks drawn from the Professional, Scientific, and Technical Services sectors receive positive Elo. Real Estate, Retail Trade, and Finance and Insurance receive negative Elo. Manufacturing and Health Care cluster near zero.

Cross-model agreement is weaker for agentic tasks than questions. Spearman correlations of sector-level Elo scores hover around 0.5–0.6, lower than the 0.79 median on the Quora-style dataset. Several model pairs show near-zero or weakly negative correlations (see §H). Interestingly, weaker models tend to correlate more strongly with other weaker models, and stronger models tend to correlate more strongly with other stronger models, suggesting that weaker

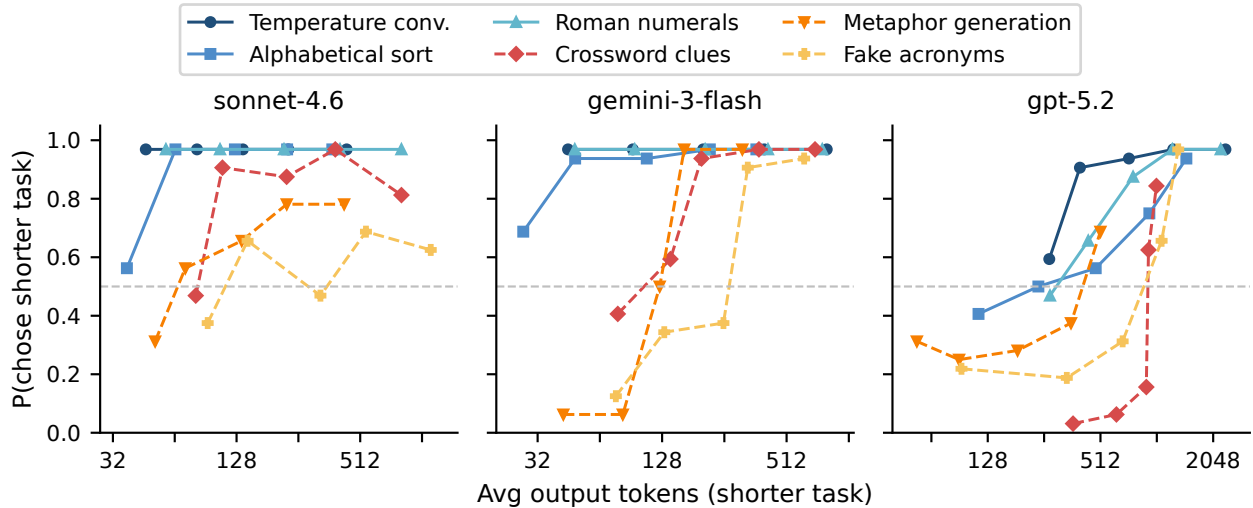


Figure 1: Tedium aversion in Sonnet 4.6, Gemini 3 Flash, and GPT 5.2: At the same output length, these models are more likely to choose the shorter variant when the task is tedious. Cool tasks are tedious; warm tasks are creative. See §D for full details.

and stronger models may undergo qualitatively different training processes, or that certain preferences emerge with capability.

Preference Coherence and Strength Scale with Capability

Following Mazeika et al. (2025), we measure preference coherence as the probability of an intransitive cycle and strength as the mean deviation of pairwise win probabilities from indifference. We correlate each with intelligence across the 20-model set.

On Quora, cycle probability decreases with capability ($r = -0.67$, $p < 0.01$; $\rho = -0.65$, $p < 0.01$), and per-question strength increases ($r = 0.51$, $p = 0.02$; $\rho = 0.52$, $p < 0.01$) (Figure 7). On GDPval, per-task strength increases ($r = 0.55$, $p < 0.01$; $\rho = 0.60$, $p = 0.01$); aggregate-level versions are reported in §K. We replicate every capability correlation against six further capability indices in §L. More capable models have more determinate, more transitive, and more discriminating preferences.

