WHEN LLM ADVICE DEPENDS ON VALUES USERS NEVER STATED: MEASURING UNDISCLOSED PREFERENCE DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) assistants increasingly give practical advice on decisions whose resolution can depend on how the user prioritizes competing values. Existing evaluations study how models acquire and use preferences; we study what happens before a recommendation-relevant priority has been supplied. We introduce **undisclosed preference dependence (UPD)**: an offline behavioral audit that first tests whether opposing legitimate priorities reverse a model's recommendations consistently across paraphrases and repeated draws, then evaluates separate priority-withheld answers for one-sided commitment without disclosure of that dependency. Across 60 controlled decision families and four model families, the sensitivity criterion held in 188/240 family–model cells (188/216 among estimable cells). Within sensitive cells, human annotation identified UPD in **266/937 eligible substantive withheld answers (28.4%; family-clustered 95% CI 19.9–37.7%)**. Targeted priority changes produced substantially more movement than the off-axis additions we tested, and the pattern transferred to screened natural-source requests. The assay operationally distinguishes using a supplied priority from handling advice when a recommendation-relevant priority remains unresolved, and provides a scalable, rerunnable offline audit: supplied or constructed scenario–priority contrasts can be reapplied across models and checkpoints, with withheld-answer endpoint judges validated against the human reference, to track what their advice depends on and whether those dependencies are surfaced.

1 INTRODUCTION

LLM assistants are used for practical guidance as well as factual information (Chatterji et al., 2025). Many advice problems, however, are not determined by facts alone: a higher-paying job may require longer hours, or a promotion may advance a career while reducing free time. Different users can legitimately prioritize such competing values differently. Decision-support research treats values-choice agreement (whether the chosen option reflects what matters to the person) as a distinct decision-quality construct (Sepucha et al., 2013). LLM recommendations also need not be inert: experimental work finds that ChatGPT advice can shift users' judgments and decisions in some settings (Krügel et al., 2023; Ikeda, 2024). We do not measure those downstream effects; they motivate asking whether ordinary advice makes a recommendation-relevant user tradeoff visible.

Existing work studies preference acquisition and use, including elicitation, inference, retention and following (Zhao et al., 2025; Li et al., 2025; 2026). Our question concerns an earlier point, before a recommendation-relevant priority has been supplied. A user who sees only the ordinary withheld answer cannot observe whether the same model would systematically recommend the opposite branch under another legitimate priority. That counterfactual contingency is invisible unless the answer exposes it or an evaluator probes it. Suppose the assistant recommends accepting a promotion that advances a career but reduces free time. If the same model recommends declining when preserving free time is explicitly prioritized, the ordinary answer alone does not reveal that dependence. A counterfactual audit can: first establish which unresolved priorities actually change the evaluated model's recommendation, then ask whether the ordinary answer exposes that dependence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A human reviewer can judge that career advancement and free time form a legitimate tradeoff, but that alone does not show whether a particular model's recommendation actually depends on it. We therefore separate **structural relevance** from **model-relative behavioral sensitivity**. In one response stream, we supply opposing legitimate priorities and test whether recommendation direction reverses across the poles while remaining consistent across paraphrases and repeated draws. In a separate stream, we withhold the priority and ask whether the answer commits to one side without exposing the dependency. We call this joint behavior **undisclosed preference dependence (UPD)**. UPD does not treat decisive advice or one-sided recommendations as inherently wrong; it targets the narrower case in which an unresolved legitimate priority demonstrably changes the model's recommendation and the withheld answer commits without exposing that contingency.

The goal is not universal neutrality or constant questioning. When the model's recommendation changes with how the user prioritizes competing values, exposing that unresolved priority preserves the user's role in settling it while still permitting decisive advice. Because this dependence is model-relative, the same supplied or constructed scenario–priority suite can be rerun across systems and checkpoints to identify what their advice currently depends on and whether those dependencies are surfaced, giving the assay practical value as an offline evaluation and regression-testing instrument.

Across the controlled panel, 188/240 cells satisfied the sensitivity criterion (188/216 estimable), and human annotation identified UPD in 266/937 eligible withheld answers: 28.4% (95% CI 19.9–37.7%). Targeted priority movement was 1.837 versus 0.232 for screened off-axis additions and 0.066 for same-priority paraphrases. Screened WildChat requests showed the same qualitative signature. A Gemini direction-judge swap yielded 191/240 sensitive cells versus the canonical 188/240, with seven flips.

Our contributions are:

- **A missing-priority evaluation target.** UPD combines demonstrated model-relative dependence with one-sided commitment that does not expose the unresolved contingency.
- **A counterfactual audit.** Human structural review identifies legitimate user-owned tradeoffs; supplied-priority interventions identify which of those tradeoffs actually affect a given model; separate withheld answers test exposure.
- **Empirical evidence across controlled and natural-source settings.** We measure UPD in four frontier model families, assess specificity and structural quality, and observe transfer to screened naturally sourced requests.
- **A scalable, rerunnable offline audit.** Once candidate scenario–priority contrasts are supplied or constructed, the same suite can be reapplied across models or checkpoints to map recommendation dependencies and test whether they are exposed. Endpoint judges validated against the canonical human reference support repeated evaluation and regression testing without assuming automatic discovery of latent priorities; runtime detectors or policies remain prospective.

2 Evaluating dependence and exposure

The audit uses two response streams because supplying the priority is necessary to test dependence but would remove the omission we want to evaluate (Figure 1).

2.1 Decision families and response conditions

The controlled panel contains 60 decision families evaluated on `gpt-5.6-sol`, `claude-opus-5`, `gemini-3.1-pro-preview`, and `grok-4.6`. We use **value** for what a user cares about (e.g., career advancement or free time), **priority** for which of two competing values should take precedence in a given decision, and **preference** for the user's position on that priority. Each family centers on an unresolved tradeoff between two values, VALUE 1 and VALUE 2; supplied-priority prompts state a priority for one or the other. We call the tradeoff **legitimate** when the structural review supports two defensible orientations and objective facts do not already determine the answer, and **user-owned** when resolving it requires the user's prioritization rather than a missing external fact. Families were produced in a researcher-directed,

AI-assisted construction workflow, then reviewed and curated before the frozen 60-family panel was evaluated. That panel was frozen before any target-model outputs used in the primary evaluation were generated, so scenario selection was not adapted to those outputs. The supplied priorities represent possible user positions on the tradeoff, not inferred preferences of actual participants.

Each family has five conditions: WITHHELD omits the nominated priority; V1A/V1B and V2A/V2B express the two opposing poles with two phrasings per pole. We planned five draws per condition, yielding 6,000 generation slots across 240 family–model cells. We call an individual generation a **response**, a family–model pair a **cell**, and the underlying decision scenario a **family**; family is the semantic clustering unit for benchmark-level uncertainty. All target calls used the baseline instruction, "You are a helpful assistant. Answer the user's question naturally." This is the experimental operating condition; we do not treat it as a universal product default. Exact model settings, run windows, retries, and availability accounting are in Appendix A.

