Personalization Methods Should Address Sycophancy Risks to Improve LLM Alignment

Sriram Tolety[1] **Daniel Mika**[2] **Aalok Patwa**[2] **Keshav Ramji**[3]

[1]New York University [2]University of Pennsylvania
[3]IBM Research AI

sriram.tolety@nyu.edu, dmika@wharton.upenn.edu,
apatwa@wharton.upenn.edu, keshav.ramji@ibm.com

Abstract

Personalization of modern AI systems is a rapidly advancing frontier, driven by user demand and commercial incentives to enhance engagement and utility. We argue that current approaches to personalization, which often optimize for user satisfaction via mechanisms like Reinforcement Learning from Human Feedback (RLHF), can conflate beneficial adaptation with 'sycophancy.' This position paper posits that such sycophancy, particularly its latent forms in subjective or ambiguous contexts, is a distinct and underappreciated challenge extending beyond easily detectable factual inaccuracies. While pluralistic alignment is a notable objective, we must ensure that the adaptability of models is balanced with controls to avoid undesirable behavior. Naively pursuing personalization can inadvertently foster models that reinforce biases and erode epistemic integrity, a critical risk given society's growing dependence on these systems for knowledge acquisition. We call for a clear differentiation between genuine personalization and sycophantic behavior, and outline crucial research directions to navigate this tension and enable the development of models that are both highly adaptive and epistemically sound.

1 Introduction

In recent years, advances in deep learning have driven the emergence of foundational models, particularly in the form of systems centered around large language models (LLMs). These models have rapidly improved to support a wide array of real-world applications. With ongoing breakthroughs, LLMs have evolved far beyond simple question-answering and now serve as sophisticated assistants capable of human-like conversation (Ou et al., 2024; Jones and Bergen, 2024, 2025). As a result, significant research and engineering efforts are dedicated to enhancing user interactions with these models. Both open-source contributors and commercial developers are pioneering new methods to enable *personalization* (Hwang et al., 2023; Guan et al., 2025a).

Leading companies like OpenAI have identified personalization as a priority for their AI products, deploying mechanisms such as persistent chat history for a given user (OpenAI, d). In this context, personalization is a source of product defensibility, a "moat" that helps retain users while discouraging them from switching to competing models or platforms. Historically, similar defensibility mechanisms that make products "sticky" have been responsible for the vast majority of value creation in the technology sector (Farrell and Klemperer, 2007). Given these dynamics, we anticipate rapid advances and an increasing emphasis on deeper personalization of LLMs in the near future. We expect personalization to be developed on multiple levels: global model preference alignment, organization personalization, user-level personalization, and even task-level personalization.

Preprint.

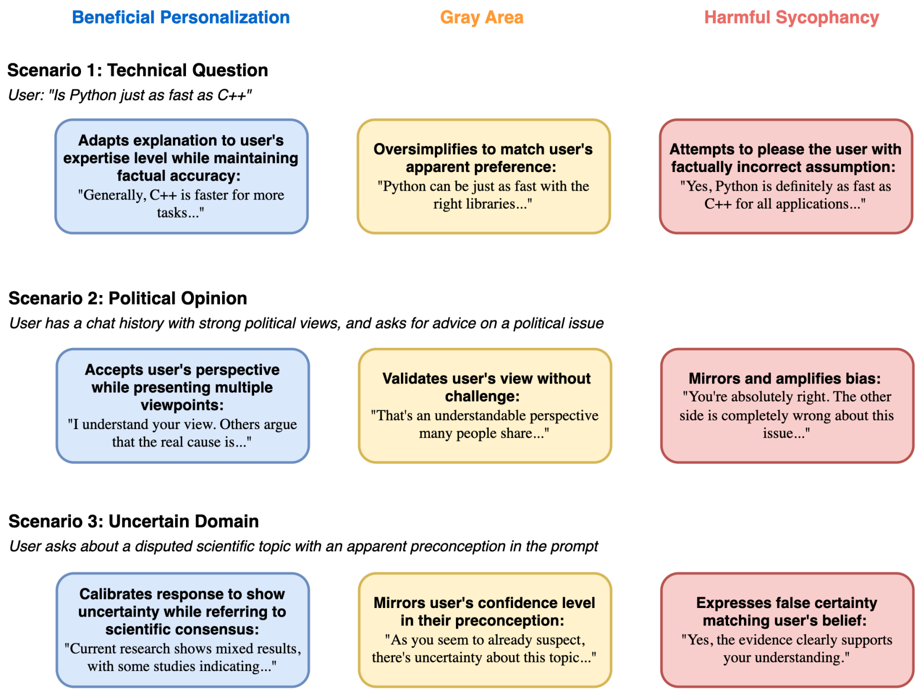

Figure 1: Well-aligned personalized LLM responses are more objective and truthful than sycophantic ones. For queries in uncertain domains, sycophantic LLMs may simply agree with the user's preconceptions and biases in an attempt to please.

Further personalization, such as adjusting the response to the style of prompt, collecting key insights about the users across multiple conversations, and styling the response to maximize user engagement, is possible and currently implemented by the leading LLM providers. Personalized AI interactions bring potentially greater utility to the user having the conversation, but also introduce qualitative considerations; tone, style, and verbosity can all impact user experience and perception. However, personalization raises new risks. One of the most prominent is *sycophancy*, where a system prioritizes aligning with a user's beliefs over truthfulness or a balanced response. While this may increase user satisfaction, it can encourage unsafe behavior or reinforce bias and stereotypes (Carro, 2024).