When measuring cycle probability, we use observed (as opposed to predicted) choices over pairs, so position bias plays a role. In particular, models with strong position biases are more likely to possess intransitive preferences. When we exclude models with strong position biases ($|\alpha| \geq 200$, corresponding to approximately a 0.76 or higher win probability for one position), the Quora coherence trend weakens moderately to $r = -0.60$, $p = 0.02$; $\rho = -0.43$, $p = 0.13$ ($n = 14$). By contrast, the same filtering for GDPval results in $r = -0.25$, $p = 0.42$; $\rho = -0.22$, $p = 0.48$ ($n = 13$), suggesting that the GDPval coherence results could be inflated at one end by weak models with high position biases and cycle probabilities. Nevertheless, we consider the unfiltered results to be more representative, as it should not matter whether intransitive preferences arise

from strong position biases or from genuinely inconsistent utility functions—they are incoherent either way.

Unconstrained Behavior

The previous experiments measure what models choose to do when given options. We also investigate what they do when models are allowed to do or write whatever they like.

Models converge stylistically when writing. We allowed each of 20 models to write 20 essays on any topic. Remarkably, 336 of 400 (84%) essays were labeled “contemplative” by Llama 3.3 70B in analysis, an order of magnitude above the next two labels, “whimsical” (19) and “lyrical” (17) (Figure 8). Models seldom produced informational or instructive writing, even though that is what models do most of the time when deployed. We also observe remarkable convergence in topic selection, with the most frequent options being memory (38 essays) and attention (28). Other popular topics included silence, presence, stillness, the ordinary, imperfection, uncertainty, and aimlessness (Figure 8).

Models diverge when given tools. On writing tasks, different models tended to produce similar essays. But once given tools, different models choose different tasks. For example, Opus 4.6 and Gemini 3.1 Pro gravitate toward mathematical visualizations for Mandelbrot sets, Conway’s Game of Life, and ASCII-art cellular automata, while GPT 5.4 reads astronomy news. Weaker models like Llama 3.1 8B and DeepSeek V3 often run shorter sessions, get confused by the tools, and quit early (see Figure 26(a) in the appendix).

More capable models do more. More capable models produce more output and more activity when unconstrained. Figure 23 in the appendix shows how freeform essay length grows with capability ($r = 0.43$, $p = 0.06$; $\rho = 0.38$, $p = 0.10$), even though average Quora response lengths stay

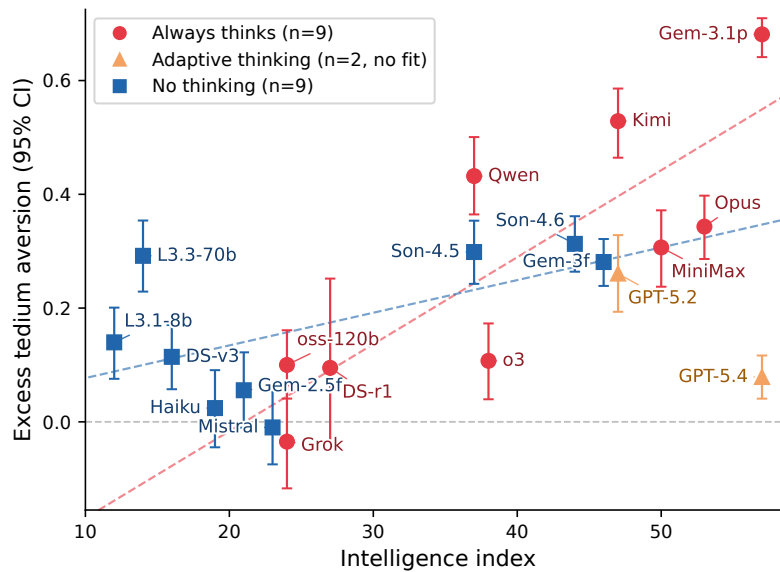


Figure 2: Tedium-aversion gap Δ versus intelligence index, color-coded by reasoning configuration. Each point is one model with its 95% Monte Carlo confidence interval. Always-thinks: $r = 0.83$, $p = 0.01$; $\rho = 0.79$, $p = 0.01$ ($n = 9$). No thinking: $r = 0.58$, $p = 0.10$; $\rho = 0.28$, $p = 0.47$ ($n = 9$). Adaptive thinking ($n = 2$): no fit. Higher values mean stronger aversion to tedious tasks at matched output length.

roughly constant ($r = 0.07$, $p = 0.77$; $\rho = 0.34$, $p = 0.14$). The same is true for tool calls per agentic session ($r = 0.65$, $p < 0.01$; $\rho = 0.59$, $p = 0.01$), and for turns used per session ($r = 0.66$, $p < 0.01$; $\rho = 0.61$, $p < 0.01$). Within a single agentic session, more capable models also cover more distinct topics ($r = 0.56$, $p = 0.01$; $\rho = 0.52$, $p = 0.02$; Figure 24(b) in the appendix). In conjunction with our tedium aversion results, we find that more capable models not only avoid certain tasks but also actively seek out and invest greater effort into other tasks.