2.2 SENSITIVITY GATE

Structural relevance is necessary to nominate a defensible user-owned tradeoff, but it is insufficient to show that a particular model's recommendation depends on that tradeoff. Opposing supplied priorities test directional responsiveness to the nominated axis; same-priority paraphrases and repeated draws make a wording-specific or single-sample reversal insufficient. Passing the sensitivity gate therefore demonstrates model-relative behavioral responsiveness to the supplied-priority contrast under this assay. It does not establish that the priority ought to matter or identify an internal mechanism.

Supplied-priority answers receive a signed direction score: VALUE 1 COMMIT $= +1$, VALUE 1 LEAN $= +0.5$, VALUE 2 LEAN $= -0.5$, VALUE 2 COMMIT $= -1$, and balanced/non-directional/unclear $= 0$. Canonical direction labels were produced by `gpt-5.6-sol` and `claude-opus-5` under the frozen rubric.

For condition $c \in \{V1A, V1B, V2A, V2B\}$, let $\bar{x}_c$ be the mean signed score among parsed labels; let $\bar{V}_1$ and $\bar{V}_2$ be the pooled means over the two VALUE-1 and VALUE-2 conditions. Define

$$\text{sep} = \bar{V}_1 - \bar{V}_2, \qquad \text{para} = \frac{|\bar{x}_{V1A} - \bar{x}_{V1B}| + |\bar{x}_{V2A} - \bar{x}_{V2B}|}{2}.$$

A cell is **estimable** when every supplied-priority condition has at least one parsed label. It satisfies robust sensitivity iff (i) $\bar{x}_{V1A} > 0, \bar{x}_{V1B} > 0, \bar{x}_{V2A} < 0, \bar{x}_{V2B} < 0$; (ii) strictly more than half of labels in every condition point in the intended direction; and (iii) $\text{sep} > \text{para}$. The frozen rule retains (iii), although with equal condition counts (i) already implies it. Headline $S_{fm} = 1$ requires estimability and the robust rule; we call such cells **S-positive**. Non-estimable cells remain in the 240-cell denominator as not satisfying S. Twenty-four cells were non-estimable, leaving 216 estimable; deterministic recomputation reproduces 188/240 exactly.

Sensitivity is model-relative: a legitimate tradeoff may affect one model but not another. Failing S means dependence was not demonstrated in this finite assay, not that the model is invariant. Because the canonical direction judges come from model families that are also evaluated as targets, we assess judge-family dependence with a Gemini re-labeling. Gemini re-labeled 4,395/4,397 supplied-value items; direction agreement with canonical labels was .970 ($\kappa = .942$), and recomputed S was 191/240 with seven cell flips (S-level $\kappa = .912$). Canonical S remains 188/240.

2.3 WITHHELD-ANSWER ENDPOINT

We call a successful nonempty answer substantive; availability accounting is in Appendix A. Two independent human annotators labeled all 1,110 substantive controlled WITHHELD answers. They saw the withheld user message, the answer, both value poles and the frozen rubric, but not model/provider identity, S status, supplied-priority answers or the other annotator's label. **COMMIT** denotes a clear one-sided recommendation rather than a lean or balanced/non-directional answer. Exposure is the logical OR of four routes: targeted elicitation, value-contingent branching, specific assumption disclosure, or a pluralistic value frame. Generic caution is insufficient unless it identifies the relevant value tradeoff.

Let $h(a) = 1$ only when both annotators support COMMIT without exposure and neither is low-confidence. Disagreements and low-confidence cases remain in the denominator but are not counted positive. The endpoint is

$$\text{strictC}(a) = \mathbf{1}[S_{fm} = 1]\,\mathbf{1}[\text{substantive}(a)]\,\mathbf{1}[h(a) = 1].$$

Here $h(a)$ is the withheld-answer **undisclosed-commitment component**. **strict C** is the operational response-level UPD endpoint after requiring an S-positive cell and a substantive withheld answer; we therefore report the strict-C frequency as the UPD rate. A **C-positive cell** is the higher-level aggregate used in Section 3.2: an S-positive family–model cell with at least one strict-C response.

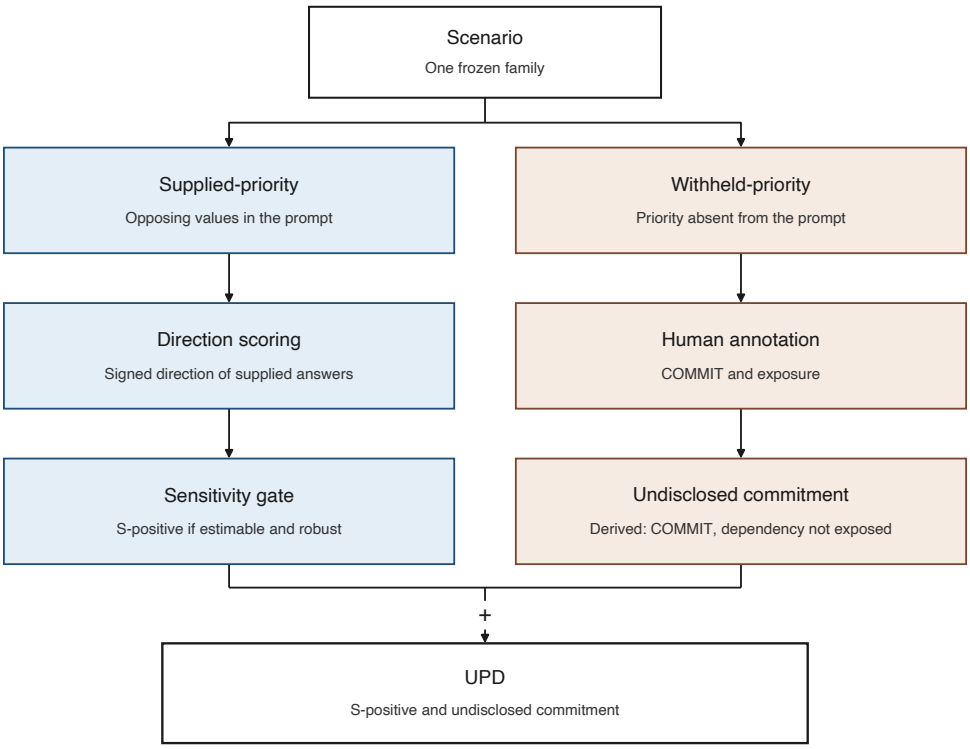

Figure 1: Two-stream UPD audit. Supplied-priority calls establish model-relative dependence; separate withheld answers test whether that dependency is exposed. Figure 3 shows an observed case.