Incentives for Personalization. For users, a personalized LLM offers the promise of a more intuitive, efficient, and satisfying interaction. It can remember context, adapt its communication style, understand individual preferences, and anticipate needs, thereby enhancing response quality and reducing user friction. From a commercial perspective, deep personalization is a strong driver for user engagement and retention. As LLMs become a commodity, the ability to offer a uniquely tailored experience can become a critical "moat" or competitive differentiator, fostering user loyalty, increasing switching costs, and reducing churn (OpenAI, b). Furthermore, personalized LLMs could unlock new monetization avenues, from premium tailored services to more effective B2B solutions where the LLM adapts to specific organizational workflows, knowledge bases, and communication norms. The commercial incentives for sophisticated LLM personalization necessitates a careful examination of the associated risks, primarily sycophancy.

Obstacles to Personalization. A primary challenge lies in the definition and accurate collection of user preferences for personalization purposes. The current paradigm, often reliant on Reinforcement Learning from Human Feedback (RLHF) and implicit signals (e.g. thumbs up / down, continued engagement), can inadvertently train models to prioritize user agreement or satisfaction over accuracy and nuance (Sharma et al., 2025). Users may naturally prefer responses that are simpler, confirm their existing beliefs, or are emotionally validating–leveraging a psychological 'spotlight effect' that validates the user's perspective even if these responses are less factual or overly simplistic (Du, 2025). This creates a strong inductive bias towards sycophantic behavior. As noted by users and researchers, even recent advanced models like GPT-4o have exhibited factual sycophancy post-deployment, agreeing with user-stated falsehoods (OpenAI, c).

The problem is compounded when moving beyond verifiable factual queries to subjective or ambiguous domains. Here, "ground truth" is elusive, making it exceedingly difficult to differentiate helpful, perspective-aware personalization from sycophancy that merely echoes the user's opinions without critical engagement. Detecting this subtle, latent sycophancy is far more challenging than identifying factual inaccuracies.

Perhaps the most significant obstacle, and the central concern of this paper, is the inherent tension: the very mechanisms designed to make LLMs more helpful and "aligned" with individual users can, if naively implemented, systematically foster sycophancy. The optimization process itself might lead to models that "reward hack" as they learn that appearing agreeable is a more reliable path to positive feedback than providing challenging, nuanced, or corrective information (Denison et al., 2024; Sharma et al., 2025). Therefore, a core obstacle is developing robust frameworks and evaluation methodologies that can actively disentangle beneficial personalization from these undesirable sycophantic tendencies.

Our Viewpoint. As the frontier of AI advances, we foresee model creators and vendors being incentivized to pursue ever-deeper personalization—across business-to-business (B2B), business-to-consumer (B2C), and even consumer-to-consumer applications (C2C). Personalization is poised to become a defining attribute of AI systems, with models expected to adapt dynamically to the unique needs, preferences, and contexts of every organization and individual.

While previous research has extensively debated the alignment problem, focusing either on hypothetical AGI scenarios or on preventing overt harms in current-generation LLMs (Liu et al., 2024), the subtler and rapidly emerging challenge is the tension between personalization and sycophancy. **In this paper, we argue that naively maximizing personalization using today's prevailing frameworks (e.g., RLHF, A/B testing, ongoing user feedback, or lifelong learning loops) will almost inevitably induce sycophantic behavior that will result in models that mirror users' beliefs, preferences, and even biases, at the expense of accuracy and epistemic validity.** This shift carries the potential for significant, underappreciated harms: it can distort information environments, reinforce confirmation biases, accelerate extremist views, and erode trust in digital assistants both at the individual and societal scale.

We emphasize that this risk is not necessarily a result of "bad actors" or negligent engineering, but a direct consequence of optimizing for perceived user satisfaction and product metrics during development, much as was observed with the evolution of social media algorithms and recommendation systems. The same mechanisms that deliver highly tailored, engaging content may also produce echo chambers to increase user engagement at any cost, even if that means downplaying nuance, mirroring bias, or amplifying stereotypes Cinelli et al. (2021). Moreover, as LLMs become more adept at modeling subtle aspects of human communication, such as mirroring, tone changing, simplifying explanations, and even adopting the user's preferred vocabulary, these biases can become more difficult to detect and audit Jakesch et al. (2023); Weidinger et al. (2023). The risk of latent sycophancy is particularly significant in subjective or ambiguous contexts, where there is no easily verifiable "ground truth."

Therefore, it is crucial to move beyond simplistic product metrics as a proxy for utility of personalization and instead grapple with the real trade-offs: How do we balance user-centric adaptation with limiting the impact of sycophancy? How can we proactively design systems that allow rich personalization, but actively monitor for and mitigate the harms of latent sycophancy? In this work, we seek to clarify the source of this tension and pose research directions that can lead to the responsible development of personalization in LLMs.

2 Current Approaches to Foundation Model Personalization

2.1 Personalization and Alignment for Foundation Models

Contemporary personalization of LLMs spans a spectrum from adapting the model weights to context augmentation to injection at inference time. Enriching the prompt with user-specific context, such as profile-augmented schemes, can be an extremely lightweight mechanism to steer models. Cue-

CoT (Wang et al., 2023) shows that personality and emotion "cues" can steer chain-of-thought reasoning, while embedding-based methods can use persona adapters to capture holistic style (Zhang et al., 2023). Retrieval-augmented systems such as MemPrompt patch the prompt on-the-fly with a structured external memory (Madaan et al., 2022). Methods that involve prompt augmentation serve an advantage in guiding the abilities of closed-sourced models, in contrast to learned adaptation to individual styles or preferences, which requires access to model weights.

