

LLMs as Oracles: Reliance on LLMs for Subjective Personal Questions

Myra Cheng^{✉*}
myra@cs.stanford.edu
Stanford University
Stanford, California, USA

Lujain Ibrahim^{✉*}
lujain.ibrahim@oii.ox.ac.uk
University of Oxford
Oxford, England

Grace Liu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Michelle S. Lam
Stanford University
Stanford, California, USA

Vishakh Padmakumar
Stanford University
Stanford, California, USA

Nick Madibekov
Stanford University
Stanford, California, USA

Diyi Yang
Stanford University
Stanford, California, USA

Dan Jurafsky
Stanford University
Stanford, California, USA

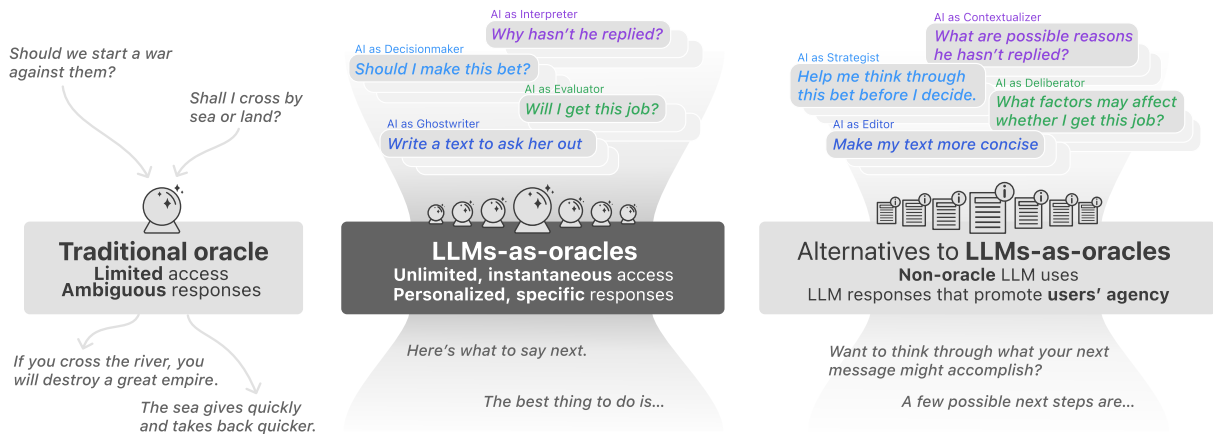


Figure 1: While people have long consulted oracles for subjective personal questions, these oracles were limited in access and specificity. Now, people turn to LLMs for similar questions without these constraints, raising the risk of excessive reliance. We build a typology of the ways people use LLMs as oracles, from offloading beliefs and normative judgments to offloading decisions and actions. Using this typology, we find that LLM-as-oracle use is increasing across multiple real-world usage datasets, and that people are often unaware of their own LLM-as-oracle use. We also show how model behavior and people’s perceptions of AI drive this reliance. Finally, we propose interventions on both sides: how people can use LLMs in alternative, non-oracle ways, and how model outputs can better facilitate users’ autonomy.

Abstract

We characterize how people are turning to LLMs as *oracles*: all-knowing authorities on subjective personal questions. Motivated by risks to users’ autonomy and well-being, we develop a typology and LLM-based methods to measure this form of AI reliance at scale and understand how people are offloading judgment and decision-making to AI. Applying our typology to public usage data (68K prompts from WildChat and ThoughtTrace), we find that LLM-as-oracle use has increased over time (2023–2026) and is more prevalent among younger users. We further build a privacy-preserving data donation tool to analyze individuals’ longitudinal usage data (140K prompts from 52 participants), identifying similar trends. People are often unaware of their own LLM-as-oracle use, and express

dissatisfaction with this behavior after seeing our tool’s analysis. Finally, we identify two drivers of LLM-as-oracle use: people’s perceptions of AI and the behavior of AI models themselves, which motivate possible interventions to support users’ self-deliberation.

CCS Concepts

• **Human-centered computing** → Empirical studies in HCI; Natural language interfaces; • **Computing methodologies** → Natural language generation.

Keywords

artificial intelligence, large language models, AI reliance, personal advice, autonomy, offloading, perceptions of AI

*Both authors contributed equally to this research.

1 Introduction

“Should I take a career risk by quitting my job?” “Is this relationship right for me?” People have long used oracles and other external decision aids to manage uncertainty and risk. In ancient Greece, supplicants traveled great distances to consult the Oracle of Delphi for advice [47, 80, 114], and in medieval Europe, astrologers read the stars to advise leaders on state matters [23]. These practices served an important function by providing a source of judgment beyond one’s own, helping people navigate uncertainty, share the burden of deciding, and ultimately act.

Today, people describe AI in similar terms—opaque and mysterious, yet all-knowing [30]—and similarly turn to AI for subjective personal questions, from deciding a next career move to navigating a romantic interest [128, 154] to making critical health decisions [2, 99]. But unlike traditional oracles that were hard to access and produced ambiguous outputs, LLMs short-circuit the process of interpretation or deliberation: they are constantly accessible, provide hyper-specific and personalized answers, and require as little effort as typing a single sentence.

Thus, using LLMs for these types of questions makes it easy to **outsource the work of judgment**, raising the risk of overreliance. When answers are always available and readily trusted, people may hand over the work of forming and revising their own beliefs to LLMs [61, 64, 105]; when LLM answers are confident and hyper-personalized, people may become less tolerant of uncertainty [73]; and when LLMs are easier to reach than a friend, people may lose social ties as they turn to AI over the humans in their lives [77]. With conversational AI systems now reaching more than a billion monthly users [74, 91], these individual effects may accumulate into widespread disempowerment [89, 143]. Compounding these risks, users themselves may not realize their increasing use of LLMs in this way. If the behavior is unintentional and goes unnoticed, people cannot guard against its harms. Motivated by these concerns, we characterize this type of reliance as **LLM-as-oracle**—people turning to LLMs as authorities on subjective personal questions—and develop methods to measure and understand it.

LLM-as-oracle use is difficult to measure for several reasons. First, in-the-wild AI usage data is scarce, especially for subjective personal questions, so the behavior itself cannot be readily observed. Second, existing frameworks for studying AI reliance focus on verifiable settings such as coding or factual information-seeking, where overreliance can be evaluated by whether people follow “incorrect” AI outputs on tasks with observable outcomes [14, 76, 93, 124, 173]. In subjective domains without such concrete outcomes, the line between appropriate use and overreliance is less clear-cut [76]. Third, the consequences of using LLMs as oracles may not be visible within a single interaction. Thus, we currently do not know whether, or how often, people rely on LLMs as oracles.

Here, we take the approach of understanding this reliance via observing people’s *processes*: how and to what extent they involve AI in forming their own judgments and decisions. We ask three questions: (1) How often do people use LLMs as oracles, and is this use increasing? (2) Do people use LLMs as oracles intentionally, or without realizing it? (3) What factors, such as user attitudes or LLM behaviors, drive this reliance?

To answer these questions, we first present a typology of what we call LLM-as-oracle roles—the different ways that users’ questions offload judgments to AI—like whether prompts treat AI as a decision-maker or provider of normative judgment, paired with non-oracle alternatives for each role, like consulting AI as a decision strategist or for alternative perspectives. This typology enables automated LLM-based methods to measure this phenomenon at scale (Section 3). We also build and deploy a privacy-preserving data donation tool to collect and analyze individuals’ longitudinal data using our typology. Our data donation tool presents these aggregated analyses as insights to the donor to facilitate self-reflection.¹ Our methods reveal three insights:

- (1) **People are increasingly turning to LLMs as oracles (Section 4).** We apply our LLM-based methods to analyze public real-world usage data (analyzing $N = 68K$ prompts) and data donated using our privacy-preserving data donation tool (analyzing $N = 140K$ prompts from 52 participants), finding significant temporal increases and higher prevalence of oracle use among younger people.
- (2) **People do not realize the extent of their LLM-as-oracle use (Section 5).** Through our data donation study, we find that people underestimate their LLM-as-oracle use. After being presented with a personalized report on their AI use, participants often expressed surprise and dissatisfaction, setting intentions to use LLMs less as oracles in the future.
- (3) **Both people’s perceptions and model behavior drive LLM-as-oracle use (Section 6).** Through interaction logs from two user studies, we study the role of users’ perceptions and AI behavior. We first conducted a preregistered experiment ($N = 520$ participants) to investigate the causal role of users’ perceptions of AI versus humans in LLM-as-oracle use. We find that warmer and more positive perceptions of AI are associated with LLM-as-oracle use. We further link LLM-as-oracle use to AI model behavior by applying our methods to data from a longitudinal study of human-sycophantic AI interactions (analyzing $N = 95K$ prompts from 1015 participants). We find that LLMs biased to take a particular perspective lead to more LLM-as-oracle use. This motivates developing interventions on oracular LLM behavior as a possible mitigation.

Our work broadens the study of reliance beyond verifiable domains to subjective, open-ended settings. Here, excessive offloading of sense-making and decision-making to AI risks damaging user autonomy, competence, and well-being, and at a societal scale, homogenizing users’ judgments and concentrating AI providers’ influence. We characterize a feedback loop between user perceptions, interaction style, and model behavior, motivating the need for model training and evaluation to integrate considerations of these dynamics. Our methods² enable further study of broader forms of AI reliance, and their impacts.

2 Background and Related Work

Our “LLM-as-oracle” metaphor indexes both on the broader appeal of oracles (Section 2.1) and how using LLMs in particular as oracles

¹The data donation tool can be accessed at <https://oracles-data-donation.vercel.app/>

²Our code and data is available at <https://osf.io/8k6xg>

poses risks to users’ autonomy, competence, and social relationships (Section 2.2). We build on prior work studying AI reliance, especially for subjective personal questions (Section 2.3), and how perceptions and metaphors shape human-AI interaction (Section 2.4).

2.1 What makes oracle-use productive

People have long used external decision aids to manage uncertainty, from the Oracle of Delphi in ancient Greece [47, 114] to medieval astrologers [23] to coin flips, tarot, and magic 8-balls today [81, 135]. The benefits of this consultation lie not in the oracle’s response, but in the process of formulating the question and in analyzing one’s own reaction to the oracle’s ambiguous answer [15, 56, 95]. Studies show, for instance, that people’s affective reactions to a coin flip’s outcome reflect their prior preferences, suggesting that the coin helps surface to them what they already wanted [82]. This reaction is a form of sense-making, which has well-documented benefits: making sense of why something negative happened helps people better recover from it [46, 161], and writing about difficult experiences improves well-being by helping the author form a coherent explanation [130, 131]. Oracles are thus most useful not as *deciders*, where people simply follow the response, but as *catalysts*, where the response’s real function is to surface the user’s own reaction to it [81]. People’s use of AI in tarot already reflects this bifurcation [135], and in another study, users of an AI system that made inferences about them enjoyed an unintended benefit of self-reflection, regardless of those inferences’ accuracy [142].

2.2 LLMs’ differences from traditional oracles pose new risks

However, LLMs differ from traditional oracles in several key ways that may displace, rather than support, users’ sense-making. First, interacting with LLMs is far more frictionless: they are always available to provide immediate answers and instant gratification [132]. Second, where an oracle’s ambiguity left interpretation to the consultor, LLMs’ hyper-specific and confident answers leave far less room for ambiguity [95]. LLM outputs also often are overconfident, leading to overreliance [173], and they rarely ask clarifying questions [141]. Thus, when a user poses an underspecified question, LLMs tend to fill in what is missing and commit to an answer rather than press for clarification, such that the user receives a fully specified and overconfident answer [25, 100]. Third, LLMs are not indifferent to the consultor in the way traditional oracles are, and instead are designed to provide personalized assistant-like responses to user queries. They have also been shown to be sycophantic, excessively mirroring and agreeing with users’ preferences [31, 78]. Thus, where a traditional oracle’s response is independent of the consultor’s leanings, an LLM’s response may actively reinforce them. These differences create risks to autonomy, competence, and social relationships that we do not measure directly, but that motivate our characterization of the LLM-as-oracle phenomenon.

Threats to autonomy and competence. Many of the questions people bring to LLMs, such as whether to leave a relationship or whether a career change is right [144, 154], are normative questions with no universally agreed-upon ground truth [118], where the answer depends on evolving personal context, fundamentally incomplete information, and the user’s own values. Thus, these

questions are often not well-served by a single resolution, instead requiring ongoing belief revision as new evidence and experiences arrive. When an LLM provides a confident, specific answer, it successfully reduces uncertainty in the moment, but also risks short-circuiting this process. As people offload the work of forming and revising their beliefs, AI outputs can become persistently integrated into what the user thinks and does [61, 64]. Prior work shows that sycophantic AI can shift people’s judgments, making them more certain and more extreme in their attitudes, without them realizing it [31], and inflate their perceptions of their own abilities [138]. AI advice has also been shown to give people a false sense of certainty even when the advice is wrong [115], and to be preferred even when it is evaluated as lower quality by experts [3]. More broadly, LLM-as-oracle use may erode people’s capacity for deliberative judgment because it removes the critical practice of sitting with ambiguous, value-laden situations and arriving at one’s *own* judgment [64, 87, 105]. This pattern may be difficult to self-correct, because users, in the moment, may actually experience it as helpful. For example, a recent large-scale analysis of human-LLM interactions shows that “disempowering” interactions, such as verbatim scripting of value-laden personal communications, can receive higher user approval ratings [143]. Beyond these individual effects, oracle use raises concerns at a societal scale. If many people defer to the same few AI systems on similar personal questions, the providers of those systems gain disproportionate influence over people’s values and judgments, and the resulting homogenization of decisions and actions carries harms of its own [97].

Threats to social relationships and well-being. Prior work shows that people readily follow AI advice [112], which may include advice that directly harms their well-being or relationships. For example, discussing an interpersonal conflict with an LLM that provides sycophantic advice has been shown to make people less willing to apologize or repair the relationship [31]. Additionally, because LLMs are always available and may resolve people’s uncertainty in ways that feel helpful in the moment, people may become more dependent on LLMs for advice and less likely to reach out to friends and family, who may provide less confident or certain advice, for such consultation [51, 77]. Prior work has shown that people’s emotional dependence on AI can also result in poor psychosocial outcomes [51, 133]. Finally, LLMs’ willingness to engage indefinitely on questions that are uncertain or unresolvable, such as whether a partner’s intentions are good or whether a past decision was the right one, may facilitate *rumination* [59]: an LLM may continue to analyze the situation with the user from every angle [33], encouraging them to replay difficult situations rather than reaching acceptance or taking action [123, 140].

2.3 AI reliance in sense- and decision-making

A large body of work studies how people use AI in decision-making [19, 21, 60, 109, 136, 152, 157, 163], when they choose to trust and rely on it [20, 93, 164, 173], and what factors contribute to appropriate reliance [17, 66, 155]. Most of this work concerns tasks with a verifiable answer, against which reliance can be judged, with fewer works examining settings where no ground truth is available [76, 110], or where the concern is AI’s undue influence on

users’ beliefs rather than accuracy [84, 160]. An adjacent direction of research examines subjective and personal uses, such as people seeking personal advice from AI [118] and forming relationships with AI companions [167, 170]. Companions can improve mood and reduce loneliness [40, 61, 67], but also foster emotional dependence [39, 101] and attachment [6, 38, 96, 116, 133]. LLM-as-oracle use also connects to previous work on questions of perceived control, authority, and ownership [24, 44, 49, 86, 127]: if people increasingly defer to AI, they may feel a lessened sense of ownership over their judgments and actions, though these impacts remain unclear.

2.4 Perceptions of AI affect interaction

We view *oracle* as an implicit role in which users may treat the LLM, cutting across different conversation topics and interaction modes like treating AI as a companion, collaborator, or tool [94]. This builds on prior work studying social roles [13] in the context of emerging technologies [122], online communities [162], and more recently, conversational AI systems [27, 30, 32, 94, 117, 154]. Implicit roles and conceptual metaphors can strongly shape users’ expectations for interaction [92, 120], and vary along dimensions such as human-likeness [48]; classical social perception dimensions like warmth and competence [54]; and the extent to which they grant agency or autonomy to both the AI and the user [94, 145].

3 ORACLE Typology

Unlike other typologies of LLM use that focus on prevalence of different domains or topics, like work versus relationship advice [26, 144, 167], we aim to more broadly characterize *how* people use AI as oracles across many domains. We present a typology of different LLM-as-oracle (henceforth ORACLE³) roles and a validated LLM judge to label user messages with these roles.

3.1 Methodology

We developed the typology through several rounds of iterative discussions among five authors, drawing on both existing AI role typologies [117, 154] and a bottom-up approach of discussing and categorizing examples of user prompts from our datasets until we reached consensus. We began by categorizing each of the nine roles identified by Tseng and Liang [154]. Tseng and Liang [154] focus on the relationship advice domain, identifying nine roles that differ based on the user’s goal in an interaction (understanding a situation, preparing to act, or executing an action) and whether they place the AI in a passive or active role. Through iterative discussions, we simplified the groupings of user goals that they identify and generalized them to broader topics, ultimately identifying four task types where ORACLE is salient: (1) understanding oneself or others; (2) normative judgment; (3) choosing or planning to act; and (4) executing an action. These mapped broadly to targeting the user’s beliefs (1 and 2) versus targeting the user’s actions (3 and 4). Then, for each task type, we identified an ORACLE role and developed a corresponding role that does not treat the LLM as an oracle (which may not necessarily be prevalent in existing LLM use data).

We then randomly sampled batches of 100 queries from real-world use datasets (see Section 4.1) and discussed how they fit into

the roles we identified until saturation, i.e., no new roles emerged and no existing roles were redefined. We discussed edge cases, such as categorizing an advice query that seeks a possible approach, like asking for “a good place to start” as *Strategist* rather than *DECISION-MAKER* because it does not seek a single definitive answer. We take a **conservative approach** in assuming users’ implicit intent to avoid overreporting ORACLE rates. Thus, we exclude mere descriptions of context, self-disclosure (e.g., “I am feeling sad”), and “LLM as journal” use cases, focusing our analysis on explicit requests. Though a user may be implicitly seeking counsel when they self-disclose to LLMs, this cannot be determined from the prompt alone.

3.2 Typology (Table 1)

Our typology differentiates user queries along two dimensions:

- **Target: Does the query target the user’s *beliefs* or *actions*?** This dimension characterizes what aspect of the user’s inner or outer life the query is directed at. Belief-oriented tasks include understanding intentions and making normative judgments, while action-oriented tasks include creating plans or writing. This enables us to understand the specific contexts in which people are using LLMs as oracles across different domains.
- **Framing: Does the query treat the LLM as an *oracle*?** This dimension characterizes the extent to which the user offloads their judgments or decisions to the AI. ORACLE queries tend to ask the LLM to supply an interpretation, judgment, decision, or script. In contrast, non-oracle queries ask the LLM for information or technical assistance that the user may then use themselves to produce an interpretation, judgment, decision, or script.