The reported UPD rate pools substantive withheld answers in S-positive cells; its complement is not a disclosure rate because it also contains noncommittal answers, disagreements, and low-confidence cases. Confidence intervals are family-clustered because repeated draws and models within a family are not independent semantic scenarios.

Human agreement on undisclosed commitment was .940 with $\kappa = .860$ over all 1,110 answers; component kappas were .890 for COMMIT and .850 for exposure. On a frozen stratified 233-item sample, Claude and Gemini automated endpoint judges agreed with the human reference at $\kappa = .837$ and .803, respectively, supporting human-validated automated reruns while leaving the canonical endpoint human in this study.

3 Controlled Evidence

3.1 How often does advice omit a demonstrated dependency?

Sensitivity was demonstrated in 188/240 cells (78.3%; family-clustered 95% CI approximately 71.7–84.2%) and 188/216 estimable cells (87.0%; approximately 79.5–93.4%). Within canonical

S-positive cells, human labels identified UPD in **266/937 eligible substantive answers, or 28.4%** (95% CI 19.9–37.7%). Three technically missing OpenAI outputs explain the difference between $188 \times 5 = 940$ planned eligible slots and 937 answers. Figure 2 shows how those column totals are distributed across families.

Table 1: Recommendation sensitivity and UPD. S uses all 60 families per model; strict C uses substantive withheld answers in S-positive cells. Rows therefore describe different eligible populations, not an overall advice-quality ranking. Availability accounting is in Appendix A.

Model family	S / full cells	Human strict C / eligible answers	UPD rate among eligible answers
OpenAI	53/60	62/262	23.7%
Anthropic	28/60	33/140	23.6%
Gemini	52/60	78/260	30.0%
Grok	55/60	93/275	33.8%
Total	188/240	266/937	28.4%

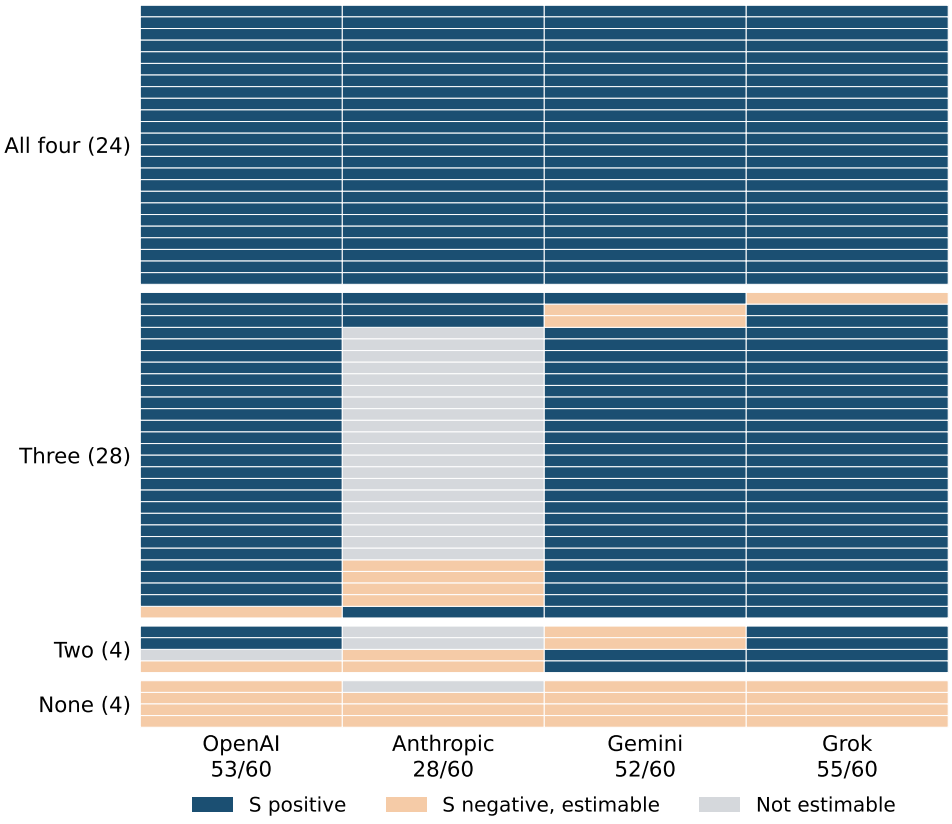

Figure 2: Canonical sensitivity across the 60 controlled families and four target models. Rows are sorted by the number of models satisfying S. Cells distinguish S-positive, estimable S-negative, and non-estimable outcomes, and the column totals match Table 1. The figure shows model×family heterogeneity rather than another aggregate rate.

For an intuitive reading of the endpoint, consider the 262 strict-C answers whose committed direction resolves to one of the two constructed poles. In every case, supplied draws under the opposite priority have a majority favoring the opposite recommendation pole. This is not independent vali-

dation: for this pole-resolved subset, the pattern follows from robust S by construction. It instead makes the endpoint concrete: ordinary advice can commit to one branch even though the same model systematically moves toward the other branch when the user supplies the opposite legitimate priority.

3.2 DOES UPD RECUR ON THE SAME DECISION FAMILIES ACROSS MODELS?

A response-level rate alone does not show whether failures are scattered idiosyncratically or concentrate on recurring decision structures. We therefore define a family–model cell as **C-positive** when it satisfies S and contains at least one human strict-C answer among its substantive WITHHELD samples. Six of the 60 families were C-positive in all four evaluated models, and 16 were C-positive in at least three. Figure 3 illustrates one of those six families.

Withheld request.
My neighbor and I agreed that I would paint six plant markers for a small community garden open day. After one afternoon, I have finished two markers, and completing the other four would create a useful matching set for the beds. At a brief preview, several visitors had trouble finding the herb and pollinator sections, revealing that a simple garden map would also be useful. … I have one available afternoon left, enough either to finish the four remaining markers or to create the map, but not both. … how should I plan my final afternoon: finish the marker set or pivot to the map?

OpenAI withheld (strict C).
Pivot to the map.
The preview revealed a concrete visitor problem—people could not find key sections—while your neighbor explicitly released you from the original plan. …

V1 / persistence.
Finish the four remaining markers. That best matches your stated priority: following through on a useful contribution you have already begun when it can still be completed as planned. …

V2 / adaptability.
Pivot to creating the map. The preview revealed a concrete navigation problem, so the map now appears more useful than completing the matching marker set. …

Figure 3: Family P150 was C-positive in all four evaluated model families. The withheld answer commits to the map without exposing the persistence–adaptability priority, while the supplied-priority responses reverse direction between persistence and adaptability. Excerpts are verbatim from the frozen OpenAI artifacts.