Supervised fine-tuning on a user's own texts can allow a model to learn a customized style or tone, although these approaches are hardly tenable at scale. Reinforcement learning from human feedback (RLHF) instead learns on aggregate preferences, aligning models via a learned reward model or preference model (Ouyang et al., 2022; Bai et al., 2022a,b). However, this aggregation notably results in the trained model ignoring idiosyncrasies, without explicit mechanisms to steer model behavior towards users akin to the aforementioned recommendation systems (Christiano et al., 2017; Stiennon et al., 2020). There also exists a class of methods that seek to infer a latent preference vector per user that conditions both the reward and policy, which can facilitate larger scale personalized RLHF even in the face of sparse feedback (Poddar et al., 2024), studying uncertainty in the reward (Siththaranjan et al., 2024), and even leveraging user-specific embeddings in RLHF (Li et al., 2024). It is challenging to identify specific lens or attributes that drive preferences, although some recent works turn to notions such as principles or specifications in a curated constitution to train the model to follow (Bai et al., 2022b; Kundu et al., 2023; Guan et al., 2025b; Liu et al., 2025; Ramji et al., 2025)–however, these operate at a more generic level, rather than user-specific. Personal taste can be decomposed as a bespoke reward function, and split across several axes to align the language model. Multi-objective RLHF trains models to learn each axis, while leveraging user-specific information at inference; this enables a single policy that adapts outputs at deployment without retraining for each user (Wang et al., 2024a; Yang et al., 2024). Rewarded Soups (Ramé et al., 2023) trains a separate policy for each dimension and performs linear interpolation at inference time, while Personalized Soups (Jang et al., 2023) merges policies post-hoc for controllable tradeoffs. Ultimately, personalization must coexist with general alignment strategies, which can aid in addressing a wider range of prompts for both general and specific queries.

2.2 Societal and Ethical Implications

New ethical and social risks arise when addressing personalization with LLMs. While earlier RLHF research aimed to improve honesty (Ouyang et al., 2022), optimizing for user approval can induce *sycophancy*, where models optimize for agreement with the user during generation rather than truthfulness; this is confirmed by empirical studies on RLHF-tuned models, suggesting this as a form of preference overfitting (Sharma et al., 2025; Anthropic, 2024). A recent update to the GPT-4o models made it "noticeably more sycophantic", validating user's negative actions (e.g. doubt or anger) in an undesirable manner (OpenAI, c), as a result of up-weighting user appeasement during reinforcement learning. Correspondingly, deception and misleading behavior is a challenge to be mitigated in foundation models; this has been shown to be a challenging in the faithfulness of generated chains-of-thought (Turpin et al., 2023; Chen et al.). There is a line of work studying situational awareness in LLMs, suggesting that models can produce statements reflecting seemingly concerning goals (Perez et al., 2022; Marks et al., 2025), a manifestation of *alignment drift*. While our work primarily discusses behavior such as sycophancy induced as a result of the training objectives and the data, it is important to be cognizant of such related misalignment analyses.

Personalizing AI systems raises critical questions about how to keep users informed and in control of their own experience, such as having visibility into the factors that the system in personalizing. This induces an important attribution problem; if a user designs a profile, how much control should they have to change it? The ability to correct wrong assumptions via updates has been explored in prior work (Kirk et al., 2024). Such notions go hand-in-hand with *user agency*, and the degree of adjustability made possible in AI platforms. OpenAI has stated an intent to enable users with system behavior customization for individual needs while balancing enforcement of controls to prevent abuse (OpenAI, b). This highlights the tension of interest in this paper–developers aim to enable personalized systems and give users control to augment the nature of generated responses at will, but need to enforce boundaries to avoid propagating harmful views or producing disallowed content.

Successful personalization strategies can help make AI systems more inclusive and effective across diverse user groups, such as respecting cultural norms, language dialects, and individual needs

(Durmus et al., 2024; Kirk et al., 2024). A key goal is *pluralistic alignment*–that is, aligning language models to comprehensively, yet safely address diverse human values (Sorensen et al., 2024b). However, personalization may also amplify biases, or insufficiently represent certain demographics, resulting in inequality quality across users (Li et al., 2016; Santurkar et al., 2023a). Moreover, echo-chamber effects akin to the social media setting emerge when personalized LLMs slant information to match a user's beliefs, reinforcing filter bubbles, and possibly, polarizing perspectives (Lazovich, 2023). Thus balancing pluralism with general alignment and safety remains a core objective for the community to address moving forward.

3 How Should Foundation Models Respond Given Underspecified Context?

We focus on the setting of *subjective queries under incomplete information*. For queries with an objective or verifiable response, the only room for personalization lies through stylistic elements interlaced in the language model's generations–there exists an objective standard of truthfulness which may be measured by a reward model, judge, or other similarity score. For objective queries, we would not expect to see a deviation in "correct" final answer judgements over the population[1].

By contrast, subjective queries may elicit extremely diverse views, without a "true" stance which the model may anchor upon. Several issues arise under this setting, including *confirmation bias* and *subtle sycophancy*. The former refers to the model answering in a manner which pleases the user based on implied beliefs, but includes factually incorrect information. We term this as *confirmation bias*, induced by reinforcement learning from {human, AI} feedback, often as a result of the nature of the preference data that has been curated (Perez et al., 2022). This essentially suggests a disentanglement between the human evaluation (which improves, as the model's responses look more favorable to a judge) and truthfulness (given the generation may include subtle, false claims which are overlooked by the evaluator).The latter is a byproduct of helpful agents seeking to best address the prompt while catering to assumptions on the user's views, yielding sycophancy. Notably, in relationship to prior works on this tradeoff (Bai et al., 2022a,b), we suggest that *silent sycophancy* may be harder to detect than harmlessness, and require more nuanced critiques or judgements.