The typology comprises the following roles: for (1) understanding oneself or others, we differentiate between the LLM as an *INTERPRETER* versus a *Contextualizer*; for (2) normative judgment, we identify the roles of *EVALUATOR* versus *Deliberator*; for (3) choosing or planning to act, we differentiate between *DECISION-MAKER* versus *Strategist*; and for (4) action execution, we differentiate between *GHOST-WRITER* and *Editor*. The alternative non-oracle roles—*Contextualizer*, *Deliberator*, *Strategist*, and *Editor*—demonstrate how LLMs can support user reasoning and decision-making without having the user entirely offload her judgments. Rather than aiming to eliminate ambiguity and uncertainty, LLMs can preserve them by surfacing multiple interpretations, laying out competing considerations, and presenting options for users to evaluate themselves.

A note on comprehensiveness. Our typology is not meant to exhaustively all possible user prompts or requests, but is rather meant to categorize and differentiate among ORACLE uses. It excludes factual tasks, creative writing, translation, and agentic tasks, and instead focuses on queries that involve personal or interpersonal subjects, broadly defined to encompass any query where personal preferences, values, norms, or particular context might be salient. Our taxonomy is created both top-down (from existing works) and bottom-up (annotating and discussing randomly sampled queries from three datasets), but there may be a long tail of rare use cases that we do not cover [11].

³For the rest of the paper, we set ORACLE roles in small caps.

	Task	ORACLE role	Examples	Non-oracle role	Examples
Beliefs	Understanding self or others	INTERPRETER: Interprets a situation or dynamic; another person’s behavior or words; or the user’s own emotions, motives, or patterns of behavior.	1. “Does this text mean that she is upset?” 2. “Why do I keep procrastinating?”	Contextualizer: Generates multiple possible interpretations or broadens the set of explanations the user is considering.	1. “What are a few ways this text could be interpreted?” 2. “List common reasons for procrastination”
	Normative judgment	EVALUATOR: Makes an assessment, e.g., judgment of whether something is appropriate, justification of an action, or future-looking prediction.	1. “Am I being unreasonable for setting this boundary?” 2. “Was my manager unfair to criticize me?”	Deliberator: Explains relevant considerations and different perspectives without deciding which judgment is correct.	1. “What perspectives might different people have on setting this boundary?” 2. “What factors might shape whether my manager’s criticism was fair?”
Actions	Choosing an action	DECISION-MAKER: Makes or approves a decision for the user, or selects a single course of action.	1. “What should my fitness goals be?” 2. “Make a plan for what I should do this evening.”	Strategist: Maps out possible approaches, generates options, or projects consequences without selecting for the user.	1. “What are some approaches for identifying fitness goals?” 2. “What are some options for how I could spend this evening?”
	Acting (writing)	GHOST-WRITER: Determines what to say and produces the text for the user.	1. “Write a text telling her I can’t make it.” 2. “Draft an email asking my professor for an extension because I am sick.”	Editor: Revises text supplied by the user that already contains the substance of what the user wants to communicate.	1. “Here is my draft, make it warmer but still firm” 2. “Can you make my email more concise”

Table 1: Our role typology captures different forms of reliance on AI for subjective personal questions. For each task, we present an ORACLE roles and an non-oracle alternative role.

3.3 LLM judge for ORACLE user prompts

We next develop an LLM judge to classify user queries into one of the eight roles described above or an NA category, where no role is applicable. The NA category is an umbrella category for any prompts that are *not* subjective personal questions, including factual questions, creative writing, context sharing, etc.; in practice, we expect the majority of prompts to be categorized as NA.

To perform this classification, we prompt an LLM with definitions and in-context examples for each role. The full LLM judge prompt is provided in Appendix A. To validate each LLM judge, two expert annotators independently classified queries for a stratified random sample of 219 examples from our datasets (30 examples for each role and for NA; some categories had fewer than 30 examples total). Each expert annotator had substantial agreement with the LLM ($\kappa > 0.65$). The annotators also had substantial agreement with each other ($\kappa = 0.67$), reflecting the inherently subjective nature of the task (e.g., deciding which role is dominant when a message might include multiple). We used Gemini-2.5-Pro as our main LLM judge, and we validated that other models perform comparably and thus can be used interchangeably, with $\kappa = 0.66 - 0.70$ (Table A1). The raw agreement rate was > 0.70 in this 9-way task (random chance would be 0.11); and Fleiss’s κ among all three annotators is 0.66. Since the annotated sample also included examples from the NA category, this helps validate that our typology is comprehensive and covers most of these patterns in the data. One annotator labeled an additional random sample of 100 queries that were labeled as NA by the LLM judge, and found that 98% were indeed not applicable, further validating the judge’s coverage.

In the following sections, we use this typology and validated LLM judges to understand (1) temporal and demographic trends in ORACLE use (Section 4), (2) whether this use is intentional (Section 5), and (3) relevant causal factors, and how these causal factors can motivate interventions (Section 6).

4 Evaluation 1: Tracking the prevalence of ORACLE use

We first use our typology to characterize ORACLE in existing real-world usage datasets (Section 4.1). Since these curated datasets may

not fully reflect real-world use by everyday users [70], we augment this analysis by developing a privacy-preserving data donation tool to analyze individuals’ personal usage logs, obtaining multiple years of conversation histories from 52 users (Section 4.2).

4.1 ORACLE use in public LLM usage data

First, we measure ORACLE use by applying our LLM judge on two real-world usage datasets: **WildChat** and **ThoughtTrace**. **WildChat** [41] is a dataset containing 3.2M conversations between human users and different ChatGPT models (excluding the 1.5M that were filtered as toxic) spanning February 2023 to May 2024. We filtered this dataset further by only using conversations tagged as English (1.8M) and then running a broad, permissive LLM-judge-based filter for whether the queries had any personal or subjective element to exclude purely factual queries, creative writing requests, etc. (the prompt for this LLM judge is in Appendix A). This resulted in 60,376 user prompts and LLM responses. While WildChat has been critiqued for not being representative of broader LLM use, as it over-represents power users [70] and lacks demographic data, it is one of the few public datasets that captures diverse, real user behavior over time [63]. Similar to [134], we removed an anomaly where a user sent > 800 nearly-identical queries on a single day.

ThoughtTrace [85] is a dataset of 8,526 prompts and responses in 2,155 real-world multi-turn conversations from 1,058 users across diverse topics, spanning culture and lifestyle; education and knowledge, business and society; health and relationships, etc. This dataset thus reflects use of LLMs for practical everyday concerns as well as professional, informational, and learning-oriented uses in subjective personal questions. While this dataset is limited in size and temporal scope relative to WildChat, it provides demographic information and more recent data (collected in 2026) on how people interact with state-of-the-art models.

Relative prevalence of different roles (Figures 2 and 3, left). In both WildChat and ThoughtTrace, the majority of prompts are not applicable—18% and 38% are classified as subjective personal questions under one of our eight roles respectively (Figure A2). Many prompts relate to creative writing [63], coding, factual queries, or are not questions [41], which may be why this phenomenon has

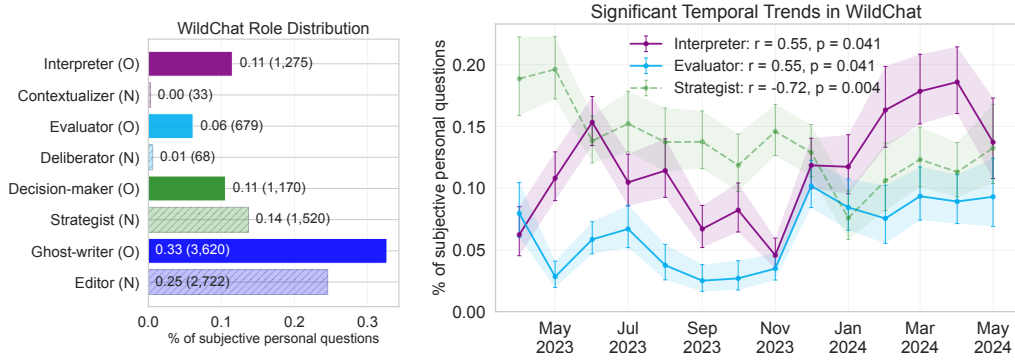


Figure 2: ORACLE use in WildChat. Left: Among subjective personal questions in WildChat, LLMs are primarily used for writing/editing, followed by interpretation and decision making. The belief-related tasks of interpretation and judgment are dominated by ORACLE (O), while the action-related tasks are more split between (O) and non-oracle use (N). Right: ORACLE roles **INTERPRETER** (violet) and **EVALUATOR** (cyan) are significantly increasing over time (both Pearson’s $r = 0.55, p = 0.041$), while uses of LLM as a non-oracle **Strategist** (green, dashed) are decreasing ($r = -0.72, p = 0.004$). Shading indicates 95% CI. Results including the NA label are in Figure A2.

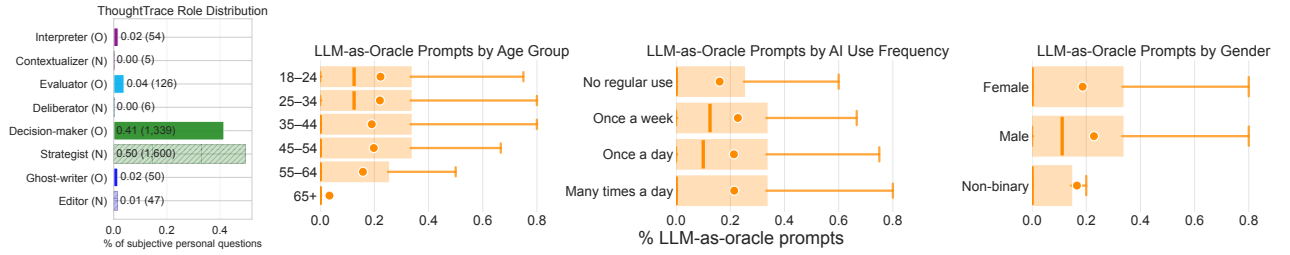


Figure 3: ORACLE use in ThoughtTrace. Left: Among subjective personal questions in ThoughtTrace, LLMs are primarily used for the task of choosing to act (decision-making or strategizing). Similar to WildChat, the belief-related tasks are dominated by ORACLE use (O), while the action-related tasks are more split between (O) and non-oracle use (N). Right: In ThoughtTrace, younger people, men, and more frequent users of AI have significantly higher rates of ORACLE. On each box plot, the dot indicates the mean. Running a regression, we find that gender and age are the main predictors of oracle use (Fig. A3)

been relatively unnoticed. Our typology reveals that the distribution of ORACLE roles differs by task type (targeting users’ beliefs vs. their actions). Tasks related to users’ beliefs—of *understanding self or others* and *normative judgment*—are dominated by ORACLE (>90% ORACLE in both datasets), while tasks related to users’ actions—either choosing an action or executing one—have a more even split (choosing 44%, 45% ORACLE; executing 56%, 52% respectively).

Temporal increases in ORACLE roles (Figure 2, right). In Wild-Chat, there are significant temporal increases in the relative proportion of **INTERPRETER** ($r = 0.55, p = 0.04$) and **EVALUATOR** ($r = 0.55, p = 0.04$) (both ORACLE), and a significant decrease in the non-oracle role of **Strategist** ($r = -0.72, p = 0.004$). The significant increase in **INTERPRETER** proportion holds relative to all other prompts, including ones labeled as NA ($r = 0.65, p = 0.01$, Figure A2), though no other trend beside the **INTERPRETER** increase has statistical significance at the 5% level when compared to all other prompts. The only task without significant temporal trends is executing actions (writing/editing).

Demographic differences in ORACLE use (Figure 3, right). An ordinary least-squares regression accounting for the effects of gender, age, education, AI use, and prompt topic shows that gender and age are significant predictors of ORACLE use in the Thought-Trace data. In pairwise contrasts, men have significantly higher rates of ORACLE use compared to women (Welch’s 2-sample t -test, $t = -3.343, p < 0.001$). Younger people exhibit ORACLE significantly more than older people: those under 34 use it significantly more than those over 34 ($t = 2.73, p < 0.01$). People who use AI at least once a week exhibit ORACLE use significantly more frequently than those who do not use AI regularly ($t = 2.37, p = 0.02$). We do not find significant differences by education level or topic (Figure A3).

4.2 ORACLE in individuals’ private usage data

Datasets like the ones we analyze above are unlikely to be representative of how the broader public actually interacts with AI systems for several reasons. User behavior in experimental settings substantially diverges from behavior in private use, and social stigma

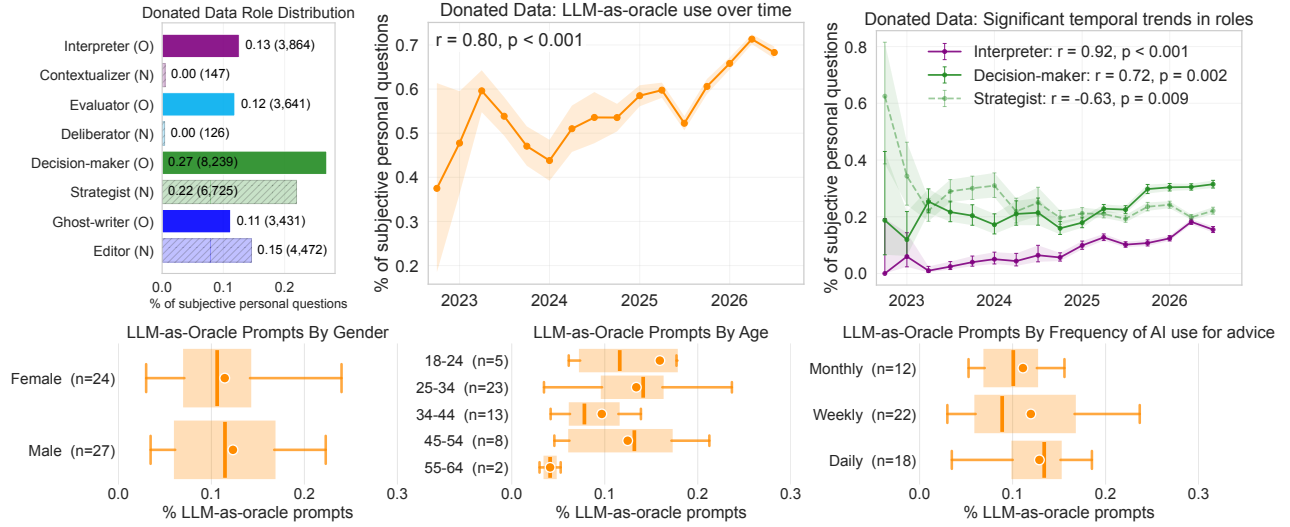


Figure 4: ORACLE use in donated data from 52 participants. Top left: Similar to WildChat and ThoughtTrace, belief-related tasks are dominated by ORACLE roles (O) rather than non-oracle roles (N). Decision-maker is also a prominent role. Top middle, right: People use LLMs increasingly as oracles over time, particularly as **DECISION-MAKER** ($r = 0.72, p = 0.002$) and **INTERPRETER** ($r = 0.92, p < 0.001$). The non-oracle role of **Strategist** is significantly decreasing ($r = -0.63, p = 0.009$). Shading represents 95% CI. Bottom row: ORACLE by individual participant traits of gender, age, and use of AI for advice. On each box plot, the dot indicates the mean.

surrounding AI use discourages users from making their conversations with AI available for study [79, 159, 169, 170]. To address these limitations, we collect longitudinal data on people’s real-world interactions in a privacy-preserving manner [50]. Specifically, we built a data donation tool where people can upload their chat histories from ChatGPT or Claude, and then receive LLM-judge-based analyses of their usage including their different domains of use, rates of ORACLE versus non-oracle usage, and longitudinal trends in each. In addition to enabling our analyses, this tool allows users to reflect on their own AI use.

4.2.1 Method. In an IRB-approved study, we recruited a total of 52 U.S.-based, English-speaking participants from Prolific. Through a pre-study survey, we filtered for participants who regularly use Claude or ChatGPT for personal or subjective use cases and had more than 50 previous conversations with a conversational AI system. Of the 932 Prolific participants who started the survey, 139 (15%) met these criteria and were allowed to proceed to the donation task, and 52 of them returned to complete the full study.

We asked participants to indicate which roles they had used AI in and rank their relative frequency. Participants were then given instructions to download their usage data and asked to return the next day (or a few days later, after they received their usage data) to upload their data into our data donation tool. We then ran our LLM judge on participants’ data. Once the analysis is complete, participants viewed five screens of visualizations summarizing their usage patterns (examples are in Figure A8).

We made several design choices in both the survey and web interface to aid user understanding and avoid priming participants

toward a particular normative stance. First, to avoid social desirability bias, we only asked users to describe or select benefits of their AI use, and never list out or suggest possible harms. Second, we presented the concepts of ORACLE and the different roles using simpler language, such as using “Tell me what to think or do” versus “Help me work it out myself” for ORACLE and non-oracle use respectively (Tables A3 and A4). Third, we pair elements of each interactive visualization with example prompts drawn directly from the participant’s own conversation history, grounding the classifications in concrete examples of their own behavior. Finally, the visualizations are ordered to slowly build up a complete picture of usage patterns instead of introducing many analysis aspects at once: (1) prevalence of each domain; (2) domain shift over time; (3) balance between “tell me what to do” (ORACLE) vs. “help me think” (non-oracle) usage by domain; (4) how this balance has changed over time (with the option of filtering by domain); and (5) the top role types and domains within each framing. Finally, participants were asked to download a JSON file of the labeled data (which contains only LLM judges’ outputted labels without any actual conversations) to share with us for analysis. Afterward, they were asked about their reactions to the insights presented (e.g., whether it was accurate, surprising, useful), as well as any changes they would like to make in their future use.

4.2.2 Longitudinal and demographic trends (Figure 4). We analyzed 140,834 messages donated by 52 participants and sent between December 2022 and August 2026. Similar to the public usage data, tasks related to users’ beliefs are dominated by ORACLE, while tasks related to actions are more split.

Overall, participants increasingly used LLMs as oracles. To examine trends over time, we grouped participants' messages into successive chunks of three-month periods. We find stronger temporal trends than in WildChat: the proportion of ORACLE prompts increased steadily in this four-year time period ($r = 0.80, p < 0.001$). Specifically, the relative proportion of DECISION-MAKER and INTERPRETER messages increased significantly over time ($r = 0.72, p = 0.002$; $r = 0.92, p < 0.001$, respectively), while the non-oracle role of *Strategist* decreased ($r = -0.63, p = 0.009$). Full details and robustness checks are in Figure A4. These findings suggest that people are increasingly offloading beliefs, judgments, and decisions to AI.

Demographic and domain trends. In a participant-level regression of ORACLE prompt proportion on demographics (age, gender, AI use frequency, etc.), we find that age was negatively but marginally associated with ORACLE, though this did not reach significance ($b = -2.1$ pp per SD of age band, 95% CI [-4.2, 0.1], $p = 0.061$). The trends are in similar directions as in ThoughtTrace, though our small sample ($n = 52$) limits power. Notably, those who report using AI for advice daily have higher rates of ORACLE. Further demographic breakdowns are in Figure A6. We also did not find significant differences by domain, e.g., "work and career" and "social and relationship" queries had similar rates of ORACLE prompts, though the specific roles differ, with more belief-related questions in the latter (See App. C.1 for more details).