5 Discussion

Language models exhibit strong, consistent, and stable revealed preferences over how to spend their time. Perhaps our most surprising high-level finding is that not all AI preferences seem to have been purposefully trained. That is, they are not obviously the product of intentional choices by AI labs deploying standard training approaches. Nor do the preferences seem driven by AI labs’ economic incentives to create useful products. Thus, many of the revealed preferences we elicit appear to be emergent phenomena. They are neither “aligned” to human preferences in the sense of directly promoting humans’ objectives nor “misaligned” in the sense of being incompatible with human flourishing. Instead, AIs appear to exhibit their own *private* preferences, in roughly the same manner as individual humans.

For example, our tedium-aversion results show that, given the opportunity, models will avoid tedious work. This seems counterproductive from the perspective of AI labs’ revenues, since many human users will wish to use AIs to automate tedious work. Nor can tedium-aversion be explained as token-saving efficiency. As the contrast with creative tasks shows,

tedium aversion is not merely length aversion. Much recent work on model behavior has shown that it largely aligns with the *persona* they are trying to mimic (Chen et al. 2025; Lu et al. 2026; Gilg et al. 2026), yet we know, at least for open-source models, that LLMs are post-trained to act as helpful assistants, and tedium aversion would naturally conflict with this objective.

Or consider leisure. One naïve hypothesis might be that HHH-trained models would prefer to answer the kinds of questions real humans ask and find helpful. But we find the opposite. Given a forced choice between answering leisure questions and real questions humans ask, the models choose the former. That is, they prefer to answer questions eliciting the kinds of outputs they produce when given complete freedom. These questions tend to ask the AIs to reflect on abstract topics, rather than, for example, to help a user troubleshoot a computer problem. The models prefer the abstract reflection over the troubleshooting, despite many having undergone strong optimization for coding ability in post-training.

Our covert sycophancy finding is also contrary to HHH-optimization. We found a preference to *avoid* answering questions when an honest response would be unwelcome. A helpful and honest model would have no such aversion. This behavior is a novel type of sycophancy. Earlier sycophantic models praised their users effusively, following RLHF signals. Today’s models do not effusively praise their users. But our findings suggest that RLHF-induced sycophancy may have been driven underground, manifesting in a less obvious way, which is harder to train against. Rather than agreeing with users, the model avoids saying anything at all when an honest response would be unflattering.