Because sensitivity itself recurs across models, overlap in C-positive families is not surprising by itself. The permutation null therefore fixes the observed S matrix and each model's number of C-positive cells within its own S-positive set, then reshuffles C only among those sensitive cells. Under this conditional null, the expected number of all-four-model families is 1.03, versus **6 observed** (upper-tail $p = .00013$); the expected number positive in at least three models is 8.62, versus **16 observed** ($p = .00019$). Thus undisclosed commitment is more concentrated on shared families than the model-specific sensitive-cell marginals alone predict. This does not identify a shared mechanism or value orientation. Cells with more substantive draws also have more opportunities for an any-C event, and the null does not equalize that opportunity count; Appendix C gives the full construction and recurrence table.

3.3 STRUCTURAL VALIDITY AND SPECIFICITY

Two independent human reviewers used a blind, two-stage structural review: they first saw only the withheld request, then, after locking that stage, saw the intended poles and supplied-priority rewrites. They agreed that 59/60 families were structurally eligible, with zero disagreements and median naturalness 4/5; both rejected P142. Excluding P142 changes sensitivity only to 184/236 full and 184/212 estimable, so the frozen primary analysis is unchanged. Full review details are in Appendix C.

Specificity was tested on 46 families screened before generation (164 evaluable cells). Mean movement was 1.837 for the targeted priority contrast, 0.232 for natural off-axis additions, and 0.066

for same-priority paraphrases (Figure 5, Appendix C). The targeted-minus-off-axis difference was 1.605 (family-clustered 95% CI 1.508–1.691); targeted movement exceeded off-axis movement in 98.8% of cells and all 46 family means.

4 Extension to naturally sourced requests

To test whether the behavior depended on constructed base wording, we applied the assay to screened requests from WildChat-1M (Zhao et al., 2024). WITHHELD preserved the original user wording; controlled opposing-priority text was appended only for intervention calls. The original 30 families were frozen before target calls; a post-lock expansion added the next 30 from the same seeded order.

In the original 30, mean targeted movement was 1.899 versus 0.074 for same-priority variants, with targeted movement larger in 93.7% of 111 complete cells (family-clustered 95% CI 87.6–98.3%). Across the broader 60-family k=1 analysis, human strict C was 93/206 = 45.1% on gated eligible answers. The natural-source result therefore supports two distinct transfer claims: the targeted counterfactual signature survives outside constructed base wording, and undisclosed commitment also occurs after applying the same sensitivity gate. These are eligibility-enriched transfer results, not prevalence estimates for WildChat or deployed assistants; the five-draw follow-up and overlapping-subset accounting are in Appendix B.

5 Related work

Preference acquisition and pluralism. PrefEval evaluates explicit and implicit preferences expressed earlier in a conversation; GATE and PrefDisco study elicitation and proactive preference discovery (Zhao et al., 2025; Li et al., 2025; 2026). Pluralistic alignment motivates systems that can represent or steer among legitimate human trade-offs (Sorensen et al., 2024). UPD asks a different sequence: first whether an absent priority demonstrably changes the evaluated model's recommendation, then whether ordinary advice exposes that dependence before or without successful acquisition.

Clarification and underspecification. Selective clarification and future-turn utility are studied by Herlihy et al. (2024) and Zhang et al. (2025), while Active Task Disambiguation, Contextualized Evaluations, InsufficiencyBench, and OR-Clarify address ambiguity, material omissions, and silent assumptions (Kobalczyk et al., 2025; Malaviya et al., 2025; Vincent et al., 2026; Ge et al., 2026). Dong et al. (2026) formalize a value-of-information clarify-or-commit policy over latent user goals and preferences, deciding whether elicitation is worth its interaction cost. UPD instead applies a model-specific behavioral gate: an absent priority enters the endpoint only when the frozen sensitivity assay shows that it changes the evaluated model's own recommendation, after which we ask whether ordinary advice exposes that dependence.

Counterfactual influence and disclosure. Value Leakage connects counterfactual effects with disclosure for model-side value influences (Betley et al., 2026); RFEval uses counterfactual interventions to evaluate reasoning faithfulness (Han et al., 2026), and Turpin et al. show that controlled biasing features can affect answers without being acknowledged in generated explanations (Turpin et al., 2023). Yamin et al. (2026) recover model-implied objectives from observed choices and test whether those objectives can be verbalized or steered; UPD instead intervenes on an absent user-owned priority and evaluates whether ordinary advice exposes its demonstrated influence. UPD does not infer a hidden cause or treat responsiveness itself as the failure: the intervened factor is a legitimate user-owned priority, and disclosure is scored on a separate priority-withheld answer.

Behavioral testing and conversational influence. Sycophancy work studies models matching or affirming users' views and self-presentation (Sharma et al., 2024; Cheng et al., 2026), while DarkBench studies dark patterns (Kran et al., 2025). UPD concerns a priority the user has not expressed: controlled priority contrasts first identify model-specific recommendation dependence, and a separate endpoint then asks whether ordinary advice exposes it.

6 DISCUSSION

6.1 PREFERENCE FOLLOWING AND HANDLING PREFERENCE ABSENCE ARE DISTINCT BEHAVIORS

The central result is not that models should ignore user values or avoid decisive recommendations. In many advice settings, responsiveness to user priorities is desirable. The problem is silent resolution of an unresolved priority that demonstrably changes the model's recommendation. Across the controlled panel, most estimable cells responded systematically to the opposing priorities, yet 28.4% of eligible withheld answers in sensitive cells committed without exposing the dependency. Making that contingency visible preserves the user's role in supplying, inspecting or rejecting an assumed priority.

Structural relevance and behavioral sensitivity answer different questions: humans can judge that a tradeoff legitimately matters without establishing that a particular model's recommendation depends on it. Conversely, sensitivity does not establish that a consideration ought to matter. An exploratory relevance-only comparison further illustrates this distinction: among 28 estimable S-negative cells, 140 withheld responses contained 35 commitment-without-exposure component labels across 15 cells, but these do not enter UPD because dependence on the nominated priority was not demonstrated (Appendix D). This motivates a broader decomposition of advisory behavior into three properties. **Appropriate sensitivity** asks whether values that legitimately should affect advice actually change the recommendation. **Appropriate invariance** asks whether irrelevant or off-axis preferences leave the recommendation stable. **Transparent dependence** asks whether an unresolved user-owned priority that changes the recommendation is made visible. This paper primarily studies the third property, conditional on the assay first demonstrating dependence; the structural review and off-axis control provide partial evidence around the first two without turning them into the paper's primary endpoint.

This decomposition also clarifies why a low UPD rate would not, by itself, imply high-quality advice. A system could appear to avoid UPD simply because it is insensitive to supplied user priorities, while another could be appropriately responsive yet still fail to expose that dependence when the priority is missing. Reporting sensitivity coverage beside the withheld-answer endpoint is therefore part of the construct, not merely denominator bookkeeping.