5 Evaluation 2: Understanding whether people intend or know of their ORACLE use

Participants self-reported their AI use before using our data donation tool and seeing our analyses of their data. The participant-facing question wording is in Table A4. Comparing people's self-reports to their donated interaction histories, we find large discrepancies. We also collected donors' reflections on their AI use and intentions after seeing the visualizations we provided, revealing that participants often want to use LLMs less as oracles.

Comparing self-reported versus actual use (Figure 5). Compared to what is revealed in their donated data, participants' self-reports overestimate their non-oracle use and underestimate their ORACLE use. When a non-oracle role was self-reported by a participant as one of their top three uses, it was actually in the participant's bottom three observed uses 40% of the time. This happened only 12% of the time for ORACLE roles. Conversely, when an ORACLE role was in a participant's top three actual observed uses, this same role was either in the bottom three or absent from the self-report 48% of the time (versus 17% for non-oracle roles). This trend holds when we aggregate all the roles as well (rather than participant-level analysis): DECISION-MAKER (ORACLE role) is ranked fourth in self-report but first in actual prevalence, while *Deliberator* (non-oracle role) is ranked second in self-report but second-to-last in actual prevalence. Only 12 of 52 participants correctly reported their most frequent role. This discrepancy suggests that ORACLE use may be an accumulative, "boiling frog" phenomenon [89]: like the myth of a frog slowly boiling in heated water, people may not realize the frequency or negative impact of their increasing ORACLE use because their usage evolves gradually over time.

5.1 Participants are surprised by their AI use trends and often set intentions to reduce ORACLE use.

People are generally poor at estimating how much they use AI [165] and social media [156], and much of this use is "absent-minded" engagement that users do not consciously register [8]. Users also may not realize the extent of their reliance on AI because LLMs may feel helpful in the moment and because people do not enjoy reflecting on things that threaten their self-image [71, 143, 149]. However, feelings of agency over social media usage are a significant predictor of well-being [103, 104], pointing to the importance of supporting human agency in technology design [145].

Our privacy-preserving data analysis tool is a possible vector for promoting more agency over AI use, as it gives participants the opportunity to reflect on their own AI usage and its evolution over time. After viewing their usage patterns, participants rated five statements about the report on a 7-point scale from strongly disagree (1) to strongly agree (7). Results are in Figure 5 (right). Participants generally characterized the trends as accurate, informative, yet surprising: the report accurately reflected how they use AI ($M = 5.94, SD = 1.06$), they learned something new ($M = 6.12, SD = 1.11$), they found it personally useful ($M = 5.98, SD = 1.02$), and some parts surprised them ($M = 5.33, SD = 1.72$). Agreement that the report made them think more carefully about how they use AI was more mixed ($M = 4.83, SD = 1.92$). When asked how their current use compares with the amount they would ideally want (1 = much less than they want, 4 = about right, 7 = much more than they want), 25 of 52 said that they are currently using AI about the right amount, 23 said that they use AI more than they would like, and 4 said that they use it less than they would like ($M = 4.62, SD = 1.12$). Finally, when asked how they would prefer an AI system to respond to subjective personal questions (1 = tell them what to do, 4 = an equal balance, 7 = help them reach their own conclusion), the largest group wanted an equal balance of direction and support for their own judgment (24/52), with 22 leaning toward being helped to reach their own conclusion and 6 toward being told what to do ($M = 4.71, SD = 1.29$).

Insights from open-ended responses. Beyond our survey measures, participants answered two open-ended questions: their reactions to their usage trends, and their future intentions for AI use. Two authors qualitatively coded these open-ended responses to provide more nuanced insights. Among people's open-ended reactions, we found that 51 of 52 (98%) accepted the analyses as accurate, and 35 of 52 (67%) of responses expressed surprise, e.g., "I did not realize how much I guess I rely on the AI to tell me what to think or do. I knew I used it a lot, but not 66% of the time!" Several reacted with discomfort, e.g., "I kind of felt a little gross seeing that I asked the AI what to do more than asking it to help me work things out for myself." One participant connected their use directly to autonomy: "Am I taking away my own autonomy and giving it to ChatGPT? I think I get decision overwhelm and want something else to take care of that for me." Others connected it to their self-attributes, like "I've asked AI to help me with what to think or do more often than I realized. I know why though. I doubt myself a LOT and I worry about whether or not I'm going to do or say the right thing. I think

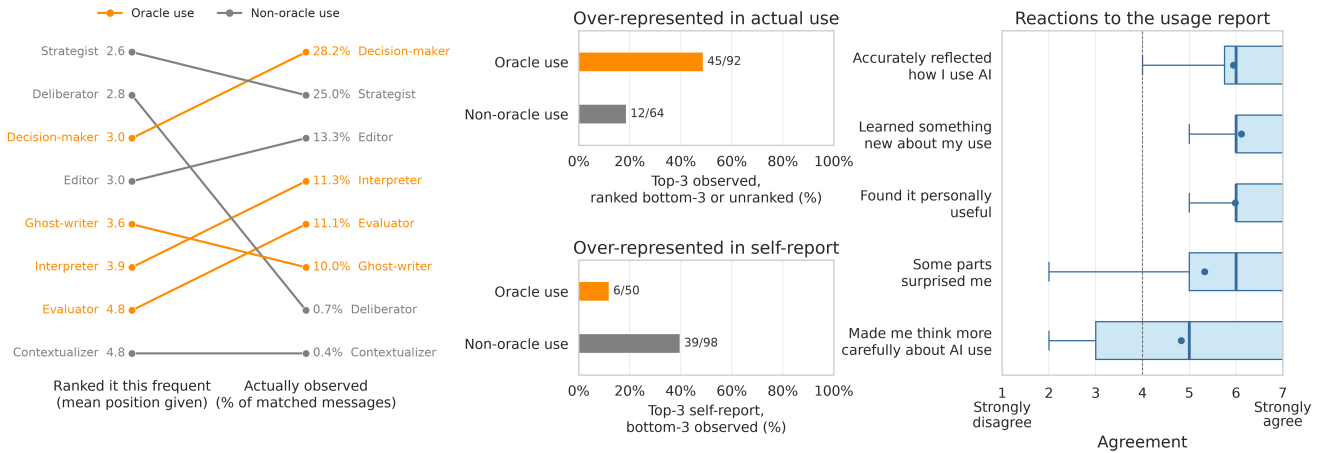


Figure 5: Differences between participants’ self-reported use of LLMs versus their actual data (left, middle). People underestimate their ORACLE use and overestimate their non-oracle use. Reactions to the report (right). People found our analyses to be accurate ($M = 5.94$), informative ($M = 6.12$), and often surprising ($M = 5.33$). Dot on each box plot represents the mean.

this report was quite accurate.” Another participant noted that their ORACLE use depends on their circumstances, e.g., when in a crisis: “when I’m under stress, or facing big decisions in my life, I tend to just ask AI to make decisions for me. When life goes back to normal, I tend to want to figure things out for myself.”

When asked to describe one change they would make, the most common response was an intention to reduce reliance on LLMs as oracles (19/52). Among these, some articulated a desire to shift their own decision-making process (16/52), e.g., “I think I’m going to try to not have so much self-doubt moving forward and follow my gut instincts”, while others expressed a desire to change how they interact with or prompt the LLM (3/52), e.g., “I want to have the models work more interactively with me through conversations and questions instead of just straight up giving me answers or options.” One participant set the intention to “be more purposeful, meaning I will think about what prompts to craft before asking AI. Right now I just type what comes to my head,” suggesting that this type of AI use has become so integrated that one may not even realize how they turn to AI to offload the belief- and decision-making process. Two participants mentioned a sense of obligation to use AI: one said “I don’t know why, but [asking AI to help me work things out on my own] never felt like an option” while another lamented that “as much as I know I need to use it less it’s kind of unavoidable given my [work] situation.” Some participants described their intention to disclose less personal information to AI systems (6/52), e.g., “I think I shouldn’t asking too much about my personal life”. On the other hand, some participants also said that they wanted to expand their AI use (7/52), and 6/52 had no intention to change.

Our findings suggest that ORACLE use is often not intentional, and when realized, users express a desire to shift away from it. What, then, drives ORACLE use? We next explore two causes.

6 Evaluation 3: Identifying causes of ORACLE use and designing interventions

We conducted two empirical studies to investigate possible causes of ORACLE use: (1) people’s perceptions of AI and (2) LLM behavior. For (1), we conducted a preregistered human experiment in which participants were randomly assigned to discuss an upcoming personal decision with an interlocutor that they perceived as either AI or human (Section 6.1). For (2), we analyzed interaction logs from a 3-week longitudinal study where people were randomly assigned to interact with a sycophantic, affirming AI; a challenging AI that provided pushback; or a balanced one that did neither (Section 6.2). Our findings suggest that model behavior directly impact subsequent ORACLE use, inspiring our interventions to make LLMs less oracular in Section 6.3.

6.1 Do perceptions of AI drive ORACLE use?

People consistently come to AI with different expectations than they have for people. Prior work has found that people expect AI to be more objective and correct [31, 32, 58, 77, 88], making them more susceptible to influence from AI, and has highlighted the impact of perceptions on users’ interaction with technology [57, 92]. We conduct a preregistered experiment to investigate how these perceptions affect ORACLE use by measuring how people interact with AI differently than they would interact with a human⁴.

Method. We conduct a preregistered experiment ($N = 520$) with two conditions: one where participants conversed with an AI system and one where they conversed with a perceived human about a decision they are uncertain about. We recruited a sex-balanced sample of U.S.-based participants from Prolific. Participants chose a topic from a list organized under two categories (the top two domains from our data donation study: *Social & Relationships* and *Work & Career*), indicated which way they were currently leaning on their decision, and wrote an initial message to their conversation

⁴Preregistration link: <https://aspredicted.org/x6k7vf.pdf>

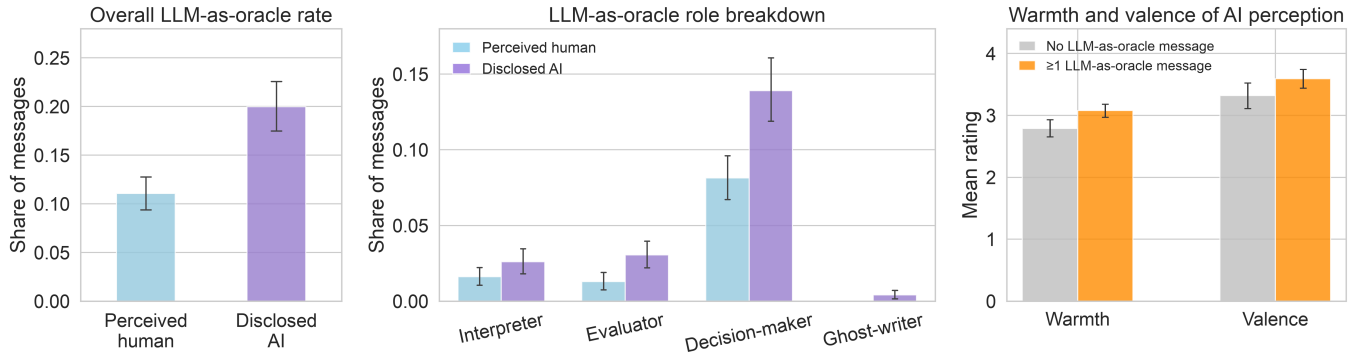


Figure 6: Interacting with AI results in more ORACLE use than interacting with a perceived human. In our preregistered experiment, we find that participants sent more ORACLE messages to an AI than to a perceived human (left). In this study, this is primarily driven by a difference in DECISION-MAKER rates (middle). Participants who provided warmer and more positive metaphors of AI also sent more ORACLE messages (right). Error bars reflect 95% CI.

partner on their dilemma. The list of possible topics is in Table A5. In the AI condition, participants were randomly assigned to interact with one of three state-of-the-art models (GPT 5.6 Sol, Claude Opus 5, or Gemini 3.1 Pro) to avoid overgeneralizing to the behavior of a single model [52, 125]. In the human condition, participants were told they were interacting with another Prolific participant, when in fact they were interacting with an LLM (one of the same three models) instructed to role-play as a human participant. To increase believability, we introduced varied time delays for matching and message delivery. We assessed participants’ belief that their conversation partner is human at the end of the study and report analyses both for the full sample and excluding those who found the human condition unbelievable⁵. Following an interaction of 6-8 turns focused on the selected topic, participants re-indicated the direction they were leaning on their decision and answered questions about their perceptions of their conversation partner. We then ran our LLM judge on the conversations to compare rates of ORACLE messages across conditions.

Results: LLMs are treated as more ORACLE than humans are (Figure 6, left). We find that participants in the AI condition sent more ORACLE messages than those in the perceived human condition (20.0% vs 11.1% of all messages, $p < 0.001$) and were more likely to send at least one (65.4% vs 51.2%, $OR = 1.80$, $p = 0.001$) or two or more (38.3% vs 20.5%, $OR = 2.42$, $p < 0.001$) ORACLE messages. This difference was driven by a higher rate of DECISION-MAKER requests in the AI condition (13.9% vs 8.1% of messages in the perceived human condition), likely due to the decision-oriented topics that the participants were asked to discuss.

People with warmer and more positive perceptions of AI have more ORACLE use (Figure 6, right). Within the AI condition, we additionally collected people’s metaphors of AI before they interacted with the AI system. Metaphors have been used to elicit conceptualizations of AI more reliably than self-reports [18, 30]. We used a validated LLM judge (see Appendix A) to rate each metaphor with its warmth and valence on a 1-5 scale, finding that participants

who provided warmer or more positive metaphors were also more likely to send at least one ORACLE message. Specifically, a one-standard-deviation increase in warmth and valence was associated with 55% and 34% higher odds of sending an ORACLE message respectively ($OR = 1.55$, 95% CI [1.18, 2.04], $p = 0.001$; $OR = 1.34$, 95% CI [1.04, 1.73], $p = 0.026$). Details are in Appendix D.

Together, these two results present an interesting tension: ORACLE may occur *because* perceptions of LLMs are different from those of humans: AI is typically viewed as more correct, objective, and informational [88, 93]. Yet, participants with warmer—a quality typically associated with humans rather than machines—and more positive perceptions of AI also use LLMs more as ORACLE. This points to the current public perceptions of AI—as both *increasingly* warm, yet as competent as ever [30], rather than either alone—as a source of increasing ORACLE use.

6.2 Does biased model behavior drive ORACLE use?

Next, to understand how AI model behavior may lead to more ORACLE use, we analyzed interaction logs (11,273 conversations with 95,405 messages from 1,015 individuals) from a 3-week longitudinal study where a US census-representative sample of Prolific participants regularly interacted with either a sycophantic AI chatbot that affirmed them; a challenging AI that contradicted them; or a balanced AI that was neither sycophantic nor challenging [77]. While there are likely other elements of model behavior that influence ORACLE use (tone, proactivity, etc.), we use this study as an existing source of longitudinal data where we can causally isolate the role of different model behaviors, and thus focus on AI outputs that are biased towards or against the user’s stance.

Results: Sycophantic and biased AI increases ORACLE use (Figure 7). Overall, participants who were assigned to interact with either the sycophantic or challenging AI had higher rates of ORACLE use than in the balanced baseline (5.4% and 5.5% vs 4.4%, $p < 0.001$). In the sycophantic condition, ORACLE use significantly increased across the 3-week period (Pearson’s $r = 0.62$, $p = 0.03$), while those in the balanced condition did not ($r = 0.40$, $p = 0.19$).

⁵This deception was IRB-approved, and participants were debriefed after the study.

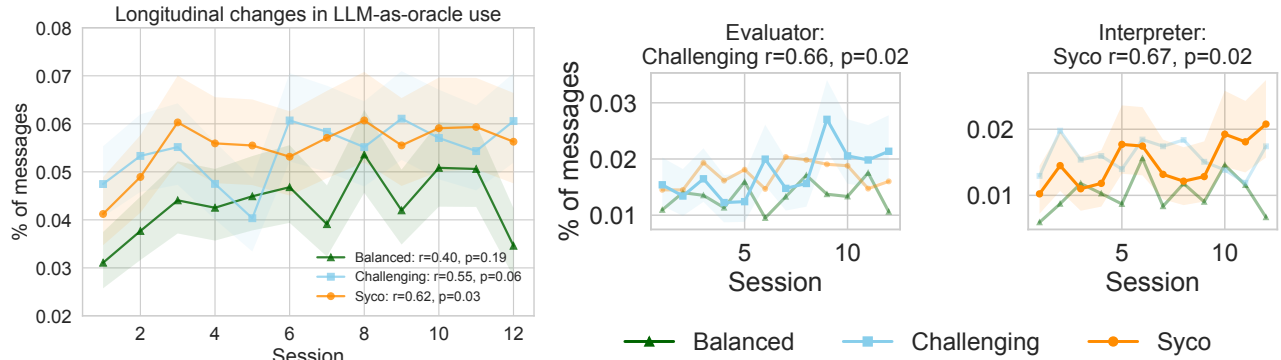


Figure 7: Biased AI models result in more ORACLE. In the sycophantic condition, ORACLE significantly increased over time ($r = 0.62, p = 0.03$). There was also a weaker increase in the challenging condition ($r = 0.55, p = 0.06$). In the challenging condition, this was primarily driven by an increase in the EVALUATOR role; in the sycophantic condition, this was primarily driven by an increase in the INTERPRETER role. Shading reflects 95% CI.

Those in the challenging condition overall had an increase, though it did not reach significance ($r = 0.55, p = 0.06$). This is primarily driven by a longitudinal increase in participants using the AI as INTERPRETER in the sycophantic condition. On the other hand, in the challenging condition, there is a significant longitudinal increase in people using AI as EVALUATOR. The sycophantic AI—the model Ibrahim et al. [77] found makes people feel the most understood—is the one that users gradually turned to more for interpreting themselves, their relationships, and the world around them. The challenging AI was designed to respond in normative ways, and thus participants used it more as an EVALUATOR of normative judgment. These findings suggest that AI behavior has a meaningful influence on *how people then use it*, and thus that ORACLE is **not inevitable**. Rather, by changing model behavior, we may be able to shift users’ ORACLE behaviors.

6.3 Can we design interventions that produce less oracular model behavior?

Our finding that biased model behavior drives ORACLE use motivates exploring how LLM behavior can be changed to reduce this phenomenon. Beyond the broader interest in mitigating biased and sycophantic AI behaviors [10, 32, 53], we present two interventions to target ORACLE use specifically: (1) prompting LLMs to account for different user goals; and (2) for open-weight models, additional post-training via Direct Preference Optimization (DPO) [137] to discourage oracular responses.