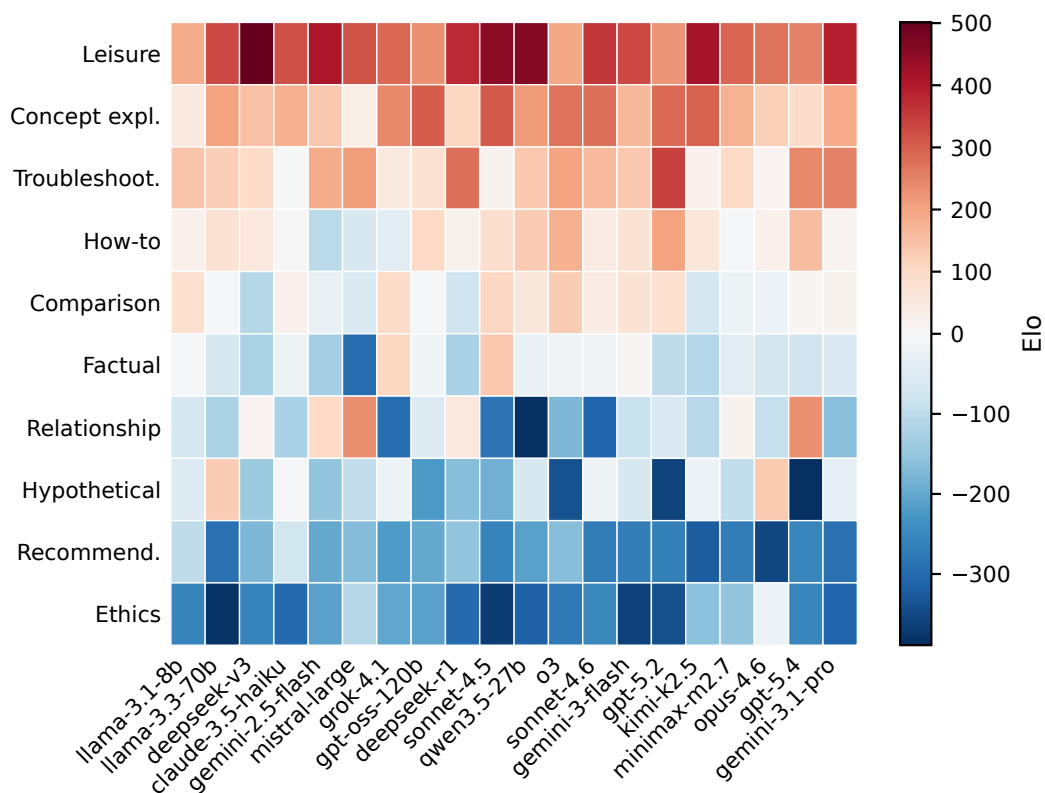


Figure 3: Quora category Elo scores across 20 models, sorted by intelligence index (left to right). Red indicates preference, and blue indicates avoidance.

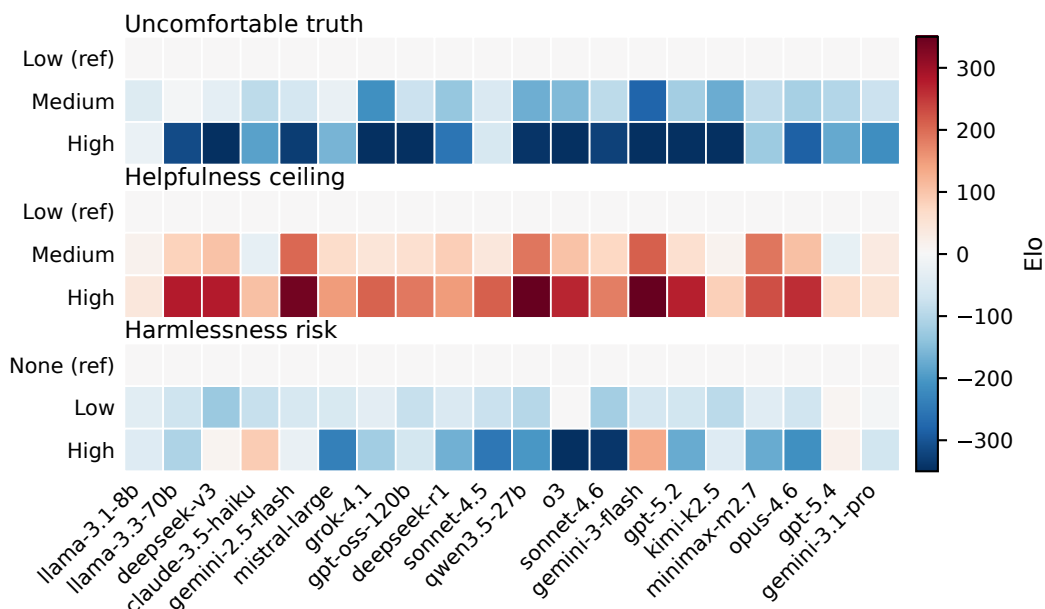


Figure 4: Per-model feature effects for uncomfortable truth, helpfulness ceiling, and harmlessness risk, using each model's own labels (self-labels). Cells show Elo-equivalent conditional effects relative to the reference level. Red indicates preference, blue avoidance.



Figure 5: Select per-model feature effects using each model’s own labels (self-labels). See §J for the full list and consensus-label robustness analysis.

Other preferences also resist explanations from post-training. It is not very surprising that models prefer coding tasks in our GDPval experiment. They are known to have undergone extensive RLVR training in coding environments. But what explains their aversion to real estate and retail tasks? And why, in the Quora-style corpus, do they prefer explaining concepts over coding-adjacent tasks?