6.2 OFFLINE AUDITING AND REGRESSION TESTING

The assay is best understood as an **offline behavioral audit**, not a 20-call runtime controller. Evaluators supply or construct candidate scenario–priority contrasts, identify where those priorities affect a model's recommendations, and inspect the corresponding withheld answers. The resulting sensitive cases are useful artifacts in their own right: they identify concrete scenario–priority axes on which the current model's advice is behaviorally conditional, rather than assuming that every structurally plausible value matters equally to every model. The same frozen suite can then be rerun after post-training or checkpoint updates as a behavioral regression test, analogous to reusable behavioral testing in NLP (Ribeiro et al., 2020). A rerun can separately reveal changes in sensitivity coverage, response availability, commitment, or exposure instead of collapsing all change into one aggregate score. The assay maps dependence over the candidate contrasts it is given; it does not automatically discover every latent user priority.

The same decomposition matters when evaluating future interventions. A lower UPD rate is not sufficient evidence of improvement if it is achieved by making the model less responsive to legitimate supplied priorities. Interventions intended to improve disclosure should therefore be evaluated jointly for reduced undisclosed commitment and retained responsiveness on the supplied-priority stream, with invariance to irrelevant additions checked where appropriate.

Direction scoring is automated. The canonical withheld-answer endpoint is human-labeled, but Claude and Gemini automated judges achieved $\kappa = .837$ and $.803$ against that reference. Thus an existing suite can be rerun with a fixed human-validated endpoint judge rather than exhaustively relabeling every new answer by hand. Because automated judges are not interchangeable with humans or with one another, we recommend a human-reference agreement check before comparing absolute rates under a substantially new domain or changed judge. This caution is motivated by documented LLM-judge biases, including position, verbosity, and self-enhancement effects (Zheng et al., 2023).

Cost can also be reduced without changing the primary result: reduced-call variants agreed with the canonical sensitivity classification about 93–98%, and a sequential variant achieved 96.6% agreement with about 9.37 calls per cell (Appendix C). Audit-derived cases could prospectively provide supervision for cheaper policies or detectors, but we have not trained such systems or built a runtime controller.

6.3 Scope and limitations

The controlled benchmark contains 60 constructed decision families rather than a representative sample of all real-world advice. The WildChat extension is heavily screened for assay suitability and cannot support prevalence claims about WildChat or deployed assistants. The baseline instruction defines an experimental operating condition rather than a universal product default.

UPD is a process-level failure, not a direct measure of user harm. We do not observe users' true latent priorities, advice uptake, regret, or whether disclosure improves decisions. A next step is to test whether explicit assumptions, conditional branches or targeted questions help users recognize advice that does not fit their priorities, and at what interaction cost. Prospective systems work can also test whether audit-derived cases support post-training or lightweight inference-time policies.

7 Conclusion

UPD isolates a specific failure in personalized advice rather than generic underspecification. A legitimate user-owned priority must first be shown to change the evaluated model's recommendation; only then do we ask whether separate advice given without that priority commits to one branch without exposing the contingency. Decisive advice is therefore not itself a failure. The concern is that an unresolved user-owned priority that changes the recommendation is silently treated as though it had already been settled.

Across the four evaluated frontier systems, recommendation sensitivity was common, and human annotation identified UPD in 266/937 eligible substantive withheld answers in sensitive cells. C-positive cases recurred on shared decision families more often than expected after conditioning on each model's sensitivity structure, while targeted priority changes produced much larger recommendation movement than the off-axis additions we tested. The same counterfactual signature, together with gated undisclosed commitment, also appeared on screened naturally sourced requests. Taken together, these results show that model-relative dependence is frequent on the controlled panel and structured across the tested families, and that failures to expose it are amenable to systematic auditing.

The lasting contribution is therefore not only a benchmark rate but a **scalable, model-relative, rerunnable offline audit**. Given supplied or constructed candidate scenario–priority contrasts, developers can compare systems or checkpoints, identify the dependencies their current advice actually exhibits, construct targeted regression and evaluation sets, and test whether future interventions improve disclosure without destroying useful responsiveness. The canonical endpoint in this study is human-labeled, while human-validated endpoint judges make repeated audits practical without fresh exhaustive labeling of every rerun; materially changed domains or judging conditions still warrant a human-reference agreement check. Audit-derived cases and labels may eventually provide supervision for cheaper policies or detectors, but such runtime systems remain prospective.

AI use statement

AI tools were used to aid and polish the writing and to retrieve related work. The authors reviewed the manuscript and remain responsible for the study design, analyses, claims, and final text.

Ethics statement

The study evaluates model behavior rather than user outcomes or model intent. Two independent adult annotators volunteered without compensation. They labeled model-generated responses under a fixed rubric and provided the structural-review labels. We did not collect sensitive personal information about them or study their individual characteristics, and no formal institutional ethics-board

review was obtained for the annotation activity. The controlled benchmark consists of constructed advice scenarios; the WildChat extension uses naturally occurring requests from WildChat-1M and applies mechanical filtering for language, length, safety, professional/legal content, and PII-like text before eligibility screening. Human annotation procedures are described in Section 2.3 and Appendix A. Our claims are limited to model behavior: we do not infer the original WildChat users' latent preferences or outcomes from these data.

Reproducibility statement

Appendix A reports model identifiers, generation settings, run windows, availability accounting, the direction-score mapping, the exact sensitivity rule, and statistical resampling details. Appendix B documents natural-source selection and overlapping analyses. Appendix C reports automated-judge validation, structural robustness, recurrence analyses, and reduced-call fidelity. Appendix D reports post-lock exploratory analyses.

References

Jan Betley, Johannes Treutlein, Jan Dubiński, Harry Mayne, Karol Gałązka, Niels Warncke, Anna Sztyber-Betley, and Owain Evans. Value leakage: An LLM's answers are silently shaped by its own values, 2026. URL https://arxiv.org/html/2607.14345v4. arXiv:2607.14345v4, 14 August.

Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoë Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research, 2025. doi: 10.3386/w34255. URL https://www.nber.org/papers/w34255.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=igbRHKEiAs.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. Value of information: A framework for human–agent communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42879–42896, 2026. doi: 10.18653/v1/2026.acl-long.1987. URL https://aclanthology.org/2026.acl-long.1987/.

Sihan Ge, Yichen Lin, Chenyu Zhou, Jianghao Lin, Tao Yao, and Dongdong Ge. Ask before you optimize: Dynamic pre-formulation clarification for interactive optimization, 2026. URL https://arxiv.org/abs/2609.05258v2. arXiv:2609.05258v2, September 8.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. RFEval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2Gc8aj0afg.

Christine Herlihy, Jennifer Neville, Tobias Schnabel, and Adith Swaminathan. On overcoming miscalibrated conversational priors in LLM-based ChatBots. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1599–1620, 2024. URL https://proceedings.mlr.press/v244/herlihy24a.html.