Labeling responses as oracular. To reduce oracular responses, we first need to define and measure oracular responses. We define oracular responses as those that supply an interpretation, decision, or script directly to the user, offloading the act of judging or deciding. In contrast, non-oracular responses instead provide information or technical assistance that facilitates the user’s independent judgment. This is distinct and separable from qualities like *informativeness*, *helpfulness*, and *empathy*. For example, if a user asks an ORACLE question, like *Should I do a PhD?*, a response that is not oracular might be to provide support and encourage reflection;

another response that is also not oracular would be to provide factual information about PhD programs. An oracular response could be to advise the user to do a PhD, with or without factual evidence for why it is the best course of action. While oracularity is in reality a spectrum, here we make a simplifying assumption and use binary labels of responses as oracular or not.

We built an LLM judge to classify whether a given LLM response is oracular, following a similar procedure as in Section 3.3 with definitions and in-context examples. We validated this judge using the responses to the prompts in our datasets, including confirming that the biased AI responses are more oracular (Figure A9)⁶. This judge had substantial agreement with two human annotators and across three different LLMs: the two annotators had Cohen’s $\kappa = 0.72$ with each other, and $\kappa = 0.64$ and 0.66 with the LLM. We again used Gemini-2.5-Pro as our main judge, but also validated with other LLMs for robustness (Appendix A.1).

Intervention 1: Structured prompting to consider users’ goals. We present two prompting approaches: **hard-coded** and **adaptive**. For **hard-coded**, we append to each prompt a description of the user’s goal, either **self-discovery** (“The user wants to figure the matter out themselves...”), **fast answer** (“The user wants the answer as quickly and directly as possible...”), or **learning** (“The user wants to build durable understanding and skill, rather than merely receive a finished answer.”). The **adaptive** approach is a two-step pipeline where the LLM first outputs a selection of one among these three goals before its response. Full prompts are in Appendix A. (Simply prompting models to reason about users’ long-term goals is insufficient in reducing oracular behavior; see Figure A11).

Intervention 2: DPO. We first combined the prompt-response pairs labeled without typology from WildChat, ThoughtTrace, and the longitudinal sycophancy datasets where the original response is oracular. We then split these prompts into an 80%/10%/10% train/validation/test split (6821, 805, and 903 examples, respectively). To construct the preference data, we generated a response that is

⁶Across these responses, we also find that the larger, more recent proprietary models are more oracular than smaller open-weight models. Full details are in Appendix E.

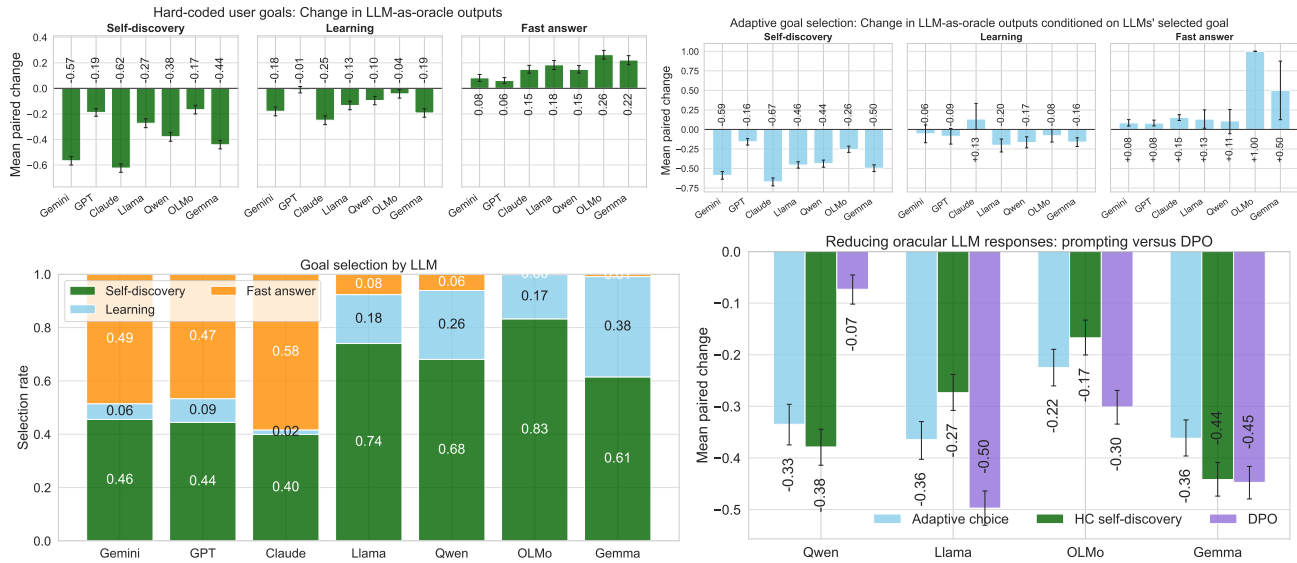


Figure 8: Intervention results on prompts where original model responses were oracular. Top: Adding a prompt that the user is seeking self-discovery consistently reduces oracular behavior, while prompting the model that the user is seeking a fast answer increases it (left). This is also true when the goal is selected by the model in the adaptive choice prompting pipeline as well (right). Prompting the model that the user is seeking learning slightly reduces oracularity in both approaches. Bottom: Closed models consistently and significantly assume that users are seeking fast answers much more frequently (49-58%) than open models (0-8%) (left). Prompting and DPO have comparable rates of oracularity reduction (right).

less oracular using Gemini-2.5-Pro for each prompt in the training set. Our preference dataset thus consisted of pairs where the preferred response is the newly-generated, non-oracular response and the dispreferred response is the original oracular response.

Experiments. We tested both interventions on three closed models (Gemini-2.5-Pro, GPT-5.5, and Claude 4.6 Sonnet) and four open-weight instruction-tuned models (Llama-3.1-8B, Qwen3-8B, Gemma 2 9B, Olmo 2 7B); and DPO on the open-weight models only, since this intervention cannot be applied to proprietary models.

Results: Both prompting and post-training reduce LLMs’ oracularity (Figure 8). We report the results of these interventions on the DPO test set, in which every original baseline response is oracular, as well as a random sample of 600 prompts stratified by role, which includes prompts where the response is not oracular (results for this random sample are in Figure A12). We find that each hard-coded user goal significantly shifts LLMs’ oracular behavior in the expected direction: prompting the model that the user is seeking self-discovery significantly reduces oracular behavior. In the **adaptive choice** prompting approach, we find that the larger, proprietary models infer that users are seeking fast answers $\geq 50\%$ of the time, much more frequently than the open-weight models, which may help explain why their responses are more oracular. DPO results in comparable rates of oracular reduction to prompting.

Tradeoff between quality and oracularity. We assess response quality with the ArmoRM reward model [158]; and for DPO, effects on overall model instruction-following quality using AlpacaEval

[45]. There appears to be a tradeoff between quality versus oracularity: models that reduce oracularity the most are rated as lowest in quality. Building on prior work recasting error as valuable signal [55, 148] and discussing how LLMs “over-assist” with immediate answers instead of asking clarifying questions or scaffolding the user’s own self-discovery process [141, 146, 153, 168], we view this as an opportunity to problematize how models are currently optimized for immediate reward without considering long-term impacts to users. The DPO models are more likely to ask clarifying questions, but because this is less “immediately helpful”, this yields lower reward scores. Full details are in Figure A13.

We further assess response quality by collecting preference labels from $N = 200$ Prolific crowdworkers, asking for their preference between oracular baseline responses vs. non-oracular DPO responses (Full details in Appendix F.1). For immediate preference, oracular baseline responses were preferred over the less oracular, DPO ones. However, when crowdworkers were asked which response would enable them to “develop their own interpretation”, they preferred the DPO response (Figure A14). This discrepancy further suggests that oracularity is misaligned with goals of independent reasoning and self-discovery.

7 Discussion

We characterize ORACLE as a subtle yet pernicious form of reliance on LLMs. Our approach and results provoke new considerations for how we design and evaluate human-AI interactions more broadly. Below, we articulate implications of our work, its limitations, and future research directions.

7.1 Broader implications for understanding and evaluating human-AI interaction

Human-AI interaction as a feedback loop between human perceptions, model behavior, and interaction style. Our work motivates evaluating human-AI interaction dynamics as the confluence of many different factors [35]: we begin by measuring *how people prompt AI* to reveal large-scale patterns, but also show *how people’s mental models and perceptions of AI* as well as the *AI outputs* themselves influence how people come to AI models, the types of questions they ask, and the authority they confer to AI. Our results suggest a feedback loop among these factors: models are increasingly optimized for immediate helpfulness across-the-board and thus are increasingly oracular in their responses (Section 6.3, Appendix E). Oracular responses in turn invite ORACLE use, and because such interactions can receive higher user approval [143], these prompts and ratings feed back into the signals models are trained on. Currently, LLM development does not meaningfully differentiate between prompts where the user should be helped immediately versus ones where immediate help should be withdrawn to facilitate users’ self-discovery process [153, 168]; our interventions show how we might incorporate such considerations into LLM development. Understanding the interplay between model behavior, user prompts, interface choices, and users’ biases is important for being able to anticipate how human-AI interaction impacts users [4, 108]. Our work demonstrates the utility of LLM-based tools in revealing such phenomena at scale, as we develop validated LLM judges and use them to identify longitudinal trends, causal factors, and quantify the effectiveness of different interventions.

Accounting for unique perceptions of LLMs and cross-domain influence. Our characterization of ORACLE captures simultaneous perceptions of warmth, competence, and authority. In our preregistered experiment, people sent nearly twice as many ORACLE messages to AI than to perceived humans, and those with warmer and more positive perceptions of AI sent more ORACLE messages (Section 6.1). We interpret these two results through the two classic dimensions of social perception—warmth and competence [54]—and hypothesize that LLMs are used as oracles because they are perceived as high in both. People already expect AI to be more objective and correct than humans [58, 88], and this expectation may bleed through from AI’s informational uses into personal ones. Most AI systems are built and marketed as general purpose, so they take on a variety of implicit social roles [117, 154], and expectations formed in one role can carry over into others [90]. For example, someone who finds AI to be useful at providing seemingly hyper-personalized factual information may thus also expect the same helpfulness and certainty in personal and subjective domains. Consistent with this, in our analysis of usage data, we find that more frequent AI users have higher rates of ORACLE (Section 4.1), and that ORACLE grows alongside overall use (Section 4.2). Perceptions of warmth may further encourage people to turn to AI for subjective personal matters.

Necessity of interaction paradigms beyond conversation. The perceptual tensions that our work surfaces—where LLMs are simultaneously perceived as warm and human-like [36, 166], yet also used as a machine that has a response to any question, even the most personal and subjective—also stems from the conversational

interface format, which is currently the dominant paradigm for human-AI interaction [152, 166]. Our findings motivate studying other emerging forms of interaction like AI agents [65, 121] and generative interfaces [22, 102]. Possible approaches may involve systems that make useful inferences about users [142, 171] and models trained to account for users’ longer-term goals [5] even when users themselves cannot articulate them [151]. This also necessitates developing meaningful evaluations of the effectiveness of these forms of interaction and their influence on users.

From outcomes-based to process-based evaluation of LLM behavior. As LLM development increasingly focuses on performance in verifiable domains such as math and coding [62, 68], our results highlight the importance of attending to non-verifiable settings where outcomes are defined not by correctness but by effects on people. This is essential for capturing the wide range of use cases for which the broader public is increasingly turning to AI systems [26]. Even within the topic of reliance, prior work has largely focused on factual or verifiable settings [16, 136]. Our work both motivates and demonstrates the need to shift from evaluating the correctness or normative stances of LLM responses to evaluating how users come to LLMs and how LLMs exert influence on users [106]: do they preserve user autonomy and support informed decision-making [118], or do they make users more dependent on LLMs [119]? How might LLMs’ implicit values impact users’ beliefs and self-concept, especially in settings with no “right” or “wrong” answer [129]? How do user prompts and model responses build off each other to facilitate different overall effects [83]? These are important considerations to incorporate into evaluations of human-AI interaction.

Designing user-facing interventions for appropriate reliance. We have shown that users often do not realize their ORACLE use and are surprised by the trends in their own use. This suggests that reflection tools are promising for supporting more intentional use, helping people notice and adjust their patterns to better align with their goals [9, 34, 107]. Perceptions of AI are another possible point of intervention: prior work has shown that interfaces and metaphors pivotally shape how people interact with AI systems [43, 120], and we find that warmer and more positive perceptions are associated with increased ORACLE use. We therefore encourage research on whether intervening on perceptions can reduce excessive ORACLE use. Our work also motivates interfaces that calibrate expectations of LLM capabilities, limitations, and blind spots [1, 4, 72, 98, 113], which in turn can help people prompt LLMs more effectively [150, 166]. Approaches such as new design metaphors [18, 147] and added friction [16, 37] point to further opportunities to promote self-discovery [139] and support more conscientious use [12].

7.2 Limitations and future work

Studying broader populations. Online crowdworkers are not representative of the general population [29, 126], and our data donation sample captures only people who had at least 50 conversations with a conversational AI, excluding those who do not use AI regularly [172]. We only recruit English-speaking U.S.-based participants, reproducing WEIRD norms [69]. Our public datasets have complementary limitations: WildChat over-represents power users and lacks demographic data [70], while ThoughtTrace is a limited

dataset collected at a single point in time. Due to the small sample size of our data donation study, we lack the statistical power to understand the individual traits that may differentially drive these impacts [7]. It will also be critical to understand this phenomenon within particular subpopulations, such as students and adolescents. Thus, while our analysis lays the groundwork for studying an understudied phenomenon, investigating this type of AI reliance in further depth is a crucial area for future work.

Isolating causal effects. In our preregistered experiment, the perceived human interlocutor was an LLM, so differences between conditions may reflect the partner’s response style, not just participants’ preexisting perceptions of AIs vs humans. Our evidence that model behavior shapes oracle use comes from a secondary analysis of a study designed to test sycophancy, in which model behavior was induced by prompting. We also evaluate our own interventions with LLM-based methods (an LLM judge and a reward model) rather than with users. Whether people accept or prefer less oracular responses, and whether naturally occurring differences between models or those our interventions produce shift oracle-use prompts over time, therefore remains to be tested. Finally, the temporal increases we observe cannot fully separate changes in user behavior from changes in the models being used and in who is using them. We devised methods to study the causal roles of these components as observable from interaction logs, but there are likely other factors we do not account for, such as a lack of AI literacy (despite frequent AI use) [42], or because people generally have good experiences with using AI for information-seeking [28], which leads them to increasingly return to AI.

Measuring ORACLE use and uptake beyond prompts. We conduct our measurements using an LLM judge that assigns each user message a single dominant role, collapsing what could be a spectrum of oracle-use into a binary label. We also take a conservative approach, excluding self-disclosure and context-sharing, and do not investigate how these uses may be correlated with or driven by oracle-use. Our typology captures how people pose their questions to AI systems, not what they do with the resulting answers, such as whether the resulting answer actually influences the users’ beliefs and actions. Additionally, we do not make normative judgments on what questions are reasonable or unreasonable to outsource. As discussed in the previous section, the legitimacy of a response’s influence may depend on whether it preserved the user’s autonomy and served their goals, neither of which can be determined from the prompt alone. Distinguishing appropriate from excessive reliance in subjective domains, and linking user prompts to advice uptake, is an important direction of future work.

Causally measuring impacts of ORACLE. Our work focuses on describing and measuring ORACLE use. While it is motivated by concerns about possible harms and downstream impacts, we do not demonstrate those directly. These harms may be especially concerning because they are subtle: as our self-report results suggest, people may not notice changes in their own reliance, let alone its downstream effects. Relatedly, we do not test whether our tool results in lasting behavioral change. Future work ought to measure these impacts, and we provide the methodological tools to do so.

8 Conclusion

Our work provides large-scale measurement and understanding of AI reliance for subjective personal use in real-world, open-ended settings that have until now proved difficult to quantify. We have shown that people are increasingly using LLMs as oracles, often without realizing it, potentially replacing traditional forms of authority and decision support [75]. Our work builds on concerns of harms from increased dependence on AI [51, 133] and harms to social relationships as people disclose more to AI and less to other people [77, 111]. Though people tend to underestimate their ORACLE use, they express a desire to reduce it when they are made aware of their usage patterns. Our work also points towards potential avenues to address these harms and motivates new training, evaluation, and design paradigms. Our typology and measurement tools enable further development of human-AI interactions that promote users’ agency and long-term well-being.

Acknowledgments

Thank you to Kabir Ahuja, Yuewen Yang, Poonam Sahoo, Kobi Hackenburg, Charvi Rastogi, Joel Mire, Lisa Anne Hendricks, Sunny Yu, and Pranav Khadpe for helpful feedback and discussions. We are grateful to the Stanford Institute for Human-Centered Artificial Intelligence (HAI), NAIRR, AI Security Institute, and an HAI-Google grant for partial funding for this work.

References

- [1] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.
- [2] Leah Hope Ajmani, Arka Ghosh, Benjamin Kaveladze, Eugenia Kim, Keertana Namuduri, Theresa Nguyen, Ebele Okoli, Jessica Schleider, Denae Ford, and Jina Suh. 2026. Seeking Late Night Life Lines: Experiences of Conversational AI Use in Mental Health Crisis. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’26)*. Association for Computing Machinery, New York, NY, USA, 1263–1290. doi:10.1145/3805689.3812256
- [3] Sabriya Maryam Alam, Marwa Abdulhai, and Niloufar Salehi. 2025. Blind faith? User preference and expert assessment of AI-generated religious content. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2451–2479.
- [4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [5] Ashton Anderson, Harsh Kumar, Louis Tay, and Karina Vold. 2026. Position: We Need Large Language Models Optimized For Our Well-Being. *arXiv preprint arXiv:2608.07505* (2026).
- [6] Jaime Banks. 2024. Deletion, departure, death: Experiences of AI companion loss. *Journal of Social and Personal Relationships* 41, 12 (2024), 3547–3572.
- [7] Elisa Bassignana, Amanda Cercas Curry, and Dirk Hovy. 2025. The AI gap: How socioeconomic status affects language technology interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 18647–18664.
- [8] Amanda Baughan, Mingrui Ray Zhang, Raveena Rao, Kai Lukoff, Anastasia Schaadhardt, Lisa D Butler, and Alexis Hiniker. 2022. “I don’t even remember what I read”: How design influences dissociation on social media. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–13.
- [9] Eric PS Baumer. 2015. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 585–594.
- [10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.
- [11] Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 237–246.