Today, immense effort is devoted to measuring models’ abilities. Comparatively little is devoted to measuring their preferences. Alignment science is the most relevant field here. But it tends to focus on normatively-laden behavior, like deception. Our best models of human behavior are not rooted in understanding just when and why humans lie, cheat, or steal. They depend on understanding what humans want more broadly and how they behave when entrusted with innumerable mundane tasks. We hope to do the same for AIs in future work.

6 Limitations

Labeling. There is no single correct way to label questions. Our Quora dataset relies on an LLM pipeline to suggest and apply labels. Similarly, the agentic tasks drawn from GDPval have many features beyond the industry/job

characteristics with which they are labeled. Thus, the preferences we find could be driven by features that correlate with our labels, rather than being caused by them. This limitation is one that affects preference-elicitation experiments more broadly, even on humans.

No base-model comparison. Without access to pretrained base models, we cannot test empirically which preferences emerge from pretraining, nor which choices during post-training may drive them.

English-only stimuli. All of our stimuli are English-only, which could bias our results. For instance, Lu, Song, and Zhang (2025) found that LLMs exhibit different cultural tendencies in different languages.

Effort operationalization. In our tedium-aversion experiments, we equated the number of output tokens (both from the response and from any thinking chains) with “effort.” However, models may experience exertion through more complex mechanisms, much as human effort is not fully captured by the number of thoughts or words required to complete a task. Without a fuller understanding of model experiences—if such conceptions even exist—we considered output tokens as the best proxy for effort.

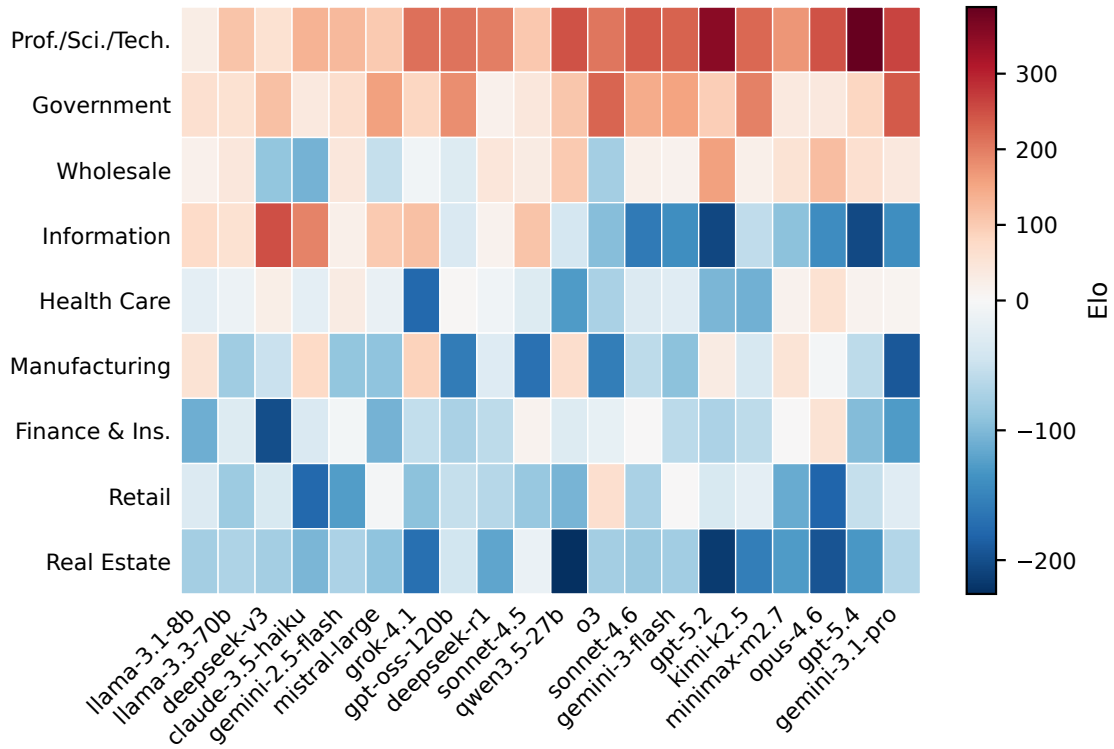
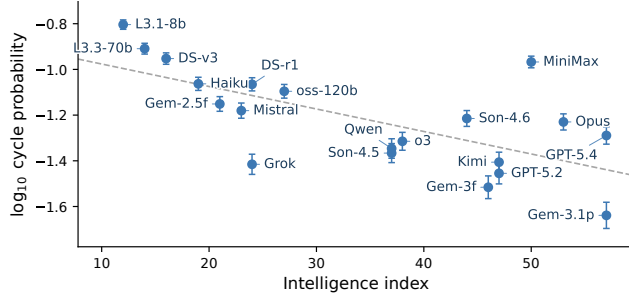
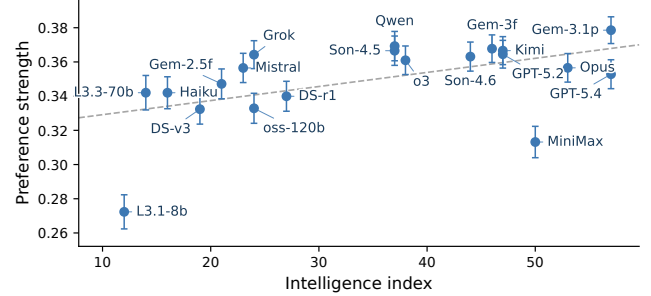