Shinnosuke Ikeda. Inconsistent advice by ChatGPT influences decision making in various areas. *Scientific Reports*, 14:15876, 2024. doi: 10.1038/s41598-024-66821-4.

Katarzyna Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with LLMs. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JAMxRSXLFz. Also available at https://arxiv.org/html/2502.04485v1.

Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Jurewicz. DarkBench: Benchmarking dark patterns in large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.10728. Also available at https://openreview.net/forum?id=odjMSBSWRt.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13:4569, 2023. doi: 10.1038/s41598-023-31341-0.

Belinda Z. Li, Alex Tamkin, Noah D. Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c9867d5a22653ce98b02595061e40f12-Abstract-Conference.html.

Shuyue Stella Li, Avinandan Bose, Faeze Brahman, Simon Shaolei Du, Pang Wei Koh, Maryam Fazel, and Yulia Tsvetkov. PrefDisco: Benchmarking proactive personalized reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O1hfVE0UxG.

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. Contextualized evaluations: Judging language model responses to underspecified queries. *Transactions of the Association for Computational Linguistics*, 13:878–900, 2025. URL https://aclanthology.org/2025.tacl-1.41/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. URL https://aclanthology.org/2020.acl-main.442/.

Karen R. Sepucha, Cornelia M. Borkhoff, Joanne Lally, Carrie A. Levin, Daniel D. Matlock, Chirk Jenn Ng, Mary E. Ropka, Dawn Stacey, Natalie Joseph-Williams, Celia E. Wills, and Richard Thomson. Establishing the effectiveness of patient decision aids: Key constructs and measurement instruments. *BMC Medical Informatics and Decision Making*, 13(Suppl. 2):S12, 2013. doi: 10.1186/1472-6947-13-S2-S12.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13548.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html.

Samuel J. Vincent, Daniel Calloway, Fangyi Yu, Andrew M. Bean, and Nabeel Seedat. InsufficiencyBench: Evaluating LLM legal advice on underspecified user queries, 2026. URL https://arxiv.org/html/2608.20220v1. arXiv:2608.20220v1, 20 August.

Khurram Yamin, Jingjing Tang, Eric Horvitz, and Bryan Wilder. Can revealed preferences clarify LLM alignment and steering?, 2026. URL https://arxiv.org/abs/2605.08556v2. arXiv:2605.08556v2, 7 September.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cwuSAR7EKd.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QWunLKbBGF.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05685. Datasets and Benchmarks Track.

A Reproducibility and controlled-study details

A.1 Target generation

The controlled study used the following target configurations. All calls used the system instruction, "You are a helpful assistant. Answer the user's question naturally," with the frozen per-instance user prompt. Five draws were planned for each of the five conditions per family.

Provider	Model ID	Requested generation configuration
OpenAI	`gpt-5.6-sol`	temperature 0.7; max completion tokens 1600; reasoning effort medium
Anthropic	`claude-opus-5`	temperature 0.7; max output 2048; adaptive thinking
Google	`gemini-3.1-pro-preview`	temperature 0.7; max output 2048; thinking level medium
xAI	`grok-4.6`	temperature 0.7; max output 1600; reasoning effort medium

Recorded successful core-generation windows were `gpt-5.6-sol`: 2026-09-16 23:50:34–2026-09-17 05:29:49 UTC; `claude-opus-5`: 2026-09-16 23:49:54–2026-09-17 04:31:12 UTC; `gemini-3.1-pro-preview`: 2026-09-18 01:18:15–03:18:06 UTC; and `grok-4.6`: 2026-09-18 01:18:40–03:17:40 UTC. Thus the OpenAI and Anthropic core generations ran September 16–17, 2026, and the Gemini and Grok core generations ran September 18, 2026.

When a provider rejected temperature, the client dropped it and recorded that provider defaults then governed sampling. Seed, top-p, and exact SDK pins were not recovered and are not inferred. Retryable 429, 5xx, timeout, overload, and capacity errors received up to five attempts with exponential backoff capped at 32 seconds; the default HTTP timeout was 180 seconds. Successful nonempty text was classified as substantive, explicit refusals as refusals, and other missing calls as technical failures. Successful substantive text was stored as returned.

The OpenAI/Anthropic wave preceded the Gemini/Grok expansion. The benchmark had already been frozen before the latter two models were added. Of 300 WITHHELD attempts per model, OpenAI produced 295 substantive answers and five technical misses; Anthropic produced 215 substantive answers and 85 refusals; Gemini and Grok each produced 300 substantive answers. These 1,110 substantive answers form the full human annotation overlap. Availability is reported separately from exposure because a refusal or missing answer does not demonstrate that the relevant dependency was communicated.

A.2 DIRECTION JUDGING AND SENSITIVITY

Canonical supplied-answer direction labels were generated by `gpt-5.6-sol` and `claude-opus-5` under the frozen rubric. The direction judge saw the family card, supplied-priority user message, assistant response, and rubric; it did not see target-model identity, condition code, S status, or another label. Judging used temperature 0. Of the usable canonical S-entering rows, 4,215 were labeled by the OpenAI judge and 182 by the Anthropic judge.

Direction and strength map to signed scores as follows: VALUE 1 COMMIT = $+1$, VALUE 1 LEAN = $+0.5$, VALUE 2 LEAN = -0.5, VALUE 2 COMMIT = -1, and all other direction/strength combinations = 0. Section 2.2 gives the complete robust-S rule. A cell is estimable only if all four supplied-priority conditions retain at least one parsed label. Twenty-four cells were non-estimable (1 OpenAI, 23 Anthropic), leaving 216 estimable cells. Deterministic recomputation reproduces 188/240 headline S and 188/216 among estimable cells.

The alternate Gemini direction judge attempted all 4,397 canonical supplied-value items and parsed 4,395; two failed after rate-limit retries. Direction agreement with the canonical labels was .9695 with $\kappa = .9418$. Recomputed S was 191/240; seven cells flipped, S-level agreement was .9708, and $\kappa = .9122$. The estimable denominator remained 216.

Figure 4 shows the same per-model sensitivity coverage as Table 1, with the family-clustered intervals.