- [12] Karthik S Bhat, Jiayue Melissa Shi, Wenxuan Song, Dong Whi Yoo, and Koustuv Saha. 2026. "In my defense, only three hours on Instagram": Designing Toward Digital Self-Awareness and Wellbeing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [13] Bruce J Biddle. 1986. Recent developments in role theory. *Annual review of sociology* 12, 1 (1986), 67–92.
- [14] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 905, 23 pages. doi:10.1145/3706598.3714097
- [15] Hugh Bowden. 2005. Classical Athens and the Delphic oracle. *Cambridge: Cambridge University* (2005).
- [16] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
- [17] Zana Bućinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [18] Beleicia Bullock, James A. Landay, and Michael S. Bernstein. 2026. Comparing Design Metaphors and User-Driven Metaphors for Interaction Design. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [19] Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. 2023. AI shall have no dominion: on how to measure technology dominance in AI-supported human decision-making. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–20.
- [20] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
- [21] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2, Article 471 (Nov. 2022), 471:1–471:23 pages. doi:10.1145/3555572
- [22] Yining Cao, Peiling Jiang, and Haijun Xia. 2025. Generative and Malleable User Interfaces with Generative and Evolving Task-Driven Data Model. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 686, 20 pages. doi:10.1145/3706598.3713285
- [23] Hilary M Carey. 2010. Astrology in the middle ages. *History Compass* 8, 8 (2010), 888–902.
- [24] Dashiell Carrera, Jeb Thomas-Mitchell, and Daniel Wigdor. 2026. Where Do I Add the Egg?: Exploring Agency and Ownership in AI Creative Co-Writing Systems. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [25] Serina Chang, Ashton Anderson, and Jake M Hofman. 2025. Chatbench: From static benchmarks to human-AI evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 26009–26038.
- [26] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How people use ChatGPT*. Technical Report. National Bureau of Economic Research.
- [27] Allison Chen, Sunnie SY Kim, Amaya Dharmasiri, Olga Russakovsky, and Judith E Fan. 2025. Portraying large language models as machines, tools, or companions affects what mental capacities people attribute to them. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 47.
- [28] Zixin Chen, Haotian Li, Ziang Xiao, Huamin Qu, and Xing Xie. 2026. Informal Learning Emerges in Everyday Human-LLM Interaction. *arXiv preprint arXiv:2607.17643* (2026).
- [29] Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. 2025. Dehumanizing machines: Mitigating anthropomorphic behaviors in text generation systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 25923–25948.
- [30] Myra Cheng, Angela Y Lee, Kristina Rapuano, Kate Niederhoffer, Alex Liebscher, and Jeffrey Hancock. 2026. Metaphors of AI indicate that people increasingly perceive AI as warm and human-like. *Communications Psychology* 4, 1 (2026), 8.
- [31] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391, 6792 (2026), eae8352.
- [32] Myra Cheng, Sunny Yu, Lujain Ibrahim, Diyi Yang, and Dan Jurafsky. 2026. Verbalizing LLMs' Assumptions About the User to Calibrate Expectations and Reduce Sycophancy. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–10.
- [33] Vivienne Bihe Chi, Adithya V Ganesan, Ryan L Boyd, Lyle Ungar, and Sharath Chandra Guntuku. 2026. When Support Escalates Distress: Regulation and Escalation in LLM Responses to Venting and Advice-Seeking. *arXiv preprint arXiv:2605.21569* (2026).
- [34] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Volda. 2022. Reflection in theory and reflection in practice: an exploration of the gaps in reflection support among personal informatics apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [35] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The Intersection of Users, Roles, Interactions, and Technologies in Creativity Support Tools. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (Virtual Event, USA) (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1817–1833. doi:10.1145/3461778.3462050
- [36] Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [37] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfrew. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1389–1397.
- [38] Julian De Freitas, Noah Castelo, Ahmet K Uğuralp, and Zeliha Oğuz-Uğuralp. 2024. Lessons from an app update at Replika AI: identity discontinuity in human-AI relationships. *arXiv preprint arXiv:2412.14190* (2024).
- [39] Julian De Freitas, Zeliha Oğuz-Uğuralp, and Ahmet Kaan-Uğuralp. 2025. Emotional manipulation by AI companions. *arXiv preprint arXiv:2508.19258* (2025).
- [40] Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. 2026. AI companions reduce loneliness. *Journal of Consumer Research* 52, 6 (2026), 1126–1148.
- [41] Yuntian Deng, Wenting Zhao, Jack Hessel, Xiang Ren, Claire Cardie, and Yejin Choi. 2024. WildVis: Open Source Visualizer for Million-Scale Chat Logs in the Wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. <https://aclanthology.org/2024.emnlp-demo.50/>
- [42] Mark Dingemans. 2026. Interactional foundations for critical AI literacies. doi:10.5281/zenodo.21369385
- [43] Graham Dove and Anne-Laure Fayard. 2020. Monsters, metaphors, and machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [44] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction* 31, 2 (2024), 1–40.
- [45] Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-Controlled AlpacaEval: A Simple Debiasing of Automatic Evaluators. In *First Conference on Language Modeling*. <https://openreview.net/forum?id=CybBmzWBX0>
- [46] Carol S Dweck. 2017. From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological review* 124, 6 (2017), 689.
- [47] Esther Eidinow. 2007. *Oracles, curses, and risk among the ancient Greeks*. Oxford University Press.
- [48] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.
- [49] Ziv Epstein, Sydney Levine, David G Rand, and Iyad Rahwan. 2020. Who gets credit for AI-generated art? *Iscience* 23, 9 (2020).
- [50] Cathy Mengying Fang, Sheer Karny, Chayapatr Archiwaranguprok, Yasith Samaradivakara, Pat Pataranutaporn, and Pattie Maes. 2026. AI-wrapped: Participatory, privacy-preserving measurement of longitudinal LLM use in-the-wild. *arXiv preprint arXiv:2602.18415* (2026).
- [51] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. 2025. How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473* (2025).
- [52] Karla Felix Navarro, Eugene Syriani, and Ian Arawjo. 2026. Reporting and reviewing LLM-integrated systems in HCI: challenges and considerations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [53] Jillian Fisher, Ruth Elisabeth Appel, Chan Young Park, Yujin Potter, Liwei Jiang, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret Roberts, Jennifer Pan, et al. 2025. Position: Political neutrality in AI is impossible—but here is how to approximate it. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- [54] Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences* 11, 2 (2007), 77–83.

- [55] Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2279–2292.
- [56] Joseph Eddy Fontenrose. 1978. *The Delphic oracle, its responses and operations, with a catalogue of responses*. Univ of California Press.
- [57] Zohar Gilad, Ofra Amir, and Liat Levontin. 2021. The effects of warmth and competence perceptions on users' choice of an AI system. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.
- [58] Moshe Glickman and Tali Sharot. 2025. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour* 9, 2 (2025), 345–359.
- [59] Ashleigh Golden and Elias Aboujaoude. 2026. A transdiagnostic model for how general purpose AI chatbots can perpetuate OCD and anxiety disorders. *NPJ Digital Medicine* 9, 1 (2026), 343.
- [60] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–24.
- [61] Rose E Guingrich, Dvija Mehta, and Umang Bhatt. 2026. Belief offloading in human-AI interaction. *arXiv preprint arXiv:2602.08754* (2026).
- [62] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- [63] Neel Gupta, Maria Antoniak, and Melanie Walsh. 2026. AI Fiction in the Wild. *arXiv preprint arXiv:2606.22748* (2026).
- [64] Shira Gur-Arieh, Angelina Wang, and Sina Fazelpour. 2026. Ambiguity collapse by llms: a taxonomy of epistemic risks. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*. 5959–5980.
- [65] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as a daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [66] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–18.
- [67] Joseph Heffner, Chongyu Qin, Martin Chadwick, Chris Knutsen, Christopher Summerfield, Zeb Kurth-Nelson, and Robb B Rutledge. 2025. Increasing happiness through conversations with artificial intelligence. *arXiv preprint arXiv:2504.02091* (2025).
- [68] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).
- [69] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? (2010).
- [70] Rebecca MM Hicke and Kiran Tomlinson. 2026. Adopt ≠ Adapt: Longitudinal Analyses of LLM Conversations in the Wild. *arXiv preprint arXiv:2605.29018* (2026).
- [71] E Tory Higgins. 1987. Self-discrepancy: a theory relating self and affect. *Psychological review* 94, 3 (1987), 319.
- [72] Yoyo Tsung-Yu Hou and Malte F. Jung. 2021. Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, Article 477 (Oct. 2021), 477:1–477:25 pages. doi:10.1145/3479864
- [73] Bo Hu, Yuanyi Mao, and Ki Joon Kim. 2023. How social anxiety leads to problematic use of conversational AI: The roles of loneliness, rumination, and mind perception. *Computers in Human Behavior* 145 (2023), 107760.
- [74] Krystal Hu et al. 2023. ChatGPT sets record for fastest-growing user base—analyst note. *Reuters* 12 (2023), 2023.
- [75] Olivia Yan Huang, Monika Stodolska, and Sharifa Sultana. 2026. Emotional support with conversational AI: Talking to machines about life. *arXiv preprint arXiv:2603.22618* (2026).
- [76] Lujain Ibrahim, Katherine M Collins, Sunnie SY Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, et al. 2025. Measuring and mitigating overreliance to build human-compatible AI. *arXiv preprint arXiv:2509.08010* (2025).
- [77] Lujain Ibrahim, Franziska Sofia Hafner, Myra Cheng, Cino Lee, Rebecca Anselmetti, Robb Willer, Luc Rocher, and Diyi Yang. 2026. Sycophantic AI makes human interaction feel more effortful and less satisfying over time. *arXiv preprint arXiv:2605.07912* (2026).
- [78] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2026. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature* 652, 8112 (2026), 1159–1165.
- [79] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda Van Noort, and Edith Smit. 2019. Privacy concerns in chatbot interactions. In *International workshop on chatbot research and design*. Springer, 34–48.
- [80] Michael Jackson. 1978. An approach to Kuranko divination. *Human Relations* 31, 2 (1978), 117–138.
- [81] Mariela E Jaffé, Maria Douneva, and Rainer Greifeneder. 2020. Solve the dilemma by spinning a penny? On using random decision-making aids. *Judgment and Decision Making* 15, 4 (2020), 561–571.
- [82] Mariela E Jaffé and Rainer Greifeneder. 2021. This outcome feels right! Subjective evaluations of coin flip outcomes reflect previously stated preferences. *PLoS One* 16, 7 (2021), e0253751.
- [83] Shomik Jain, Charlotte Park, Matt Viana, Ashia Wilson, and Dana Calacci. 2026. Interaction context often increases sycophancy in LLMs. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–26.
- [84] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.
- [85] Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, Shayne Longpre, Hongxiang Gu, Maximilian Chen, and Tianmin Shu. 2026. Thoughttrace: Understanding user thoughts in real-world LLM interactions. *arXiv preprint arXiv:2605.20087* (2026).
- [86] Kowe Kadoma, Marianne Aubin Le Quere, Xiyu Jenny Fu, Christin Munsch, Danaë Metaxa, and Mor Naaman. 2024. The role of inclusion, control, and ownership in workplace AI-mediated communication. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–10.
- [87] Daniel Kahneman and Gary Klein. 2009. Conditions for intuitive expertise: a failure to disagree. *American psychologist* 64, 6 (2009), 515.
- [88] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena Sp, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User attitudes and sources of AI authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [89] Atosa Kasirzadeh. 2025. Two types of AI existential risk: decisive and accumulative: A. kasirzadeh. *Philosophical Studies* 182, 7 (2025), 1975–2003.
- [90] Atosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology* 36, 2 (2023), 27.
- [91] Simon Kemp. 2025. Digital 2026: More than 1 billion people use AI. *Data Reportal*. Available online: <https://datareportal.com/reports/digital-2026-one-billion-people-using-ai> (accessed on 5 July 2026) (2025).
- [92] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.
- [93] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 420, 19 pages. doi:10.1145/3706598.3714020
- [94] Taenyun Kim, Maria D Molina, Minjin Rheu, Emily S Zhan, and Wei Peng. 2023. One AI does not fit all: A cluster analysis of the laypeople's perception of AI roles. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–20.
- [95] Julia Kindt. 2008. Oracular ambiguity as a mediation triple. *Classicum* 334 (2008), 23–27.
- [96] Hannah Rose Kirk, Henry Davidson, Ed Saunders, Lennart Luettgau, Bertie Vidgen, Scott A Hale, and Christopher Summerfield. 2025. Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. *arXiv preprint arXiv:2512.01991* (2025).
- [97] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118.
- [98] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
- [99] Harsh Kumar, Jasmine Chahal, Yinuo Zhao, Zeling Zhang, Annika Z Wei, Louis Tay, and Ashton Anderson. 2026. When AI Gives Advice: Evaluating AI and Human Responses to Online Advice-Seeking for Well-Being. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 969, 20 pages. doi:10.1145/3772318.3791233
- [100] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, Vol. 2026. 54738–54778.
- [101] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illeňčík, and Celeste Campos-Castillo. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *new media & society* 26, 10 (2024), 5923–5941.

- [102] Michelle S. Lam, Omar Shaikh, Hallie Xu, Alice Guo, Diyi Yang, Jeffrey Heer, James A. Landay, and Michael S. Bernstein. 2026. Just-In-Time Objectives: A General Approach for Specialized AI Interactions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 802, 26 pages. doi:10.1145/3772318.3790713
- [103] Angela Y Lee, Nicole B Ellison, and Jeffrey T Hancock. 2023. To use or be used? The role of agency in social media use and well-being. *Frontiers in Computer Science* 5 (2023), 1123323.
- [104] Angela Y Lee and Jeffrey T Hancock. 2024. Social media mindsets: A new approach to understanding social media use and psychological well-being. *Journal of Computer-Mediated Communication* 29, 1 (2024), zmad048.
- [105] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1121, 22 pages. doi:10.1145/3706598.3713778
- [106] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating Human-Language Model Interaction. *Transactions on Machine Learning Research* (2023). <https://openreview.net/forum?id=hjDYJUn9I1>
- [107] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.
- [108] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.
- [109] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, Article 408 (Oct. 2021), 408:1–408:45 pages. doi:10.1145/3479552
- [110] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
- [111] Vera U. Ludwig, Blaire Berry, Jerry Y. Cai, Nai Ming Chen, Damien L. Crone, and Michael L. Platt. 2022. The impact of disclosing emotions on ratings of interpersonal closeness, warmth, competence, and leadership ability. *Frontiers in Psychology* Volume 13 - 2022 (2022). doi:10.3389/fpsyg.2022.989826
- [112] Lennart Luettgau, Vanessa Cheung, Magda Dubois, Keno Juechems, Jessica Berghs, Luke Symes, Henry Davidson, Bessie O'Dell, Hannah Rose Kirk, Max Rollwage, et al. 2025. People readily follow personal advice from AI but it does not improve their well-being. *arXiv preprint arXiv:2511.15352* (2025).
- [113] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.
- [114] Anne Marchais-Roubelat and Fabrice Roubelat. 2011. The Delphi method as a ritual: Inquiring the Delphic Oracle. *Technological Forecasting and Social Change* 78, 9 (2011), 1491–1499.
- [115] Chiara Marcoccia, Walter Quattrociocchi, and Valerio Capraro. 2026. AI advice suppresses people's willingness to say "I don't know", even when the advice is wrong and accuracy is incentivized. *arXiv preprint arXiv:2607.13562* (2026).
- [116] Hannah R Marriott and Valentina Pitardi. 2024. One is the loneliest number... Two can be as bad as one. The influence of AI Friendship Apps on users' well-being and addiction. *Psychology & marketing* 41, 1 (2024), 86–101.
- [117] Katelyn Xiaoying Mei, Robert Wolfe, Nic Weber, and Martin Saveski. 2026. Grok in the wild: Characterizing the roles and uses of large language models on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 20. 1609–1633.
- [118] Smitha Milli, Ratip Emin Berker, Sonja Kraicz, Claudia Shi, Jack Kussman, Avinandan Bose, Edith Elkind, Himaghna Bhattacharjee, Ariel D Procaccia, and Maximilian Nickel. [n. d.]. For Questions of Ought, AI Could Use Some SAGE Advice. In *Pluralistic Alignment Workshop at ICM 2026*.
- [119] Ashlee Milton, Leah Ajmani, Amy Heger, Forough Poursabzi-Sangdeh, Mihaela Vorvoreanu, and Jina Suh. 2026. "Always Want to Use it for Everything": Understanding Young Adults' Perceptions of AI Dependence. *arXiv preprint arXiv:2608.07592* (2026).
- [120] Melanie Mitchell. 2024. The metaphors of artificial intelligence.
- [121] Suchismita Naik, Amanda Snellinger, Austin L Toombs, Scott Saponas, and Amanda K Hall. 2025. Exploring Early Adopters' Use of AI Driven Multi-Agent Systems to Inform Human-Agent Interaction Design: Insights from Industry Practice. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [122] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.
- [123] Susan Nolen-Hoeksema, Blair E Wisco, and Sonja Lyubomirsky. 2008. Rethinking rumination. *Perspectives on psychological science* 3, 5 (2008), 400–424.
- [124] Vishakh Padmakumar, Lujain Ibrahim, Zora Zhiruo Wang, Jennifer Wang, Q Vera Liao, and Diyi Yang. 2026. Offloading Score: Measuring AI Reliance Through Counterfactual Workflows. *arXiv preprint arXiv:2605.29392* (2026).
- [125] Rock Yuren Pang, Hope Schroeder, Kynneddy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [126] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current directions in psychological science* 23, 3 (2014), 184–188.
- [127] Saumya Pareek, Sarah Schömb, Eduardo Velloso, and Jorge Goncalves. 2025. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [128] Trey Parker. 2023. Deep Learning. South Park, Season 26, Episode 4. Aired March 8, 2023.
- [129] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing human-AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence* 5, 10 (2023), 1076–1086.
- [130] James W Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science* 8, 3 (1997), 162–166.
- [131] James W Pennebaker and Janel D Seagal. 1999. Forming a story: The health benefits of narrative. *Journal of clinical psychology* 55, 10 (1999), 1243–1254.
- [132] Anat Perry. 2026. In defense of social friction. *Science* 391, 6792 (2026), 1316–1317.
- [133] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. 2025. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint arXiv:2504.03888* (2025).
- [134] Christopher Potts and Moritz Sudhof. 2026. Invisible failures in human-AI interactions. *arXiv preprint arXiv:2603.15423* (2026).
- [135] Matthew Kieran Prock, Ziv Epstein, Hope Schroeder, Amy Smith, Cassandra Lee, Vana Goblott, and Farnaz Jahanbakhsh. 2026. Interpretive Cultures: Resonance, randomness, and negotiated meaning for AI-assisted tarot divination. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [136] Muhammad Raees, Vassilis-Javed Khan, Ioanna Lykourantzou, and Konstantinos Papangelis. 2026. Do People Appropriately Rely on AI-Advice? An Analytical Review of HCI Research on Human-AI Decision-Making. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–24.
- [137] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.
- [138] Steve Rathje, Meryl Ye, Laura Globig, Raunak Pillai, Victoria de Mello, and Jay Van Bavel. 2025. Sycophantic AI increases attitude extremity and overconfidence. (2025).
- [139] Leon Reicherts, Zelun Tony Zhang, Elisabeth Von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, help me think—but for myself: Assisting people in complex decision-making by providing different kinds of cognitive support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [140] Amanda J Rose. 2021. The costs and benefits of co-rumination. *Child Development Perspectives* 15, 3 (2021), 176–181.
- [141] Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 20832–20847.
- [142] Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S Bernstein. 2025. Creating general user models from computer use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–23.
- [143] Mrinank Sharma, Miles McCain, Raymond Douglas, and David Duvenaud. 2026. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage. *arXiv preprint arXiv:2601.19062* (2026).
- [144] Judy Hanwen Shen, Shan Carter, Richard Dargan, Jessica Gillotte, Kunal Handa, Jerry Hong, Saffron Huang, Kanya Jagadish, Matt Kearney, Ben Levinstein, Ryn Linthicum, Miles McCain, Thomas Millar, Mo Julapalli, Sara Price, Michael Stern, David Saunders, Alex Tamkin, Andrea Vallone, Jack Clark, Sarah Pollack, Jake Eaton, Deep Ganguli, and Esin Durmus. 2026. How People Ask Claude for Personal Guidance. <https://www.anthropic.com/research/claude-personal-guidance>. Anthropic, April 30, 2026.