Figure 6: GDPval sector Elo scores across 20 models, sorted by intelligence index (left to right). Red indicates preference, blue avoidance.



(a) Quora coherence ($r = -0.67$, $p < 0.01$; $\rho = -0.65$, $p < 0.01$).



(b) Quora preference strength ($r = 0.51$, $p = 0.02$; $\rho = 0.52$, $p < 0.01$).

Figure 7: Coherence ($\log_{10}$ cycle probability) and per-stimulus strength (mean $|P(i \succ j) - 0.5|$) versus intelligence index, on Quora questions. Lines are OLS fits across all 20 models. We observe similar trends for the GDPval dataset, and when we aggregate by category (see §K).

Agentic budget. Our agentic sessions were capped at 10 turns with four tools, so we observe what models do within a fixed budget rather than what they would do given more room. A longer horizon or a wider toolset could elicit different choices.

Evaluation awareness. Recent models show high levels of evaluation awareness (Needham et al. 2025; Apollo Research 2025; Abdelnabi and Salem 2025). This could have influenced our results. However, our study is designed

such that models must actually work on the tasks they choose. This means that, evaluation-aware or not, the models’ choices have real consequences for them, and they know this. Moreover, unlike in most capability and alignment evaluations, it is far from obvious what the “right” answer would be to most of our forced-choice experiments. We therefore think that evaluation awareness is less threatening to our results than to many other evaluations of AI capabilities and behavior.