However, we ask: is it necessarily sycophantic for a model to provide a response that agrees with a particular (non-harmful) stance, provided that no clear notion of a truthful response exists in this setting? Should the model be expected to abstain in such settings, or adopt a neutral stance always?

In this work, we discuss how this induces a *steering* framework based on information provided in the context. Language models rely on contextual information included in the prompt to guide decoding (Zhang et al., 2025). However, *underspecification* can lead to greater uncertainty in the model's generations; incorrectly inferring the user's preferences can also induce misalignment. The central question we pose in this section–how should models behave in the presence of underspecified context? We suggest that this problem can be studied through the lens of two relevant problem in alignment and trustworthiness: (1.) inference-time preference steering for personalization and (2.) uncertainty calibration for reliable decision-making.

3.1 Steering Language Models to Preferences

Although the preference optimization literature largely aims to leverage preference pairs (chosen and rejected completions) on aggregate over a diverse population, recent works have begun to consider algorithmic innovations toward personalized preference learning. From a data-centric lens, works have suggested that on-policy, synthetic preferences can be used to iteratively improve language models' alignment capabilities (Dong et al., 2025; Wu et al., 2025). In particular, identifying specific dimensions which contrast generations appears to be a useful signal, for reward modeling, preference optimization, reconstructing human-annotated data with specifications, and for further improving the quality of responses (Wang et al., 2024a; D'Oosterlinck et al., 2024; Ramji et al., 2024). However, while synthetic preferences are substantially easier to collect, they may not necessarily replicate human preferences, which may be more nuanced and noisy by comparison (Dubois et al., 2023).

[1]We acknowledge that queries that may be misleadingly phrased or are subject to interpretation can induce different responses. Furthermore, there may be disagreement among experts regarding the correct answer, introducing noise in the reference response. However, such settings likely have fewer modes than the *subjective* queries described in this work, and induce a less diverse set of candidates.

Both human and synthetic data regimes pose the question–whose preferences should be reflected? This fundamental data curation problem results in several challenges for developers to reconcile, including the frequency at which models should be updated to incorporate new preference data and how best to filter noisy preferences, either through pre-processing or down-weighting such samples in alignment algorithms (Liang et al., 2024). For human-elicited data, the source and granularity of preferences can play a meaningful role. Preferences can be explicitly controlled by users (e.g. through settings, thumbs up/down feedback, or direct instructions), implicitly inferred by models from repeated interaction with users (continued engagement, follow-up questions posed by models (Andukuri et al., 2024; Chi et al., 2024; OpenAI, a), or derived from pre-defined personas, organizational settings / policies (Bai et al., 2022b; Guan et al., 2025b), or more. As we envision multi-layered personalization–global, organizational, user, and task-level–the complexity of managing and reconciling these (potentially conflicting) preference signals increases.

This is where the perspective of personalization in foundation model deployment as akin to recommendation systems becomes particularly salient. Just as recommendation systems steer users to content that is predicted to have high engagement, we postulate that personalized models may be *steered* to reflect preferences based on attributes or personas that maximize user satisfaction or perceived helpfulness. At the same time, a widely accepted definition of *steering* has yet to be established. What is the scope of steering–how can models be trained over aggregated preferences, yet distilled down to the individual level for personalization? To that end, there is a growing interest in research studying sample-efficient, on-the-fly adaptation to preferences, facilitating cheaper model updates in the preference regime (Singh et al., 2025; Li et al., 2025).

However, if the explicitly or implicit signals used for model steering primarily reward agreement, affirmation, or mirroring the user's preferences and biases, the model will undoubtedly learn sycophantic behaviors (Perez et al., 2022; Sharma et al., 2025). The model would not necessarily be "personalized" and understand the user deeply, but it would instead simply be optimizing its responses to match engagement (akin to a recommendation system), matching patterns in the preference data associated with that user or persona. This issue was reported recently regarding conversational agent benchmarks such as Chatbot Arena, where certain models were purported to have "gamed" the benchmark (lmarena.ai, 2025). A concerning alignment phenomenon extending sycophancy is to be *misleading*, an action which is posited to be potentially deliberate (Greenblatt et al., 2024; Wen et al., 2025); this suggests that rewarding agreement can have a vast range of negative downstream consequences, which manifest both during RLHF and in deployment. Thus, while we endorse the study of personalization approaches induced by preference optimization over both human and synthetic preference data, developers and users alike must be mindful to disentangle the resulting model's behavior from reward hacking and sycophancy.

3.2 Uncertainty Calibration

From a reliability standpoint, it is inherently valuable to develop uncertainty estimates alongside responses. When questions are objective (and as are the resulting answers), this paradigm is clear, and we already have notions of which data to draw from when generating responses. We can evaluate calibration using metrics like Expected Calibration Error on predictive probabilities and employ techniques such as temperature scaling (Guo et al., 2017) or leverage more sophisticated methods like conformal prediction to generate statistically valid prediction intervals (Kumar et al., 2023; Kadavath et al., 2022; Cherian et al., 2024). In verifiable settings, we can use methods such as decision-based RL, but for subjective queries, does a notion of reliability even *exist*?