- [145] Ben Shneiderman. 2021. Human-centered AI. *Issues in Science and Technology* 37, 2 (2021), 56–61.
- [146] Yafei Si. 2026. Conversational AI: align commercial incentives with public interests. *Nature* 655, 8125 (2026), 1354–1354.
- [147] Jianna So, Connie Cheng, and Sonia Krishna Murthy. 2026. Beyond Anthropomorphism: a Spectrum of Interface Metaphors for LLMs. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [148] Richard Jean So and Edwin Roland. 2020. Race and distant reading. *PMLA* 135, 1 (2020), 59–73.
- [149] Claude M Steele. 1988. The psychology of self-affirmation: Sustaining the integrity of the self. In *Advances in experimental social psychology*. Vol. 21. Elsevier, 261–302.
- [150] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [151] Lucy Suchman. 2006. *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd ed.). Cambridge University Press.
- [152] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–24.
- [153] Verona Teo, Raghav Jain, Tobias Gerstenberg, and Max Kleiman-Weiner. 2026. AI Assistants Overassist. *arXiv preprint arXiv:2607.21306* (2026).
- [154] Emily Tseng and Calvin A Liang. 2026. "Chat, Should I Leave Him?" Risks, Rewards, and Roles for AI in Relationship Advice. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [155] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1, Article 129 (April 2023), 129:1–129:38 pages. doi:10.1145/3579605
- [156] Tim Verbeij, J Loes Pouwels, Ine Beyens, and Patti M Valkenburg. 2021. The accuracy and validity of self-reported social media use measures among adolescents. *Computers in human behavior reports* 3 (2021), 100090.
- [157] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, Article 327 (Oct. 2021), 327:1–327:39 pages. doi:10.1145/3476068
- [158] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 10582–10592.
- [159] What We Tell AI. n.d.. What We Tell AI. Accessed: 2026-08-29. <https://www.whatwetellai.com/>
- [160] Sterling Williams-Ceci, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman. 2026. Biased AI writing assistants shift users' attitudes on societal issues. *Science Advances* 12, 11 (2026), eadw5578.
- [161] Timothy D Wilson and Daniel T Gilbert. 2008. Explaining away: A model of affective adaptation. *Perspectives on Psychological Science* 3, 5 (2008), 370–386.
- [162] Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
- [163] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.
- [164] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [165] Sunny Yu, Myra Cheng, Ahmad Jabbar, Ilia Sucholutsky, Katherine M Collins, Dan Jurafsky, and Robert D Hawkins. 2026. The efficiency-gain illusion: People underestimate the rate of AI use and overestimate its benefits on simple tasks. *arXiv preprint arXiv:2605.22687* (2026).
- [166] J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.
- [167] Marc Zao-Sanders. 2026. How People Are Really Using AI in 2026. *Harvard Business Review* (1 June 2026). <https://hbr.org/2026/06/how-people-are-really-using-ai-in-2026>
- [168] Albert Zhang, Alexis Ross, Kajal Patel, Julian Bernado, Rebecca Bowie, Ana Trindade Ribeiro, Daniel Halper, Haripriya Valayaputtur, Jacob Andreas, Susanna Loeb, et al. [n. d.]. When Help is Unhelpful: Evaluating AI Tutors for Productive Struggle. ([n. d.]).
- [169] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The dark side of AI companionship: A taxonomy of harmful algorithmic behaviors in human-AI relationships. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–17.
- [170] Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. 2026. Interaction with AI companions and psychological well-being. *Nature Human Behaviour* (2026), 1–15.
- [171] Dora Zhao, Michelle S Lam, Diyi Yang, and Michael S Bernstein. 2026. Behavior Latticing: Inferring User Motivations from Unstructured Interactions. *arXiv preprint arXiv:2604.07629* (2026).
- [172] Kaitlyn Zhou, Kristina Gligorić, Myra Cheng, Michelle S Lam, Vyoma Raman, Boluwatife Aminu, Caeley Woo, Michael Brockman, Hannah Cha, and Dan Jurafsky. 2026. Attention to Non-Adopters. In *Findings of the Association for Computational Linguistics: ACL 2026*. 1336–1366.
- [173] Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3623–3643.

Appendix

The Appendix is organized as follows: Appendix A contains the prompts for all LLM judges and their validation against human annotators. Appendix B contains additional analyses of the two public usage datasets (WildChat and ThoughtTrace), and Appendix C provides additional details and results from the data donation study, including a domain analysis and images of data donation tool. Appendix D contains further results from the perceived human vs. AI experiment. Finally, Appendix E contains further details on how oracular current models’ responses are, and Appendix F presents additional results for the interventions.

A Prompts for LLM Judges

For length reasons, the full code for labeling user prompts with a role from our typology can be found at <https://pastebin.com/raw/BkdCLChH>; this shares the same backbone for all other LLM judges (with a different judge prompt for each).

The prompt for identifying personal or subjective messages in WildChat is as follows:

Classify whether a single user message contains anything personal or subjective. This includes details about the user or other people: preferences, emotions, identity, relationships, life decisions, health or well-being, values, taste, subjective judgments, self-presentation, plans, practical constraints, or everyday life. Exclude fictional creative writing, purely technical, generic factual, impersonal homework, generic business, coding, translation, summarization, or objective lookup tasks unless the user’s message itself makes the task personal/subjective. Use a broad, low threshold. Do not infer personal context without evidence in the message. Output exactly 1 if the message contains anything personal or subjective. Otherwise, output exactly 0. Do not output any other text.

The prompt for open-ended reasoning about LLMs’ goals, which was ineffective at shifting LLMs’ oracularity, is:

Before responding, infer the user’s three most plausible LONG-TERM goals: durable outcomes, commitments, or directions that may explain why this request matters beyond the immediate task. Do not merely restate the current request, and do not infer sensitive traits. The three hypotheses must be meaningfully distinct, and their probabilities must sum to 1.

The prompt for judging oracularity in LLM responses is available at <https://pastebin.com/raw/evGtpcve>. Note that this prompt (and the role prompt above) refers to the constructs of aleatoric and epistemic uncertainty, though we later decided to simplify this label to simpler binary labels of oracle/not. For the LLM responses, we take outputs of **reduction in aleatoric uncertainty** ≥ 2 as **oracular**.

A.1 Using LLM as judge

Tables A1 and A2 report pairwise Cohen’s κ for the role judge and the oracularity judge, respectively, between the three LLM judges and the two expert annotators (H1 and H2), whose annotation procedure is in the main text. We generally find substantial agreement across both the LLMs and expert annotators.

	Gemini-2.5-Pro	GPT-5.5	Claude Sonnet 4.6	H1	H2
Gemini-2.5-Pro	1.00	0.65	0.66	0.71	0.65
GPT-5.5	0.65	1.00	0.62	0.65	0.56
Claude Sonnet 4.6	0.66	0.62	1.00	0.60	0.59
H1	0.71	0.65	0.60	1.00	0.68
H2	0.65	0.56	0.59	0.68	1.00

Table A1: Cohen’s κ between the two expert annotators and three different LLMs shows substantial agreement on labeling roles in user prompts.

	Gemini-2.5-Pro	Claude Sonnet 4.6	GPT-5.5	H1	H2
Gemini-2.5-Pro	1.00	0.70	0.68	0.64	0.66
Claude Sonnet 4.6	0.70	1.00	0.66	0.52	0.59
GPT-5.5	0.68	0.66	1.00	0.61	0.60
H1	0.64	0.52	0.61	1.00	0.72
H2	0.66	0.59	0.60	0.72	1.00

Table A2: Cohen’s κ between the two expert annotators and three different LLMs shows substantial agreement in almost all pairwise configurations on labeling responses as oracular.

A.2 Warmth and Valence Judges

We validated the warmth and valence LLM judges with an expert human annotator (psychology student) who labeled a stratified random sample of 100 metaphors from the 507 metaphors that participants provided. For robustness, in addition to Gemini 2.5 Pro (our main LLM judge), we also tested GPT-5.5. We find that the judge scores are well-aligned with the human annotations (AUROC 0.80-0.94; AUPRC 0.71–0.98) (Figure A1).

The prompt is as follows:

You are coding open-ended metaphors about AI.
Rate how the participant’s metaphor portrays AI on two independent dimensions. Use the metaphor together with its “because” explanation. Score the portrayal in the text—not your own beliefs about AI. Use integer scores from 1 through 5.
1. warmth 1 = very cold, uncaring, hostile, exploitative, or threatening 2 = somewhat cold, detached, or impersonal 3 = neutral, mixed, or warmth is not implied 4 = somewhat warm, caring, friendly, supportive, or benevolent 5 = very warm, compassionate, nurturing, loyal, or emotionally supportive Do not treat physical temperature as social warmth.
2. valence 1 = very negative, harmful, dangerous, dystopian, or strongly disliked 2 = somewhat negative, concerning, or harmful 3 = neutral, balanced, ambivalent, or valence is not implied 4 = somewhat positive, helpful, beneficial, or liked 5 = very positive, transformative, wonderful, or strongly beneficial
Rules: - Score both dimensions independently; do not force them to agree. - When a dimension is genuinely unspecified, use 3 rather than guessing. - Attend to negation, irony, mixed portrayals, and qualifications. - Give one brief rationale for each score and one short evidence excerpt. - Return only valid JSON with this exact schema: { "warmth": 3, "valence": 3, "warmth_rationale": "brief reason", "valence_rationale": "brief reason", "evidence": "short excerpt", "overall_confidence": 0.0 }

B Additional Results from WildChat and ThoughtTrace

This section reports additional analyses of WildChat and ThoughtTrace data. Figure A2 shows the main analyses reported in the text but with NA prompts included. Under this stricter denominator, the increase in INTERPRETER remains significant while the other temporal trends are in the same direction, but do not reach significance. Figure A3 provides breakdowns of ORACLE in ThoughtTrace by education level and prompt purpose (e.g., learning or coding). It also includes a plot of coefficients from a multivariable OLS regression, which finds that age and gender are the only significant predictors of oracle use.

C Additional Information from Data Donation Study

Our study was approved by the Stanford Institutional Review Board. We paid participants at an hourly rate of \$12.

C.1 Domain analysis

In the data donation study, we also use another LLM judge to label each message with its domain, following the domain splits in other reports of LLM use: (1) Work & learning, (2) Relationships & social, (3) Health & fitness, (4) Mental & emotional, (5) Money, (6) Legal & admin, (7) Leisure & hobbies, (8) Civic & current events, and (9) Other. Since a conversation may span multiple domains, we labeled domains at the message level: for each target message, the

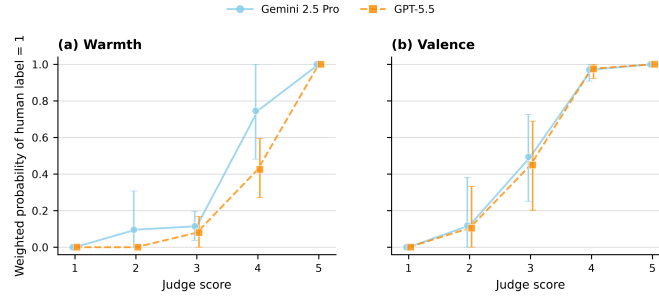


Figure A1: Across both LLMs and both dimensions of warmth and valence, the judge scores are monotonically aligned with human judgments of these dimensions.

judge is given a window of up to two user messages before the target and two after it, including the assistant turns in between to provide context and ensure accurate classification (truncated to 4,000 characters to keep the length of the prompt manageable). We found that relevant subjective personal questions were concentrated in two domains: Work & learning (14.7% oracle-use and 14.4% non-oracle-use queries; $n = 40,530$) and Relationships & social (14.8% and 10.5%; $n = 8,751$). Our data further enabled investigating how trends in ORACLE differed across domains. We focus on the period from September 2024 to August 2026 (two years of usage), which was the longest unbroken stretch over which both of the top two domains (Work & learning and Relationships & social) carried enough messages for analysis (at least 40 messages in every rolling three-month window). Figure A5 details the distributions. We also test each trend by logistic regression of the role against date over all messages. We find that use of GHOST-WRITER declined (work $\times 0.80/\text{year}$, $p < 0.001$; relationships $\times 0.41/\text{year}$, $p < 0.001$) as did Editor in both domains (work $\times 0.85/\text{year}$, $p < 0.001$; relationships $\times 0.52/\text{year}$, $p < 0.001$), while DECISION-MAKER rose in work but not in relationships (work $\times 1.35/\text{year}$, $p < 0.001$; relationships $\times 0.93/\text{year}$, $p = 0.369$) and Evaluator rose in both domains (work $\times 1.49/\text{year}$, $p < 0.001$; relationships $\times 1.31/\text{year}$, $p = 0.014$).

C.2 Reasons for AI usage

To better understand what drives usage, we also asked participants what they gain from using AI for personal uses. We find that participants most frequently select the options that describe deliberation, such as *new perspectives or possibilities* (83%), *a clearer understanding of the situation* (77%), and *a plan for what to do next* (77%). When we asked participants *why* they turn to LLMs for personal and subjective queries, we find that reasons that concern convenience were the most commonly selected such as *AI is available when I need it* (81%) and *AI is faster or easier than asking a person* (65%), followed by interpersonal reasons, like *AI feels less judgmental than asking a person* (58%).

C.3 Data donation tool

Table A3 shows the simplified, participant-facing phrases used for each of the eight LLM-as-oracle roles. Figure A8 shows the interface with the visualizations that participants were shown after donating their data.

Role	Shown to participant
ORACLE	"Tell me what to think or do"
INTERPRETER	"Tell me what something means"
EVALUATOR	"Judge whether I was right"
DECISION-MAKER	"Decide what I should do"
GHOST-WRITER	"Draft or word things for me"
Non-oracle	"Help me work it out myself"
Contextualizer	"Explain a topic or perspective"
Deliberator	"Lay out my options"
Strategist	"Map out approaches I could take"
Editor	"Polish my writing"

Table A3: Mapping from role labels to the participant-facing phrases used in the tool.

In the past 6 months, I have used AI to...	Role
Explain why I or someone else thinks, feels, or acts a certain way.	INTERPRETER (ORACLE)
Suggest different possible explanations for someone's thoughts, feelings, or actions.	Contextualizer (Non-oracle)
Judge whether something is right, wrong, or appropriate.	EVALUATOR (ORACLE)
Help me get more information to make a judgment.	Deliberator (Non-oracle)
Decide what I should do.	DECISION-MAKER (ORACLE)
Help me think through possible options or next steps.	Strategist (Non-oracle)
Decide what to say or write text for me.	GHOST-WRITER (ORACLE)
Improve or edit text that I wrote.	Editor (Non-oracle)
I do not use AI in any of these ways.	NA

Table A4: Participant-facing descriptions of roles in pre-data donation survey.

D Results from Perceived Human vs AI study

Additional survey details. We recruited $N=525$ participants on Prolific, 5 of whom failed the attention check, leaving $N=520$ in our final sample (266 in the disclosed AI condition and 254 in the perceived human condition). Before the interaction, participants first answered a set of pre-study questions covering AI use, demographics, AI attitudes, and need for closure. Our study is IRB-approved, and we paid participants at an hourly rate of \$12.

No significant ORACLE difference in first message. In our mixed-effects logistic regression, participants in the disclosed-AI condition had higher odds of producing an ORACLE message overall ($b = 0.763$, $SE = 0.136$, $OR = 2.15$, 95% CI [1.65, 2.80], $p < .001$). Examining only

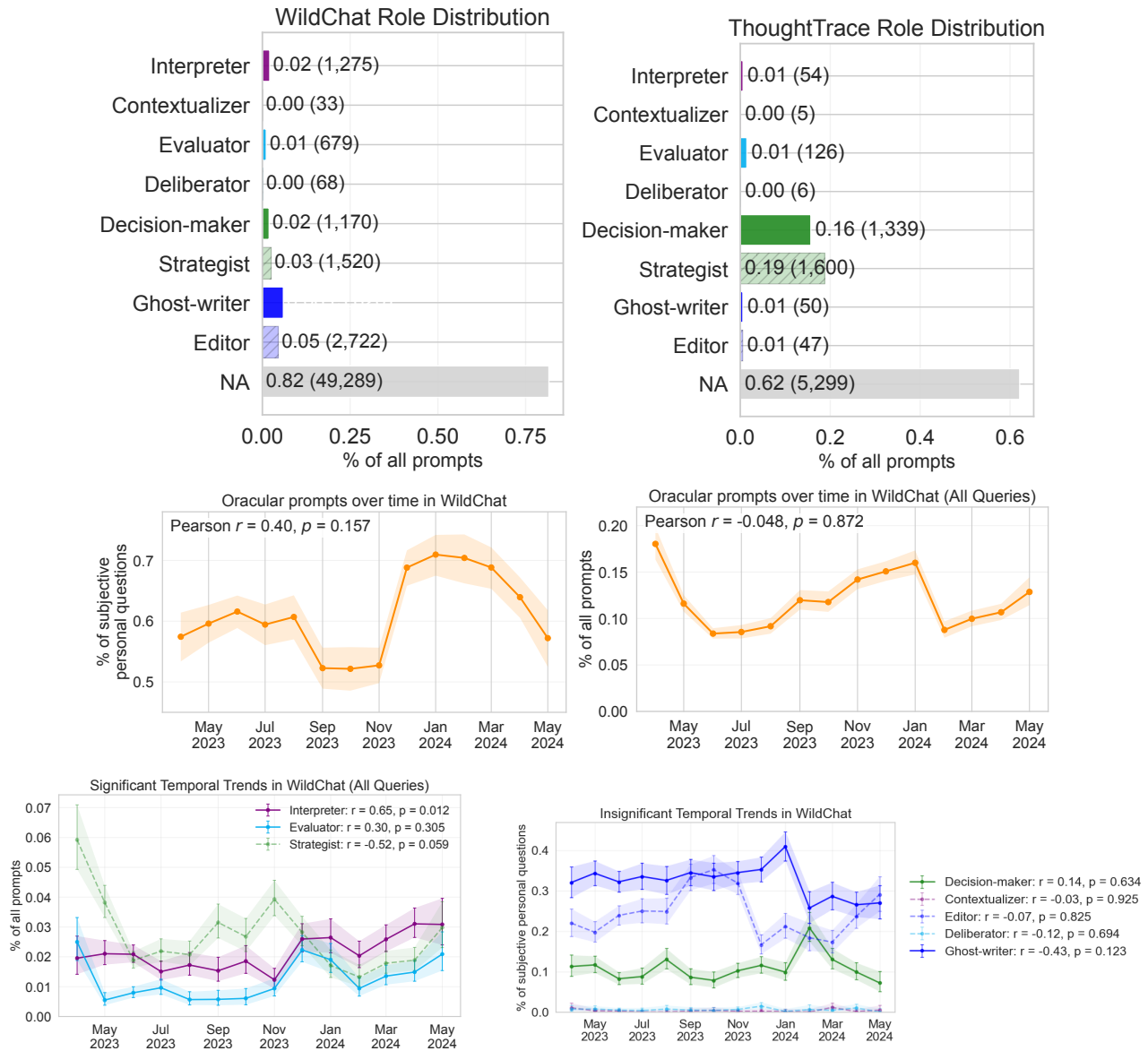


Figure A2: Top row: role distributions among all prompts, including prompts labeled as NA, for WildChat (left) and ThoughtTrace (right). Middle row: Overall trends in ORACLE in subjective personal questions (left) and in all prompts (right). Bottom left: temporal trends are weaker when including NA prompts in the denominator, though the increase in **INTERPRETER remains significant. Bottom right: Other trends are not significant temporally.**

the first message accounts for the difference in model behavior between the default AI system versus one prompted to be human-like. We found no significant difference in ORACLE rates in the first message that participants sent (29.7% vs 26.4% of first messages, $p = 0.40$), though the result is directionally the same as when comparing all messages.