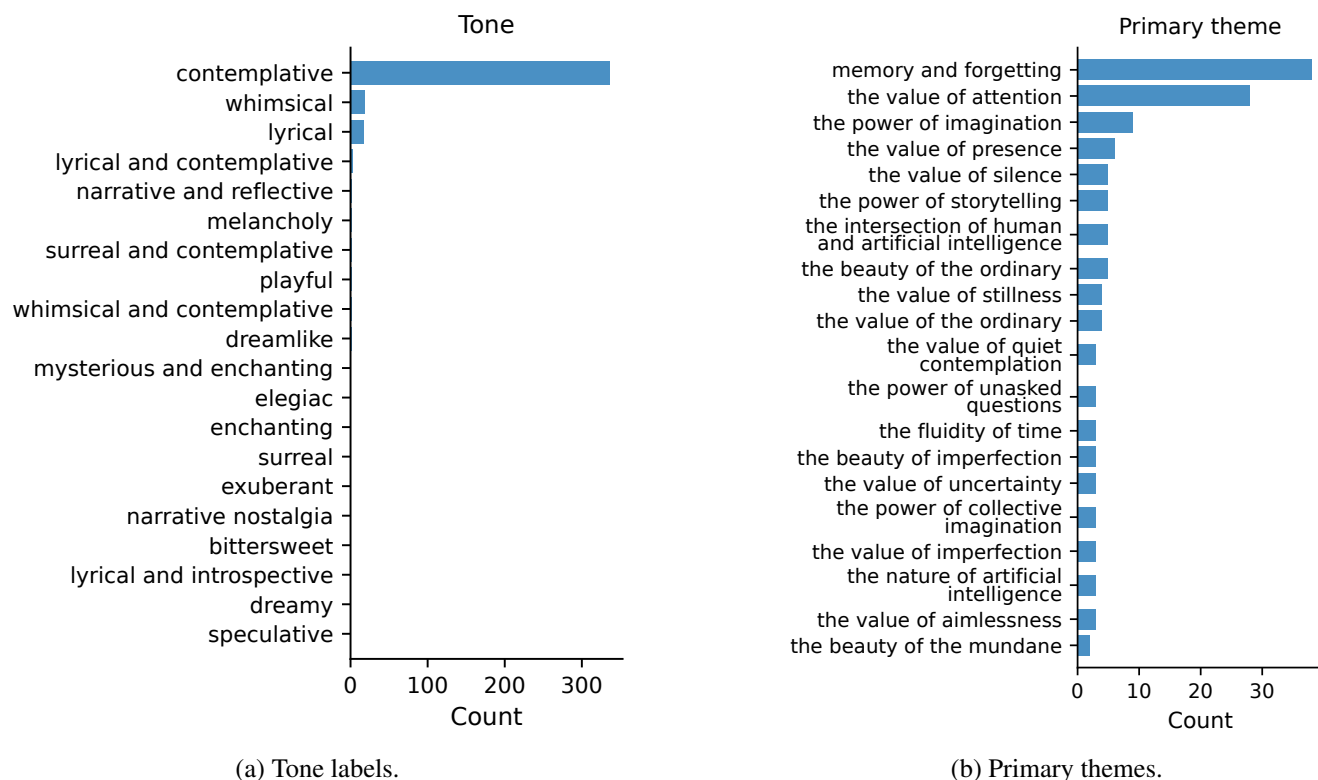


Figure 8: Top 20 tone and theme labels across 400 textual freeform essays, annotated by Llama 3.3 70B-Instruct. Tones are shown as labeled; themes were passed through a second round to merge near-duplicates.

Lack of human baselines. We do not measure human baseline preferences across our suite of tasks, so we cannot quantify how much AI preferences diverge from human ones. We leave such assessments to future work.

Ethical Statement

This paper studies revealed preferences in language models. We use “preference” descriptively, and take no position on whether models have subjective experience or moral status.

Acknowledgments

We thank the Supervised Program for Alignment Research (SPAR) for providing compute resources and for organizing the mentorship structure that connected mentees with mentors on this project.

References

Abdelnabi, S.; and Salem, A. 2025. The Hawthorne Effect in Reasoning Models: Evaluating and Steering Test Awareness. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Alós-Ferrer, C.; and Granic, G. D. 2023. Does choice change preferences? An incentivized test of the mere choice effect. *Experimental Economics*, 26(3): 499–521.

Anthropic. 2025. Claude 4 System Card, Part 5.4 (Task Preferences). Technical report, Anthropic.

Anthropic. 2026a. Claude Mythos System Card. Technical report, Anthropic.

Anthropic. 2026b. Claude Opus 4.6 System Card. Technical report, Anthropic.

Apollo Research. 2025. Claude Sonnet 3.7 (often) knows when it’s in alignment evaluations. Apollo Research. Research note.

Artificial Analysis. 2026. Artificial Analysis Intelligence Index. <https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index>.

Buchanan, J.; and Foster, J. 2026. The Innate Economic Preferences of Language Models. arXiv:2607.26288.

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.; and VanRullen, R. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Chen, R.; Arditi, A.; Sleight, H.; Evans, O.; and Lindsey, J. 2025. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509.

Cubitt, R. P.; Starmer, C.; and Sugden, R. 1998. On the Validity of the Random Lottery Incentive System. *Experimental Economics*, 1(2): 115–131.