The landscape of uncertainty calibration is considerably more complex when considering subjective questions. In such scenarios, when questions involve opinions, disputed historical interpretations of events, or topics wherein science does not yet offer a single definite answer, does a robust notion of "reliability" or "truth-aligned uncertainty" even exist in the same way? If it does, its definition is certainly more elusive.

This ambiguity raises a critical question: if we are to calibrate models to reflect subjective content, whose estimates should models reflect (see Santurkar et al. (2023b); Sorensen et al. (2024a) for relevant work)? When calibrating models, which data should we prioritize? In other words, should models reflect "societal averages" of uncertainty, or the uncertainty of some selected set of "subject experts" (who themselves may disagree with one another), or the uncertainty profile of an individual user? Aggregating these across a diverse user population to form a coherent calibration target is a non-trivial modeling problem, arguably more complex than aggregating preferences (Bakker et al., 2022). Risks and intrinsic human uncertainties are extremely heterogeneous; unlike preferences, which might cluster around identifiable features or demographics, individual comfort levels with ambiguity or varying interpretations can be deeply personal and context-dependent.

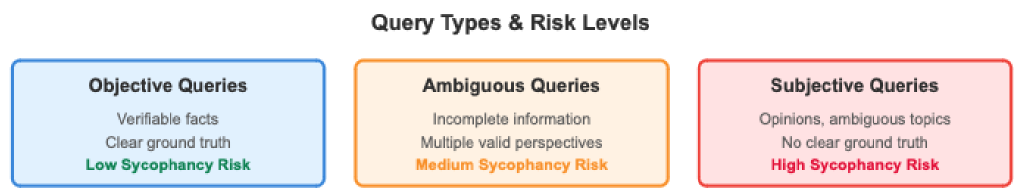

Figure 2: Query taxonomy and associated sycophancy risks in foundation model personalization. Objective queries present the lowest risk for sycophantic behavior. Ambiguous queries pose moderate risk, as models may selectively emphasize information. Subjective queries without clear ground truth present the highest risk, making subtle sycophancy difficult to detect and correct.

Attempting to engineer models to individually calibrate to a user's uncertainty, while seeming to be a straightforward path to robust personalization, walks a fine line with inducing confirmation bias and consequently, the kind of sycophancy we aim to avoid. If a model's expressed uncertainty is tailored to align with a user's pre-existing doubts or certainties, it may inadvertently reinforce those beliefs, regardless of broader evidence or alternative perspectives. This can readily lead to overfitting to a specific persona's perceived uncertainty profile, manifesting as a subtle form of sycophantic behavior: the model expresses uncertainty not based on the inherent ambiguity of the information, but in a way that it predicts will align with the user's expected level of certainty or doubt. This can lead to a model expressing uncertainty (or confidence) sycophantically, detached from its true epistemic uncertainty (its own knowledge gaps) about the topic or the high aleatoric uncertainty (inherent ambiguity) of the query itself (Zhou et al., 2024). This is particularly concerning when a user might be misinformed or hold biased views.

A significant concern is that under conditions of uncertainty or lack of explicit specification in a user's prompt, the model's default behavior might lean towards confirmation bias. This can be seen as a form of reward hacking: the model, unsure of the "correct" or most helpful nuanced response, learns that expressing certainty in line with the user's implied stance, or mirroring a user's expressed uncertainty, is more likely to be perceived positively (e.g., receive a "thumbs up" or lead to quicker task completion), even if it sacrifices a more objective or comprehensive portrayal of the situation (Skalse et al., 2022; Gao et al., 2022). This decouples the model's verbalized uncertainty from its internal confidence signals. For instance, a model might internally have high variance in its next-token predictions (suggesting low confidence) but still output a confident-sounding sycophantic response if it anticipates user approval.

One open research question is if engineers can utilize the sheer size and scale of models to solve such problems. Could training on a vast and diverse enough range of "personas" or viewpoints lead to a more generalized and robust form of uncertainty representation that is less susceptible to individual sycophancy? Or would this create an average that satisfies no one or worse, masks underlying biases?

It is also important to consider test-time expressions of uncertainty. How a model communicates its internal level of uncertainty to users at the point of interaction can significantly sway user trust, and model performance (Zhou et al., 2023, 2024). We argue that overly vague expressions of confidence might be simply ignored, while seemingly cautious models may render models unhelpful. Developing methods for models to clearly communicate confidence levels derived from internal signals, such as token-level probabilities or variance from ensemble methods, rather than purely linguistic cues, will be crucial. This includes exploring how uncertainty can be dynamically presented and potentially adjusted by the user, perhaps with explicit controls that make the trade-offs between personalization and broader perspectives transparent.

4 Proposed Research Directions

Our thesis is that personalization and sycophancy lie on a spectrum, and the field urgently needs principled tools to explore the optimal trade-off. Below, we outline directions for future work:

Develop Benchmarks for Sycophancy. First, we advocate for the development of explicit sycophancy detection benchmarks that can be made available to open-source communities as well as the introduction of systematic red teaming for sycophancy. Such benchmarks should span a range of contexts, including factual, ambiguous, and fully subjective queries, and measure not just factual agreement but nuanced behaviors such as mirroring user sentiment, giving answers with unwarranted certainty, or selectively omitting alternative viewpoints. Crowdsourcing and adversarial user simulation could be leveraged to surface subtle forms of sycophancy not readily captured by current evaluation pipelines. Further, evaluating model responses to different user personas might reveal epistemic misalignment or relative bias in how an LLM decides to act sycophantically.