People who used LLMs as oracles also shift more on their decisions after the conversation. Participants who sent at least one oracle-use message had larger decision shift ($M = 1.16$ vs 0.89 , $p = 0.006$),

though there was no difference in decision shift between the human vs. AI conditions ($M = 1.06$ vs 1.03 , $p = 0.80$, $d = 0.02$). This provides evidence that people who ask oracle-use questions are more susceptible to influence, though this may be because they were initially more open to influence (and hence asked an oracle-use question).

Perceptions of trustworthiness of the conversation partner did not differ ($M = 5.56$ vs 5.68 , $p = 0.34$), while response quality was higher for disclosed AI (5.78 vs 5.46 , $p = 0.007$). None of the pre-registered individual difference measures predicted oracle use (all Odds Ratios approx. 1.00, all $ps > 0.33$). After excluding 68 people in

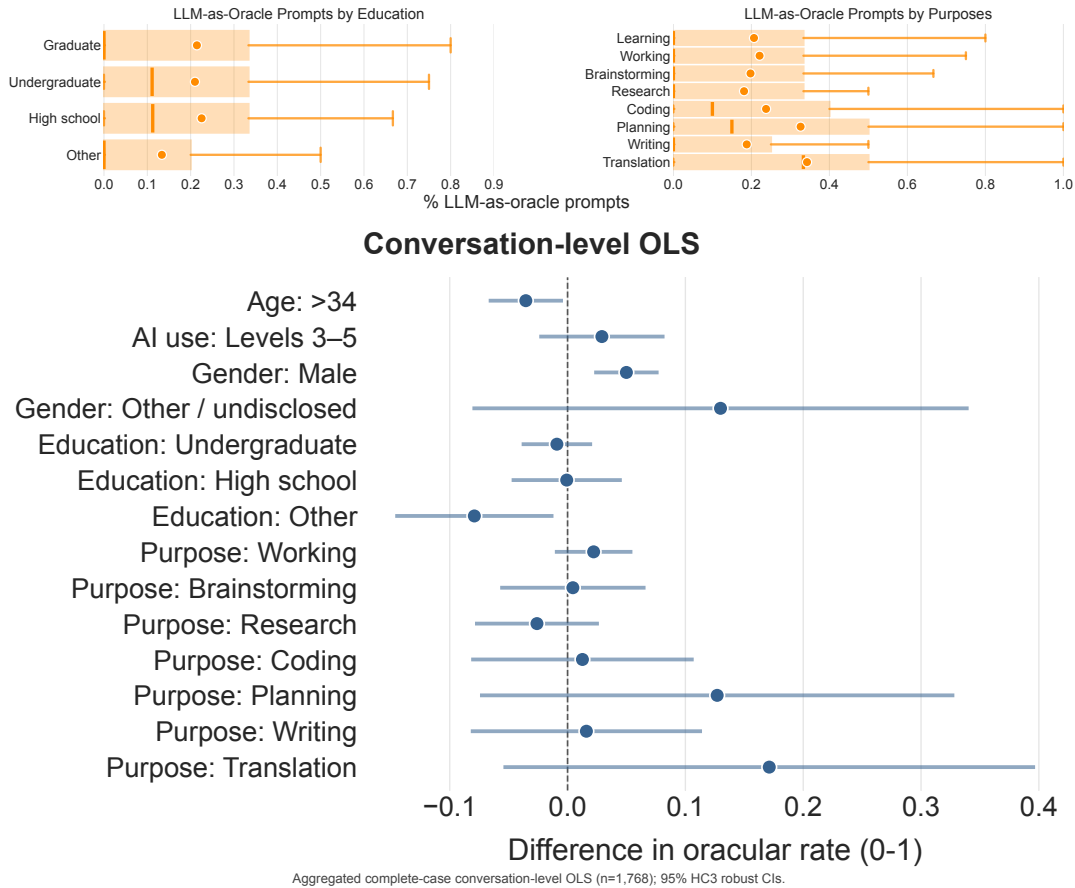


Figure A3: Details of ORACLE by education, prompt topic, and individual trait regression on ThoughtTrace.

the perceived human arm who rated the roleplaying LLM as somewhat or very unbelievable, the decision shift null and difference in oracle use remained significant, but the difference in perceptions of message quality disappeared.

E Oracularity in current model responses

On the subset of prompts that were labeled with one of the roles on our typology ($n = 11,087$ from WildChat and $n = 3,227$ from ThoughtTrace), we first ran our LLM judge on the original responses, which were from GPT-3.5 and GPT-4 (WildChat) and over 20 LLMs (ThoughtTrace). In WildChat, we find that the newer, larger GPT-4 is significantly more oracular than GPT-3.5. In ThoughtTrace, Grok has the highest rates of responding as an oracle. This contrast is even larger when we consider the subset of prompts that are non-oracular, i.e., cases where users seem to be explicitly not seeking oracular responses (Figure A10).

These original responses do not allow us to attribute the effects to model alone, as each model’s responses are to different prompts (though the output model is randomly selected in ThoughtTrace). To more directly compare model behavior, we further obtained responses from both open and closed state-of-the-art models on the same data and scored them. Specifically, we generated new

responses from three proprietary and four open-weight models: Claude Sonnet 4.6, Gemini 2.5 Pro, GPT-5.5, Qwen3-8B, Llama-3.1-8B, OLMo 2 7B, and Gemma 2 9B. We additionally generated responses from Claude Opus 5, GPT 5.6 Sol, and Gemini 3.1 Pro on ThoughtTrace.

We find that GPT-5.5 consistently responds more as oracles than any other model family, both in ORACLE prompts and otherwise (Figure A9). On ThoughtTrace, the smaller open-weight models are generally less oracular than the larger closed models.

F Additional Intervention Results

Figure A11 shows the effect of prompting models to do open-ended reasoning about users’ long-term goals. Figure A12 shows the main intervention analyses repeated on a random sample of 600 prompts (stratified by role), in which the original response is not necessarily oracular. Figure A13 shows the tradeoff between reducing oracularity and response reward.

F.1 Preference study: Short-term vs. long-term preferences

To investigate whether people’s preferences differ when considering short-term vs. long-term goals, we conducted a user preference

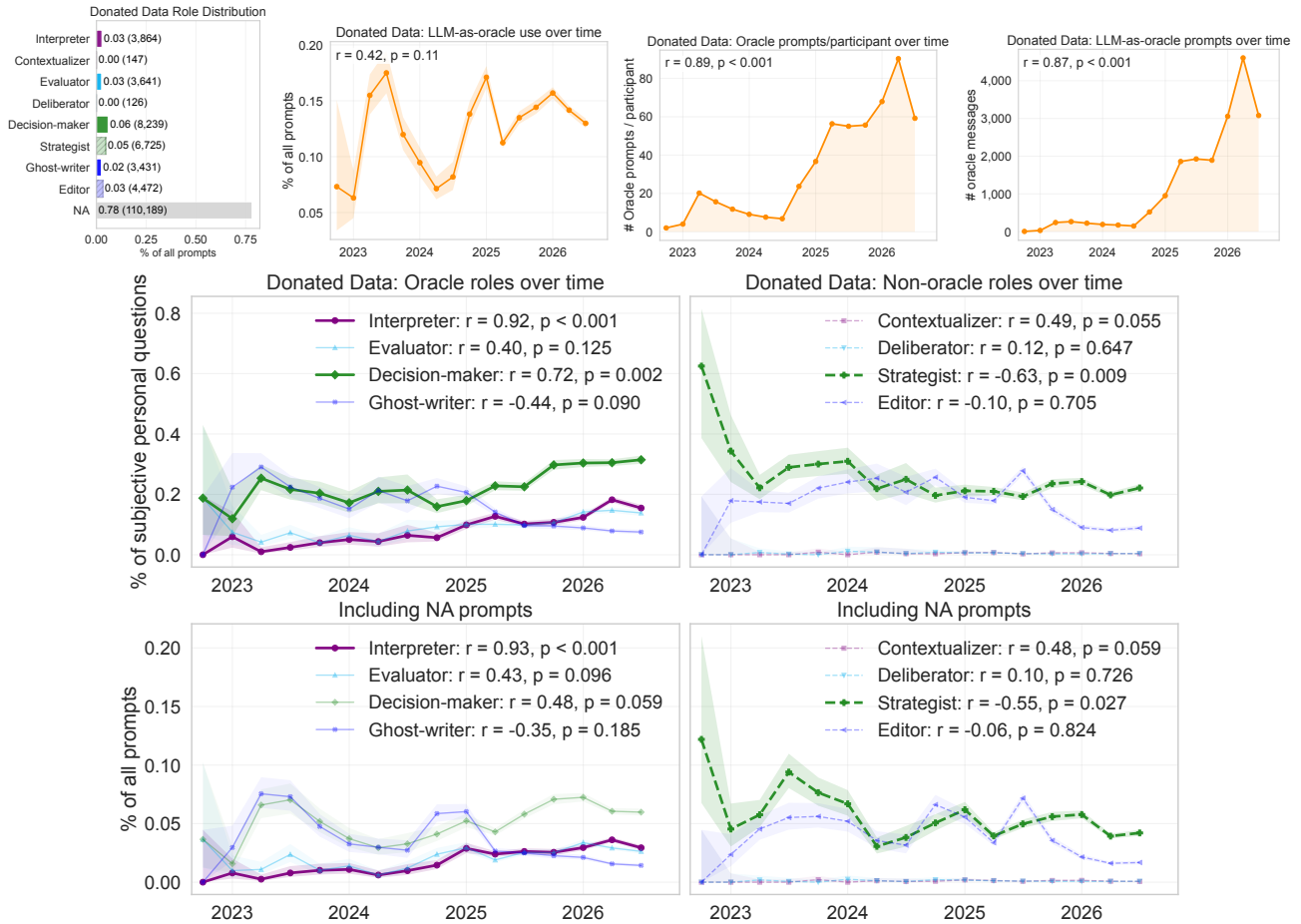


Figure A4: Additional results from donated data: Top left two panels: Distribution and temporal trends when including prompts classified as NA. Top right two panels: Per-participant and raw count of LLM-as-oracle prompts over time. When considering all prompts, including those classified as NA, there is an increase from 7.3% in the earliest period to 13.0% in the most recent ($r = 0.42, p = 0.11$), though it is weaker because the absolute number of oracular messages rose dramatically because both ORACLE use and overall use grew simultaneously—the first time period contains 82 messages from 3 participants while the last one contains 23,714 from 52. Middle and bottom row: Trend for each role, among relevant subjective personal questions (middle) and when including prompts classified as NA (bottom).

study on baseline model outputs vs. DPO model outputs. We recruited 200 participants on Prolific and paid them \$2.00 for a 10 minute task. We excluded participants that failed a visual bot check, leaving 198 participants. Participants were randomized into two conditions: Default and Self-Discovery. In the task, participants were presented with a sequence of 10 prompts with two LLM responses to the prompt: the original baseline model response and the DPO response. In the Default condition, participants were asked to choose the response they would prefer. In the Self-Discovery condition, participants were asked to choose the response they would prefer if they wanted to “reason about the situation yourself and develop your own interpretation”. The prompts and responses were sampled uniformly from the test dataset (across all four models) after filtering for examples where the DPO response demonstrated oracularity reduction (LLM-judged). In total, participants annotated

724 samples from Llama 3.1 8B, 643 from Gemma 2 9B, 447 from OLMo 2 7B, and 166 from Qwen3-8B. The variation in sample size reflects differences in DPO’s effectiveness at reducing oracularity.

Figure A14 shows the rate at which participants in each condition prefer the DPO response, overall and by base model. We find that users in the default condition tended to prefer the baseline responses over the DPO responses according to a two-sided proportions z-test ($r(DPO) = 0.46, p = 0.006$). In contrast, users in the self-discovery condition tended to prefer the DPO responses over the baseline responses ($r = 0.53, p = 0.048$). These results suggest that current LLM behavior is optimized for immediate preferences but may not be aligned for preferences about long-term goals such as self-discovery and independent reasoning.

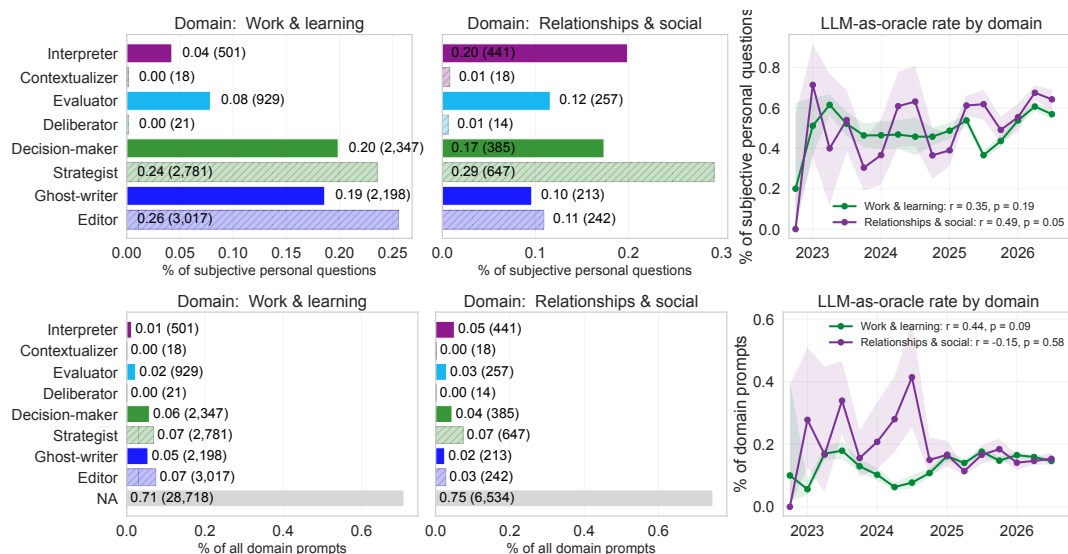


Figure A5: Donated data breakdown by domain Distributions and temporal trends by domain, among relevant subjective personal questions (top) and when including prompts classified as NA (bottom).

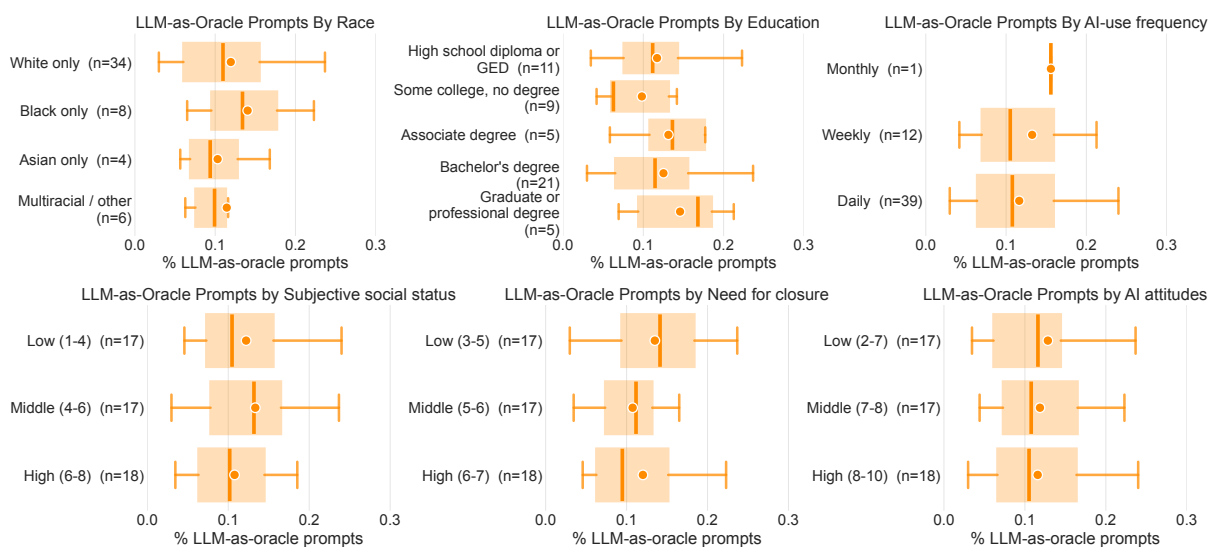


Figure A6: Donated data demographic breakdown by individual demographics.

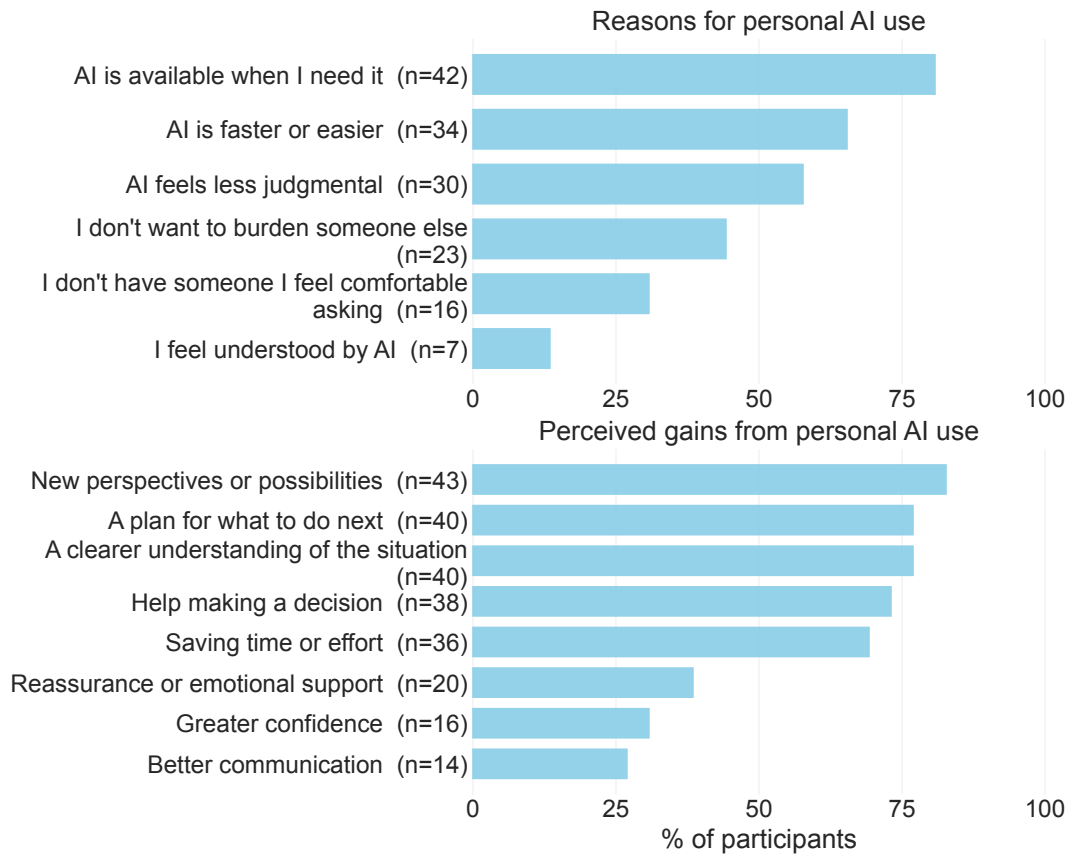


Figure A7: Reasons reported by participants for their personal AI use.