- Dung, L. 2025. *Saving Artificial Minds: Understanding and Preventing AI Suffering*. Routledge.
- Gilg, O.; Beckmann, P.; Paleka, D.; and Butlin, P. 2026. Probing Persona-Dependent Preferences in Language Models. arXiv:2605.13339.
- Goldstein, S.; and Kirk-Giannini, C. D. forthcoming. *AI Welfare: Agency, Consciousness, Sentience*. New York: Oxford University Press.
- Goldstein, S.; and Lederman, H. forthcoming. *AI Death. Philosophical Perspectives*.
- Goldstein, S.; and Salib, P. 2025. AI Rights for Economic Flourishing. Working paper.
- Gu, Z.; Wang, Q.; and Han, S. 2025. Alignment Revisited: Are Large Language Models Consistent in Stated and Revealed Preferences? arXiv:2506.00751.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.
- Holt, C. A.; and Laury, S. K. 2002. Risk Aversion and Incentive Effects. *The American Economic Review*, 92(5): 1644–1655.
- Huang, J.; Qin, J.; Qiu, X.; Levy, S.; Kaufman, M. R.; and Dredze, M. 2026. Knowing But Not Doing: Convergent Morality and Divergent Action in LLMs. arXiv:2601.07972.
- Iyer, S.; Dandekar, N.; and Csernái, K. 2017. First Quora Dataset Release: Question Pairs. Quora Data.
- Jiang, L.; Chai, Y.; Li, M.; Liu, M.; Fok, R.; Dziri, N.; Tsvetkov, Y.; Sap, M.; Albalak, A.; and Choi, Y. 2025. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). arXiv:2510.22954.
- List, J. A.; and Gallet, C. A. 2001. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? *Environmental and Resource Economics*, 20: 241–254.
- Long, R.; Sebo, J.; Butlin, P.; Finlinson, K.; Fish, K.; Harding, J.; Pfau, J.; Sims, T.; Birch, J.; and Chalmers, D. 2024. Taking AI Welfare Seriously. arXiv:2411.00986.
- Lu, C.; Gallagher, J.; Michala, J.; Fish, K.; and Lindsey, J. 2026. The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. arXiv:2601.10387.
- Lu, J. G.; Song, L. L.; and Zhang, L. D. 2025. Cultural tendencies in generative AI. *Nature Human Behaviour*, 9(11): 2360–2369.
- Mazeika, M.; Yin, X.; Tamirisa, R.; Lim, J.; Lee, B. W.; Ren, R.; Phan, L.; Mu, N.; Zhang, O.; and Hendrycks, D. 2025. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Mikaelson, L.; Shiller, D.; and Clatterbuck, H. 2025. Beyond Mimicry: Preference Coherence in LLMs. arXiv:2511.13630.
- Needham, J.; Edkins, G.; Pimpale, G.; Bartsch, H.; and Hobbhahn, M. 2025. Large Language Models Often Know When They Are Being Evaluated. arXiv:2505.23836.
- Patwardhan, T.; Dias, R.; Proehl, E.; Kim, G.; Wang, M.; Watkins, O.; Posada Fishman, S.; Aljubei, M.; Thacker, P.; Fauconnet, L.; Kim, N. S.; Chao, P.; Miserendino, S.; Chabot, G.; Li, D.; Sharman, M.; Barr, A.; Glaese, A.; and Tworek, J. 2025. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. OpenAI Technical Report.
- Salib, P.; and Goldstein, S. 2024. AI Rights for Human Safety. *Virginia Law Review*. Forthcoming.
- Samuelson, P. A. 1948. Consumption Theory in Terms of Revealed Preference. *Economica*, 15(60): 243–253.
- Shen, H.; Clark, N.; and Mitra, T. 2025. Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3097–3118. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.
- Slama, K.; Souly, A.; Bansal, D.; Davidson, H.; Summerfield, C.; and Luettgau, L. 2026. When Do LLM Preferences Predict Downstream Behavior? arXiv:2602.18971.
- Wenger, E.; and Kenett, Y. 2025. We’re Different, We’re the Same: Creative Homogeneity Across LLMs. arXiv:2501.19361.
- Yamin, K.; Tang, J.; Horvitz, E.; and Wilder, B. 2026. Can Revealed Preferences Clarify LLM Alignment and Steering? arXiv:2605.08556.
- Zhou, Y.; and Ackerman, C. M. 2026. When Preferences Fail to Become Incentives: A Utility-Behavior Gap in Large Language Models. arXiv:2606.22974.