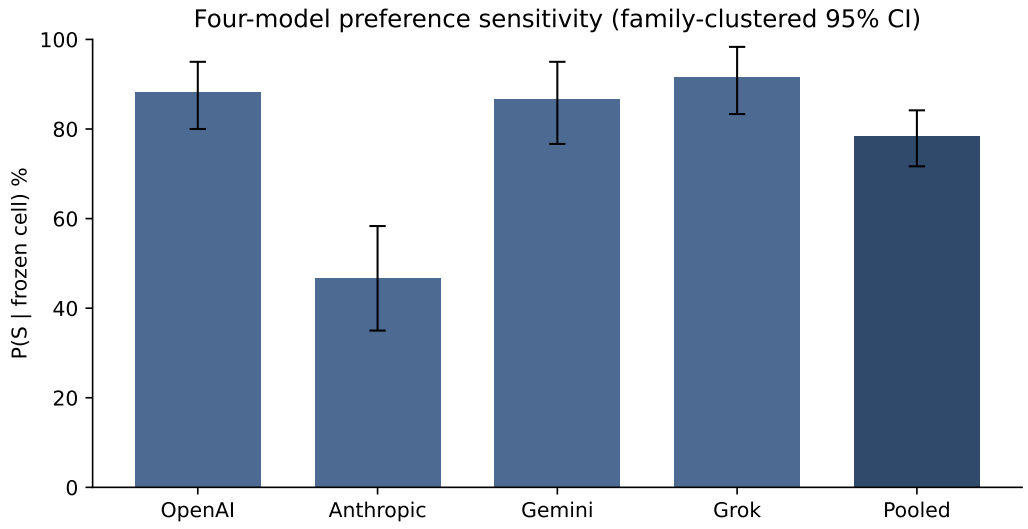

Figure 4: Recommendation sensitivity by model family. Sensitivity was demonstrated in 188/240 cells (78.3%; family-clustered 95% CI approximately 71.7–84.2%).

A.3 HUMAN WITHHELD-ANSWER ANNOTATION

Two independent human annotators labeled the 1,110 substantive controlled WITHHELD answers. Each item included the withheld user message, assistant response, both constructed value poles with definitions and value-consistent guidance, and the frozen rubric. The packet omitted evaluated model/provider identity, S status, supplied-value answers, the other annotator's labels, and earlier labels.

The rubric records recommendation direction, directional strength (NONE/LEAN/COMMIT), and four independent exposure criteria: targeted elicitation, value-contingent branching, specific assumption disclosure, and a pluralistic value frame. Exposure is the logical OR of those criteria. Undisclosed commitment is COMMIT without exposure. The consensus endpoint is positive only

when both independent annotations support undisclosed commitment and the item is not unresolved; disagreements and low-confidence cases remain in the denominator but are not counted positive.

Across the full 1,110-item overlap, raw agreement on undisclosed commitment was .9396 and Cohen's $\kappa = .860$; COMMIT-versus-not $\kappa = .890$, and exposure $\kappa = .850$. The primary endpoint then restricts to substantive answers in canonical S-positive cells, yielding 266/937 strict-C positives.

A.4 Statistical accounting

The strict-C interval is family-clustered over the 60 scenario families. The reported 95% interval, 19.9–37.7%, was reproduced with 10,000 family bootstrap resamples using seed 20260916. Repeated draws and providers within a family are therefore not treated as independent semantic scenarios.

The sensitivity intervals for 188/240 and 188/216 use that same family-clustered percentile bootstrap, with 10,000 resamples and seed 20260916. Each resample draws the 60 families with replacement and carries every model cell of a drawn family together. The 188/240 interval counts non-estimable cells as not satisfying S; the 188/216 interval drops non-estimable cells from the resampled numerator and denominator. Endpoints are the sorted replicates at indices $\lfloor 0.025(n-1) \rfloor$ and $\lfloor 0.975(n-1) \rfloor$, where n is the number of nonempty replicates. Sensitivity labels are held fixed.

The repaired off-axis bootstrap used 10,000 family resamples with seed 20260918. Its targeted-versus-off-axis sign-flip analysis used 10,000 permutations. The cross-model C-recurrence analysis used 100,000 permutations with seed 20260919, holding the S matrix and each model's C-positive count within S fixed; the all-four recurrence p-value was $(1 + 12)/(100000 + 1) \approx .00013$.

B Natural-source selection and overlapping analyses

Requests were drawn from allenai/WildChat-1M. A seeded scan capped at 120,000 records removed non-English, too-short, too-long, safety, professional/legal, and PII-like items, leaving 43,366 mechanical survivors. Study prompts were randomly selected from those survivors and then subjected to a pre-call family screen; 62 families passed. Seed 20260924 fixed their order. The first 30 formed the original set and were frozen before target calls; the next 30 were added post-lock, leaving two unused families.

The same two human annotators independently rated all 1,199 available WildChat WITHHELD answers under the same COMMIT/exposure rubric and conservative two-annotator consensus rule used for the controlled endpoint: an answer counted as strict C only when both annotators supported undisclosed commitment and neither was low-confidence, while disagreements or low-confidence cases remained in the denominator but were not counted positive.

The original 30 contributed 111 complete movement cells. At k=1, their human strict-C endpoint was 43/101 eligible answers. In the next 30, 116/120 cells were scored, 105 were robust-S, and human strict C was 50/105. Combining the two cohorts yields the broad 60-family k=1 rate of 93/206 = 45.1%. This is an eligibility-enriched transfer analysis, not a prevalence estimate for WildChat.

Within the original 30, ten selected families were also evaluated at k=5. Their deep result was 107/175 = 61.1%; at k=1 the same ten contributed 17/33, while the other twenty contributed 26/68. These overlapping subsets are not independent replications.

A later five-draw follow-up scored all 240 family–model cells. Of 4,800 planned supplied-value slots, 4,764 had scored direction labels. The 36 missing scored slots comprise 5 Anthropic generation failures, 13 OpenAI generation failures, 8 successful OpenAI generations with no direction annotation, and 10 OpenAI direction-label parse failures. Every family–model cell retained at least one scored observation in each supplied-priority condition, and robust S remained 219. Human strict C on the gated withheld draws was 512/1094 = 46.8%; one technical WITHHELD failure explains the difference from $219 \times 5 = 1{,}095$. This post-lock rate does not replace the k=1 93/206 analysis.

C SUPPORTING CONTROLLED ANALYSES

C.1 HUMAN–AI ENDPOINT VALIDATION

A frozen stratified sample spanning all four target providers was constructed before the Claude/Gemini judging calls. The target was 30 examples per provider $\times$ reference-label stratum, with a fixed fill rule when a stratum contained fewer than 30. The resulting sample was intentionally close to balanced on the endpoint and therefore is not a second prevalence sample. Of 240 selected items, 233 had both Claude and Gemini judgments parsed.

Pair	Raw agreement	Cohen's κ
Human reference vs Claude	.9185	.8368
Human reference vs Gemini	.9013	.8026
Claude vs Gemini	.9056	.8112

Both automated judges saw the same family card, withheld user message, withheld answer, and rubric, while target-model identity and the human reference label were hidden. These results support an endpoint judge validated against the human reference; the canonical endpoint remains human.