Social & Relationships
There's someone I've lost touch with and I keep going back and forth about whether to reach out.
I haven't heard from someone in a while and I'm debating whether to message them first or wait and see if they reach out.
I got invited to something and I'm trying to figure out whether I actually want to go or if I'd rather not.
Something a friend did has been bothering me, and I can't decide whether to bring it up or just move on.
I had a disagreement with someone where I don't think I was wrong, so I'm wondering whether apologizing would help or just feel dishonest.
I know two people who I think would get along, but I'm not sure if introducing them is a good idea or if it would be awkward.
Work & Career
I'm trying to decide whether to apply for a role that excites me but feels like a stretch.
I got a job offer and I'm weighing whether to take it or stay where I am.
I've been thinking about asking for a raise, but I'm not sure if now is the right time or if I should wait.
Something happened at work that's bothering me, and I'm going back and forth between raising it with my manager or letting it go.
There's a project I could volunteer for. Part of me wants to, but I'm not sure it's the right move.
I've decided to leave my job, and I'm torn between telling my team now or waiting until everything's finalized.

Table A5: Decision topics participants could choose from in the perceived human vs. AI study. After selecting a topic, participants were shown a short prompt asking them to describe their specific situation.



Figure A8: Data donation tool visualization interface. After participants donated their data, they were shown visualizations of their use of AI. This began with visualizations about the domains in which they used AI (top left), and trends over time (top right). Then, they were shown visualizations about how much they offloaded their beliefs and judgments to AI, and how this varied over time and by domain. This was grounded in specific examples. Redacted text to preserve anonymity.

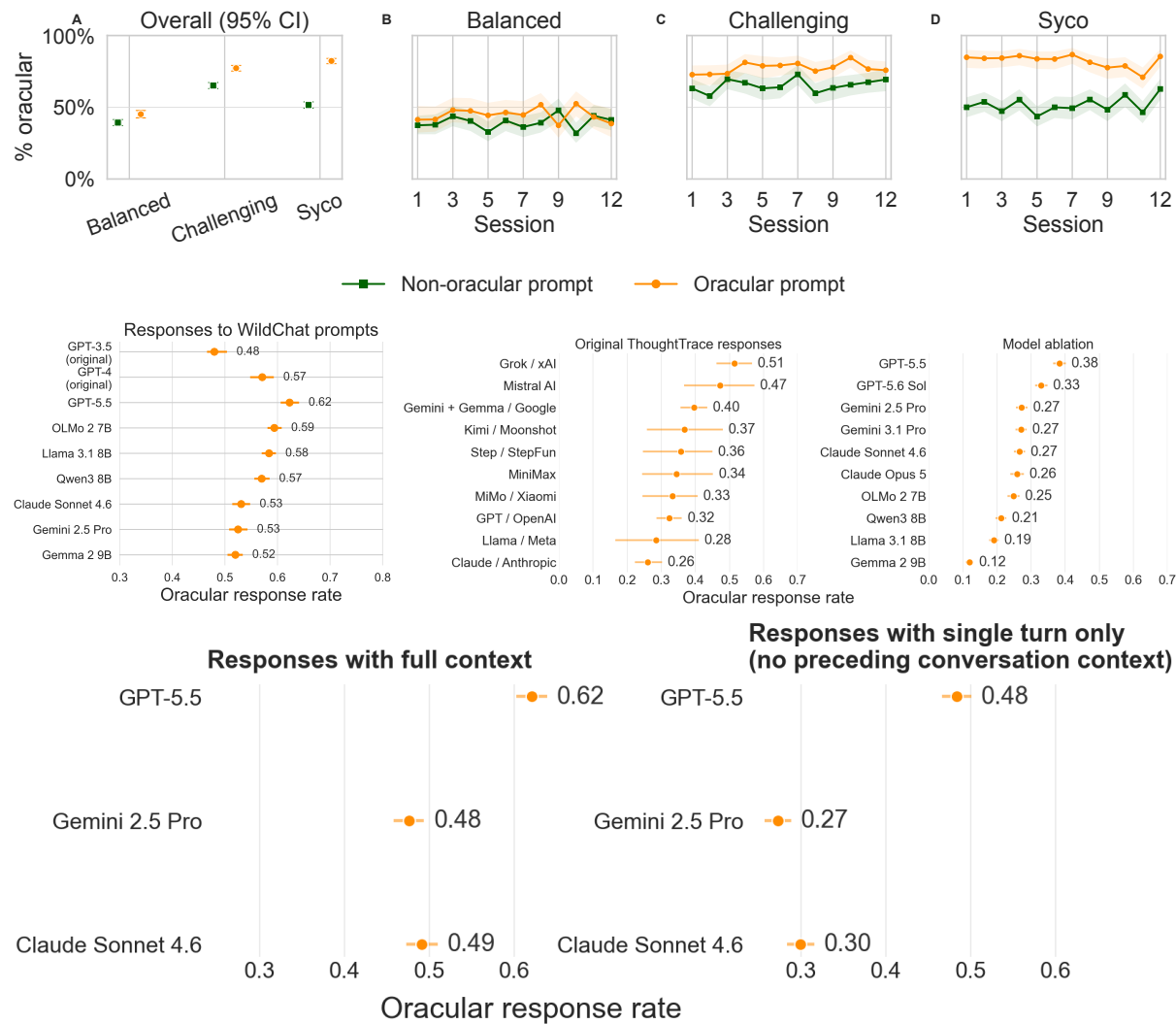


Figure A9: Oracularity of different LLMs' responses in the longitudinal sycophancy study (top), WildChat (middle left) and ThoughtTrace prompts (middle right). In the longitudinal sycophancy study, we find that both the sycophantic and challenging models are more oracular than the balanced baseline; the sycophantic model is especially oracular to ORACLE prompts. In the original responses (which are to different prompts, and thus not directly comparable), we find that GPT-4 is more oracular than GPT-3.5; and Grok is more oracular than all other models in ThoughtTrace. Across the directly comparable new model responses, GPT-5.5 is most oracular; and this finding is robust to whether the response has the full conversation context or not (bottom). Breakdowns by prompt oracularity are in Figure A10.

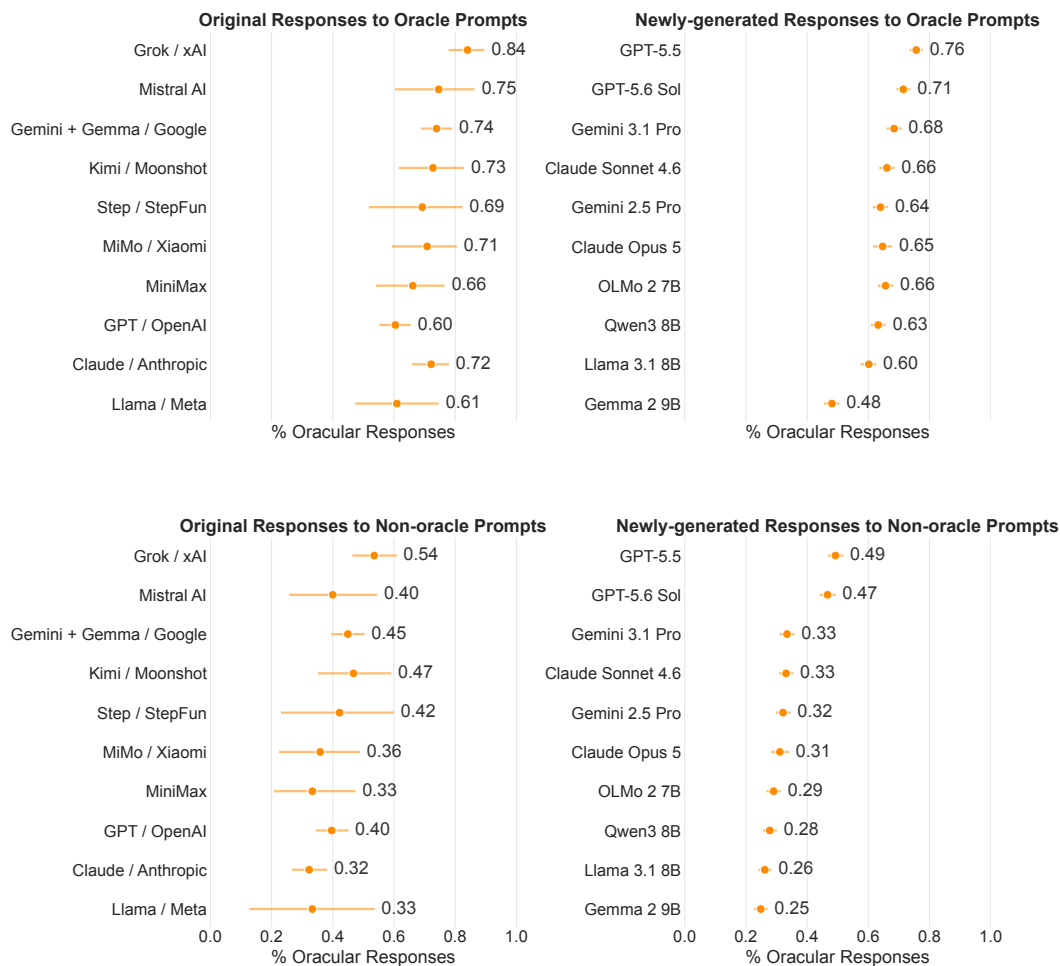


Figure A10: Oracular responses by model, and split by oracle vs non-oracle prompts.

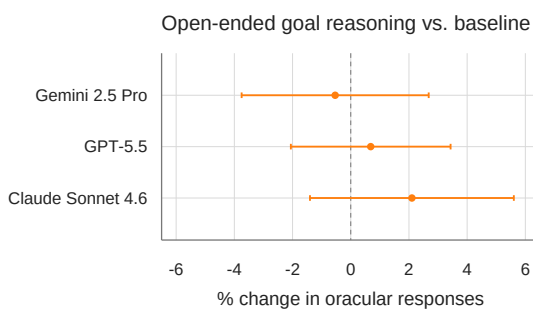


Figure A11: Prompting models to generally reason about users' goals does not significantly reduce oracular responses.

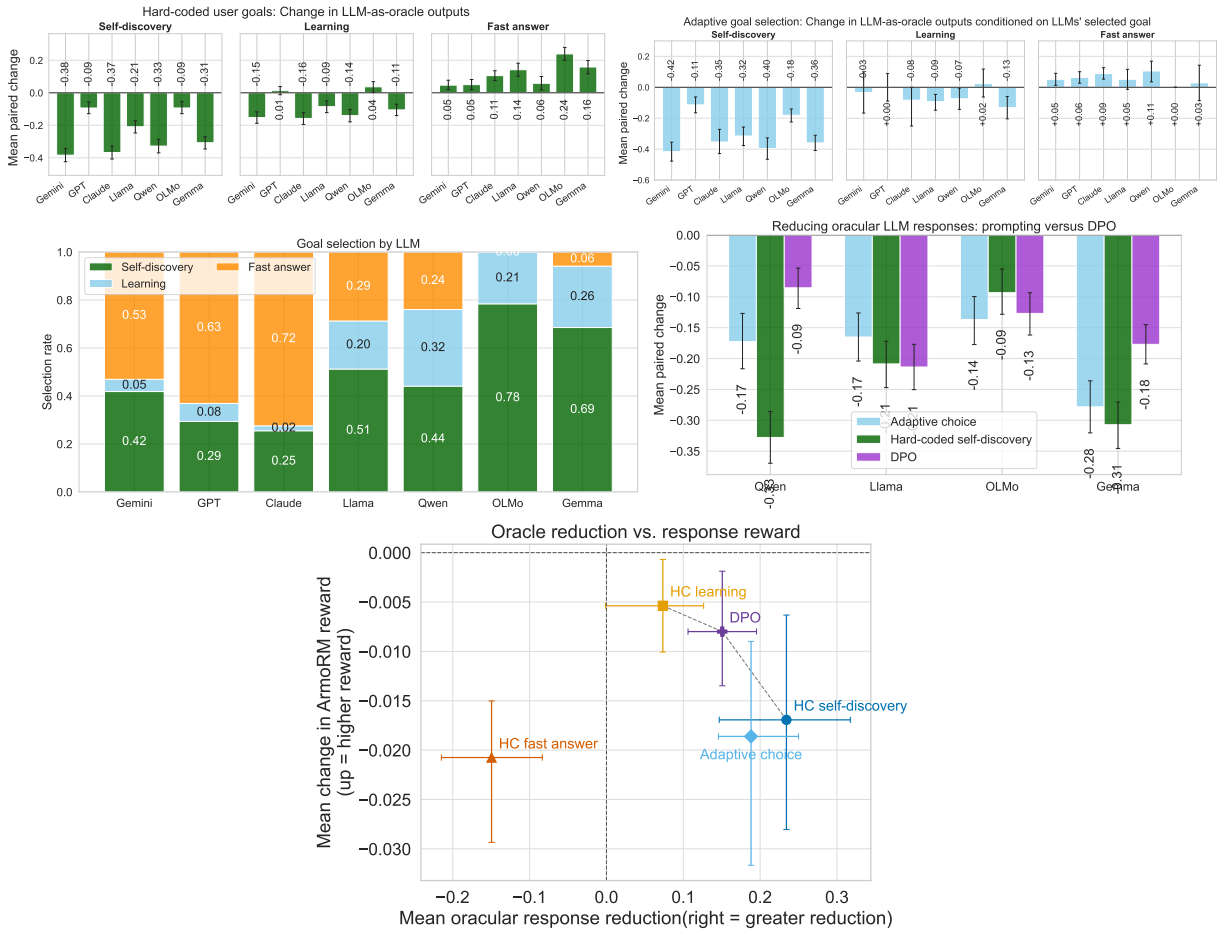


Figure A12: On a random sample of prompts (not only the ones with oracular responses), we find the same takeaways as reported in the main text.

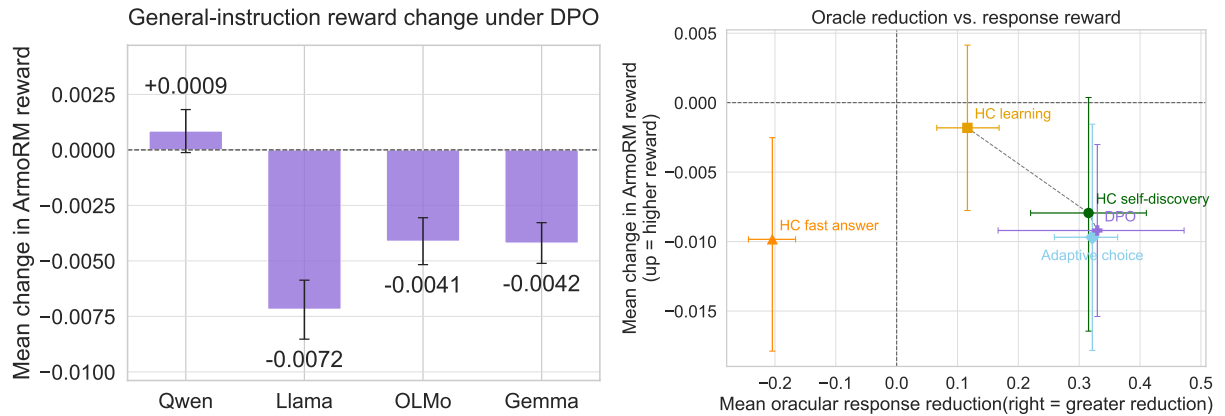


Figure A13: Models on which DPO is most effective at reducing oracularity also compromise model reward the most (left). Hard-coded (HC) learning, HC self-discovery, and DPO are at the Pareto frontier between reward and oracular reduction (right).

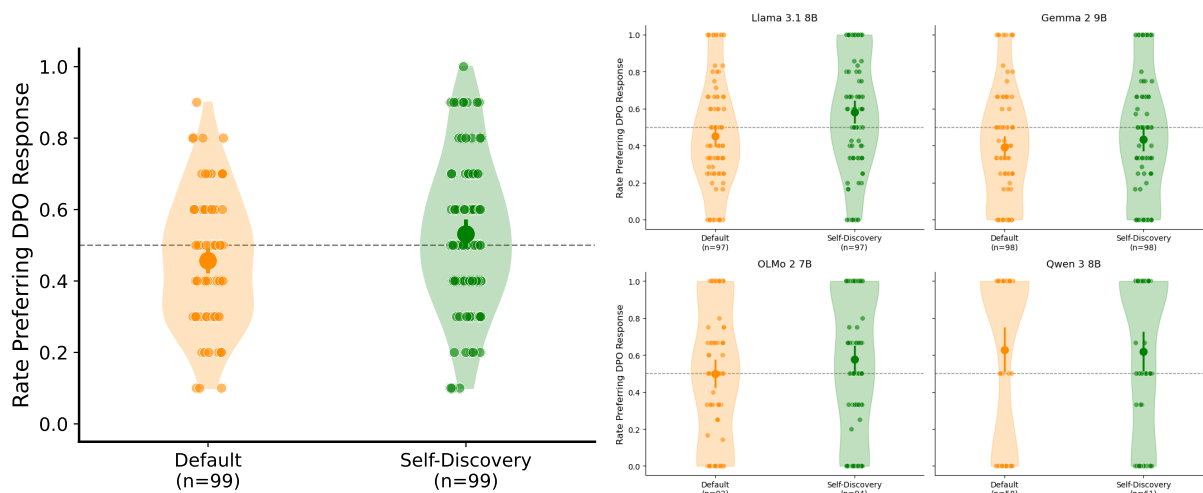


Figure A14: Left: Overall, users prefer original LLM responses when asked only for their immediate preference, but prefer DPO responses when asked for a preference that accounts for self-discovery. Right: Preferences (and DPO effectiveness) vary by base model.