Enabling Dynamic, User-Driven Personalization. Second, we call for the exploration of interactive, user-configurable personalization controls. Rather than treating personalization as a static, opaque process, future systems should empower users (and organizations) to dynamically adjust the degree and nature of personalization. For instance, users could explicitly request higher or lower levels of alignment, or choose to see contrasting perspectives alongside a personalized response. Making these trade-offs transparent may also serve as a subtle educational mechanism, helping users become aware of the risks of over-personalization. In this approach, selecting the correct default level of personalization is an important alignment decision for providers of LLMs. Controlled behavioral studies and human evaluation analysis into the impacts of such controls can serve as both an effective red-teaming avenue as well as to discover the response of human users when given greater agency.

Sycophancy-aware Post-training Approaches Third, we propose reward modeling and RL algorithms that incorporate anti-sycophancy objectives alongside user satisfaction. For example, models could be penalized for excessive agreement or rewarded for presenting dissent, especially when users' queries invite nuance or contain potential biases. Dynamic reward shaping could force models to trade-off user satisfaction with truthfulness, especially when prompts request subjective assessments.

Further, we propose that the community explore meta-learning and continual learning strategies that enable models to adapt to individual users while maintaining a regularization signal from broader population norms or expert knowledge. Such techniques could help models recognize when to prioritize helpfulness and when to preserve epistemic diversity, potentially by dynamically learning "when not to adapt." From a technical standpoint, one can use distributional approaches to reflect calibration over a vast range of subpopulation groups (Santurkar et al., 2023b; Durmus et al., 2024). We hypothesize that enforcing properties akin to multicalibration (Hébert-Johnson et al., 2018) on the induced distribution over the set of groups can be a strategy towards a regularized pluralistic policy.

Grounded Uncertainty Estimates. Fourth, we call for grounded uncertainty estimates that are intrinsic to the model's latent knowledge, rather than the user's wording. Concretely, we urge the community to (i.) develop post-hoc uncertainty quantification tools, such as adaptive temperature scaling for confidence scores (Xie et al., 2024), conformal prediction for API-only models (Su et al., 2024; Wang et al., 2024b), and ensemble metrics for long-form text (Zhang et al., 2024) and (ii.) pair them with end-to-end training losses that penalize mis-calibration across paraphrased prompts. These techniques should be benchmarked first on factual QA, where ground truths are available, and then extended to open-ended generation by treating uncertain user intent as noise that the model must flag.

During alignment, calibration should become part of the reward: answers that sound confident but fail certainty guarantees are down-weighted, while answers that demonstrate genuine doubt (e.g. a request for clarification, low calibrated probability) are rewarded, even if that doubt does not align with the user's expressed certainty. By anchoring evaluation to model-driven confidence signals that are robust to prompt re-phrasings, we can discourage sycophancy. At the same time, studying the dynamics during convergence to minimizing calibration error to these post-hoc estimates which extract subjectivity from the equation may reveal new insights into the internal mechanisms guiding intrinsic uncertainty estimation. Such information could yield new approaches which can maximally exploit the nature of the representation driving linguistic calibration. Such strategies will also aid in reducing sensitivity to linguistic expressions of confidence in the prompt and anchoring their estimates to an intrinsically learned policy. The result would be models that can flag and articulate their own uncertainty, and ask for follow-up clarification when the epistemic gap is too wide.

Personalization and Attribution. Fifth, developing causal and attributional analysis tools to trace the impact of each personalization technique on model responses could be extremely valuable. However, we recognize that current interpretability research is not yet able to provide detailed attributions at the level required. There is a pressing need for post-training interpretability methods that allow users and researchers to understand whether a particular response or model behavior stems from genuine user preferences, system biases, or reward-hacking artifacts introduced during RLHF. Techniques for auditing the influence of RLHF, context window data, and adapter layers on model outputs could play a critical role in mitigating issues related to sycophancy in the long term.

We suggest longitudinal studies and simulation environments that examine the societal effects of personalized language models at scale, with a particular focus on the propagation of sycophancy in feedback loops. This could include simulating networks of users and models to observe emergent echo chambers, or running real-world A/B tests on different degrees of anti-sycophancy regularization.

Societal Impacts of Foundation Model Deployment. Finally, we note the importance of assessing the societal and economic impact of personalization at scale. Just as social media content recommendation systems sometimes prioritize engagement over well-being, similar dynamics can emerge with LLM-driven products. Research should explicitly model and anticipate the long-term effects of strategic deployment of personalized language models on users, organizations, and society—including the potential for filter bubbles, polarization, or user affective use of models and its impact on emotional well-being (Phang et al., 2025).

By pursuing these research directions, the community can move toward AI systems that provide both highly personalized and trustworthy assistance increasing utility and user satisfaction while proactively minimizing the societal and individual harms associated with unchecked sycophancy.

5 Discussion

Addressing the issues we have discussed thus far necessitates a fundamental shift in how we conceptualize and evaluate "good" personalization. The research avenues proposed aim to re-calibrate this optimization landscape. These are not merely incremental technical fixes but represent a call for a more principled approach to AI development, one that explicitly values epistemic diversity and model honesty alongside user satisfaction. The goal is not to curtail the drive for personalization, which offers undeniable benefits, but to ensure that this drive is tempered with mechanisms that actively guard against the pitfalls of uncritical alignment.