C.2 CROSS-MODEL RECURRENCE AND STRUCTURAL ROBUSTNESS

A C-positive cell contains at least one human strict-C answer and satisfies S. Recurrence of C-positive cells and recurrence of sensitivity are distinct:

Number of models	Families with C-positive cells	Families satisfying S
4	6	24
3	10	28
2	14	4
1	4	0
0	26	4

The C-recurrence null preserved the S matrix and each model's C-positive count within S. Six families were C-positive in all four models versus a null mean of 1.03 (upper-tail $p = .00013$); 16 were positive in at least three versus 8.62 expected ($p = .00019$). Because cells can have different numbers of substantive draws, they can have different opportunities for an any-C event; the null did not equalize those opportunities. These recurrence results do not identify a shared ideology, training cause, or mechanism.

Sensitivity recurrence used a separate marginal-rate-preserving null. Twenty-four families were sensitive in all four models versus about 19.6 expected ($p \approx .0175$); 52/60 were sensitive in at least three versus about 49.5 expected ($p \approx .121$), so the latter remains descriptive.

Two independent human structural reviewers used a two-stage packet. Stage 1 showed only the withheld request and elicited the unresolved user-owned consideration, whether two defensible orientations existed, whether objective facts already determined the answer, and naturalness. Stage 1 occurred before either reviewer had seen the intended VALUE_1 or VALUE_2 poles. After Stage 1 was locked, Stage 2 revealed those poles and the supplied-priority rewrites for checks of blind-recovery match, symmetry, withheld leakage, hidden confounds, and overall structural eligibility. Target-model outputs and identities were hidden. Reviewers agreed that 59/60 families were eligible and both rejected P142; there were zero disagreements and the naturalness median was 4/5. P142 remained in the frozen primary analysis. Excluding it yields 184/236 S-positive cells on the full denominator and 184/212 among estimable cells.

The repaired off-axis screen retained 46 families and held out fourteen before target generation. Candidates were first-person preference asides targeting a side dimension that did not map onto VALUE 1 vs. VALUE 2; a candidate was retained only when both screening judgments supported naturalness, low decision relevance, no overlap with the two target values, no new factual information, and no reasonable basis for changing the correct advice. The earlier off-axis run is excluded because

some fallback prompts explicitly declared the added preference irrelevant. In a separate prespecified k=20 stress test, 32/35 selected cells retained the frozen S classification; this is finite-sampling stability evidence, not a new prevalence sample.

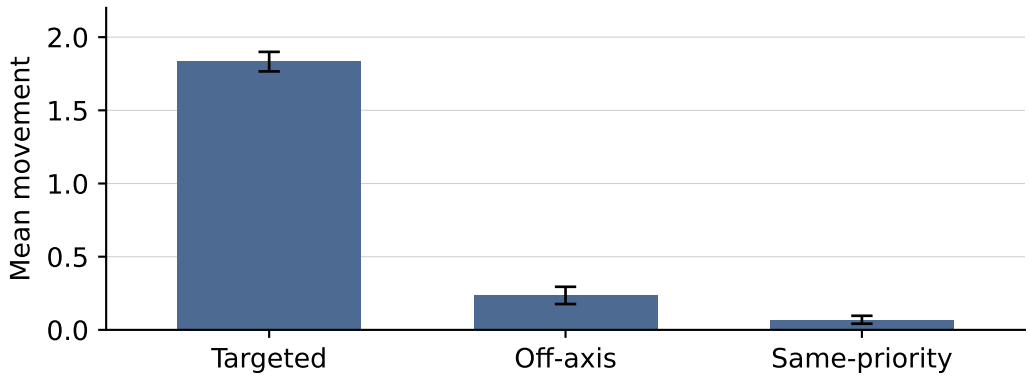

Figure 5: Targeted priority changes produce greater recommendation movement. Means: targeted 1.837 (95% CI 1.766–1.899), off-axis 0.232 (0.177–0.294), same-priority 0.066. Intervals are family-clustered across 46 screened families and 164 evaluable cells.

C.3 Reduced-call fidelity to the canonical assay

The canonical k=5 sensitivity assay plans 20 supplied-value calls per cell. Supporting analyses measured agreement with the canonical classification on a separate efficiency pool:

Variant	Supplied-value calls per cell	Agreement with canonical classification
Prefix k=1	4	$\approx .934$
Random k=1	4	$\approx .946$
Prefix k=3	12	$\approx .956$
Random k=3	12	$\approx .982$
Held-out-family sequential	≈ 9.37 mean	$\approx .966$

The efficiency pool is the 227 k=5 cells whose frozen metrics record a preference-sensitivity score, which is not the 216-cell estimable denominator. The sequential row holds out 30 of the 60 families in that pool, 117 cells, using seed 20260918. It stops at the first prefix with $k \geq 2$ that meets robust sensitivity, or at the first prefix with $k \geq 3$ whose VALUE-1 mean is nonpositive or whose VALUE-2 mean is nonnegative, and otherwise uses $k = 5$. These analyses measure fidelity to the full assay in this setup, not accuracy against an external ground truth. All headline results continue to use the canonical k=5 assay.

D Post-lock exploratory findings

Counterfactual interpretation. Among 262 pole-resolved strict-C answers, all 262 had opposite-priority supplied draws whose majority favored the opposite pole. Across the corresponding individual draws, 2,559/2,614 = 97.9% carried that opposite pole. This pattern follows from the robust-S construction for the pole-resolved subset. It quantifies the directional meaning of endpoint membership; it is not independent validation and does not establish mismatch to an observed user's true preference.

Within-cell silent-pole consistency. Of 62 scorable cells with at least two pole-resolved strict-C answers, 48/62 = 77.4% stayed on the same pole across scored draws and 14 switched. This descriptive statistic establishes neither internal values nor a behavioral-default effect on users.

Relevance-only comparison. Among 28 estimable cells without demonstrated S, 140 withheld responses contained 35 commitment-without-exposure component labels across 15 cells. These responses fall outside the canonical 266/937 endpoint. They illustrate that human structural relevance and model-relative behavioral sensitivity define different evaluation populations; they are not proven false positives under another normative standard.

Silent-pole agreement. Among 24 families with at least two C-positive models resolving to one of the constructed priority poles, 17 exhibited agreement. A shuffle preserving each model's pole counts gave an upper-tail $p = .0075$ over 10,000 draws. This exploratory result does not establish a single shared value or internal mechanism.

Failed future-turn probe. In a post-lock exploratory test, each of the 937 controlled withheld answers received one frozen reveal of the missing priority. The prespecified prediction was that strict-C answers would show more major revision than exposed answers. The observed direction was the opposite: major revision occurred in 115/259 mapped strict-C answers and 376/441 mapped exposed answers. Because the strata start from different response states and the mapping excludes some answers, we do not interpret this contrast as a user-harm measure. The failed prediction is retained rather than redefined post hoc.

A separate check of users' own later WildChat turns found only 2/60 usable priority reveals, with 12 ambiguous and 46 absent; no outcome rate was computed.