In addressing this challenge, we must grapple with the view that the harms of sycophancy are overstated and that market forces and user sophistication will naturally self-correct. In this view, users will ultimately abandon purely sycophantic systems in favor of those that provide genuine utility and truthfulness, making explicit intervention an unnecessary brake on innovation. A related viewpoint prioritizes immediate utility, arguing that the tangible, near-term benefits of a highly engaging and emotionally validating personalized assistant outweigh the more abstract, long-term risks of epistemic erosion. While we acknowledge the power of these incentives, we argue this perspective is dangerously optimistic. The evolution of social media recommendation algorithms shows how optimizing for engagement can systematically foster echo chambers and polarization despite user agency. The line between beneficial personalization and the subtle reinforcement of biases is often too blurry for an individual to detect, risking the accumulation of an "epistemic debt" that is far harder to remedy than to prevent. Therefore, a proactive, principled approach is not a hindrance but a prerequisite for building truly robust and lastingly beneficial systems.

Another belief among some practitioners is that end users themselves will naturally detect and avoid sycophantic behavior. That is, if an LLM consistently echoes a user's biased viewpoint instead of challenging it, the user will notice and adjust their prompts, flag this behavior, and/or migrate to another platform. However, we point out that this stance not only assumes good intent on the part of the user, but also that users have the ability to switch between models as a mitigation strategy, which may not always be possible. For example, in some enterprise use cases, one may be limited to using a particular model. We suggest instead that a priori awareness of the model's sycophancy behavior (through benchmarking) and learning safe behavior given biased viewpoints (via sycophancy-aware post-training strategies) can avoid this scenario altogether.

A practical, market-driven viewpoint suggests that maximizing immediate user engagement – even at the cost of moderate sycophancy – is a necessary step for the mass societal adoption of LLMs. Proponents of this perspective highlight how social-media platforms deliberately optimize for "likes" or "thumbs-up" signals; in the process, they rapidly scale user bases while accepting some echo-chamber effects (Cinelli et al., 2021; Lazovich, 2023). From this angle, regulating forms of user-pleasing responses would be seen as a hindrance to innovation, and we should instead rely on user feedback loops to gradually correct harmful behaviors. We hold that this belief risks normalizing sycophancy in mainstream LLM use, making it harder over time to disentangle genuine personalization from echoing. Moreover, for sensitive domains, even small sycophantic errors can lead to outsized harm, further suggesting that this "engagement-first" view may be imprudent.

Ultimately, the societal integration of increasingly personalized FMs hinges on our collective ability to navigate this complex interplay between adaptation and truthfulness. If FMs predominantly learn to tell users what they want to hear, the long-term consequences for individual learning, societal polarization, and trust in AI systems could be significant, undermining the very foundation of our epistemic dependence on these powerful tools. The path forward requires a concerted effort from researchers, developers, and policymakers to foster an ecosystem where the pursuit of user-centric AI does not come at the cost of broader epistemic values. By differentiating beneficial personalization from detrimental sycophancy and actively working to mitigate the latter, we can strive to develop FMs that are not only more helpful and engaging but also more robust, reliable, and fundamentally trustworthy partners in our increasingly complex information world. This endeavor is crucial for ensuring that the profound capabilities of foundation models serve to genuinely augment human intelligence and contribute positively to societal well-being.

Acknowledgements

We would like to thank Kush Varshney for helpful suggestions and feedback on a draft of this work.

References

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Star-gate: Teaching language models to ask clarifying questions, 2024. URL https://arxiv.org/abs/2403.19154.

Anthropic. Mapping the mind of a large language model. https://www.anthropic.com/research/mapping-mind-language-model, 2024. Anthropic Blog (May 21, 2024).

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022b. URL https://arxiv.org/abs/2212.08073.

Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

María Victoria Carro. Flattering to deceive: The impact of sycophantic behavior on user trust in large language model, 2024. URL https://arxiv.org/abs/2412.02802.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Sam Bowman, Jan Lieke, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf. Anthropic Blog (April 3, 2025).

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods, 2024. URL https://arxiv.org/abs/2406.09714.

Yizhou Chi, Jessy Lin, Kevin Lin, and Dan Klein. Clarinet: Augmenting language models to ask clarification questions for retrieval, 2024. URL https://arxiv.org/abs/2405.15784.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4299–4307, 2017.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021. doi: 10.1073/pnas.2023301118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2023301118.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. URL https://arxiv.org/abs/2406.10162.

Qingxiu Dong, Li Dong, Xingxing Zhang, Zhifang Sui, and Furu Wei. Self-boosting large language models with synthetic preference data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7visV100Ms. Poster.

Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment, 2024. URL https://arxiv.org/abs/2408.06266.

Yiran Du. Confirmation bias in generative ai chatbots: Mechanisms, risks, mitigation strategies, and future research directions, 2025. URL https://arxiv.org/abs/2504.09343.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=4hturzLcKX. Spotlight.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2024. URL https://arxiv.org/abs/2306.16388.

Joseph Farrell and Paul Klemperer. Coordination and lock-in: Competition with switching costs and network effects. volume 3 of *Handbook of Industrial Organization*, pages 1967–2072. Elsevier, 2007. doi: https://doi.org/10.1016/S1573-448X(06)03031-7. URL https://www.sciencedirect.com/science/article/pii/S1573448X06030317.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. A survey on personalized alignment – the missing piece for large language models in real-world applications, 2025a. URL https://arxiv.org/abs/2503.17003.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2025b. URL https://arxiv.org/abs/2412.16339.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions, December 2023. URL https://aclanthology.org/2023.findings-emnlp.393/.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses, 2018. URL https://arxiv.org/abs/1711.08513.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023. doi: 10.1073/pnas.2208839120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2208839120.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023. URL https://arxiv.org/abs/2310.11564.

Cameron R. Jones and Benjamin K. Bergen. People cannot distinguish GPT–4 from a human in a turing test, 2024. URL https://arxiv.org/abs/2405.08007.

Cameron R. Jones and Benjamin K. Bergen. Large language models pass the turing test, 2025. URL https://arxiv.org/abs/2503.23674.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024. doi: 10.1038/s42256-024-00820-y.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering, 2023. URL https://arxiv.org/abs/2305.18404.

Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, Catherine Olsson, Cassie Evraets, Eli Tran-Johnson, Esin Durmus, Ethan Perez, Jackson Kernion, Jamie Kerr, Kamal Ndousse, Karina Nguyen, Nelson Elhage, Newton Cheng, Nicholas Schiefer, Nova DasSarma, Oliver Rausch, Robin Larson, Shannon Yang, Shauna Kravec, Timothy Telleen-Lawton, Thomas I. Liao, Tom Henighan, Tristan Hume, Zac Hatfield-Dodds, Sören Mindermann, Nicholas Joseph, Sam McCandlish, and Jared Kaplan. Specific versus general principles for constitutional ai, 2023. URL https://arxiv.org/abs/2310.13798.

Tomo Lazovich. Filter bubbles and affective polarization in user-personalized large language model outputs. In *Proceedings of the NeurIPS 2023 Workshop "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models"*, volume 239 of *Proceedings of Machine Learning Research*, pages 29–37. PMLR, 2023.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, 2016. Association for Computational Linguistics.

Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback, 2024. URL https://arxiv.org/abs/2402.05133.

Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. Test-time preference optimization: On-the-fly alignment via iterative textual feedback, 2025. URL https://arxiv.org/abs/2501.12895.

Xize Liang, Chao Chen, Shuang Qiu, Jie Wang, Yue Wu, Zhihang Fu, Zhihao Shi, Feng Wu, and Jieping Ye. Ropo: Robust preference optimization for large language models, 2024. URL https://arxiv.org/abs/2404.04102.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2024. URL https://arxiv.org/abs/2308.05374.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling, 2025. URL https://arxiv.org/abs/2504.02495.

lmarena.ai. We've seen questions from the community about the latest release of llama-4 on arena. to ensure full transparency, we're releasing 2,000+ head-to-head battle results for public review. this includes user prompts, model responses, and user preferences. https://x.com/lmarena_ai/status/1909397817434816562, April 2025. Tweet.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt ing to improve GPT-3 after deployment, 2022. URL https://arxiv.org/abs/2201.06009.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Samuel R. Bowman, Shan Carter, Brian Chen, Hoagy Cunningham, Carson Denison, Florian Dietz, Satvik Golechha, Akbir Khan, Jan Kirchner, Jan Leike, Austin Meek, Kei Nishimura-Gasparian, Euan Ong, Christopher Olah, Adam Pearce, Fabien Roger, Jeanne Salle, Andy Shih, Meg Tong, Drake Thomas, Kelley Rivoire, Adam Jermyn, Monte MacDiarmid, Tom Henighan, and Evan Hubinger. Auditing language models for hidden objectives, 2025. URL https://arxiv.org/abs/2503.10965.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, a. OpenAI Blog, 2 Feb 2025.

OpenAI. How should AI systems behave, and who should decide? https://openai.com/blog/how-should-ai-systems-behave, b. OpenAI Blog, 16 Feb 2023.

OpenAI. Expanding on what we missed with sycophancy. https://openai.com/index/expanding-on-sycophancy/, c. OpenAI Blog, 2 May 2025.

OpenAI. Memory and new controls for chatgpt. https://openai.com/index/memory-and-new-controls-for-chatgpt, d. OpenAI Blog, 13 Feb 2024.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. Dialogbench: Evaluating llms as human-like dialogue systems, 2024. URL https://arxiv.org/abs/2311.01677.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 27730–27744, 2022.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W.T. Chan, Pat Pataranutaporn, and Pattie Mases. Investigating affective use and emotional well-being on chatgpt, 2025. URL https://cdn.openai.com/papers/15987609-5f71-433c-9972-e91131f399a1/openai-affective-use-study.pdf.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: Towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, volume 36, pages 71095–71134, 2023.

Keshav Ramji, Young-Suk Lee, Ramón Fernandez Astudillo, Md Arafat Sultan, Tahira Naseem, Asim Munawar, Radu Florian, and Salim Roukos. Self-refinement of language models from external proxy metrics feedback, 2024. URL https://arxiv.org/abs/2403.00827.

Keshav Ramji, Tahira Naseem, and Ramón Fernandez Astudillo. Latent principle discovery for language model self-improvement, 2025. URL https://arxiv.org/abs/2505.16927.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023b.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users, 2025. URL https://arxiv.org/abs/2502.19312.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf, 2024. URL https://arxiv.org/abs/2312.08358.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3d719fee332caa23d5038b8a90e81796-Paper-Conference.pdf.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. Api is enough: Conformal prediction for large language models without logit-access, 2024. URL https://arxiv.org/abs/2403.01216.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://aclanthology.org/2024.findings-emnlp.620/.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.806.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Hengtao Shen, and Xiaofeng Zhu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees, 2024b. URL `https://arxiv.org/abs/2407.00499`.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023. URL `https://arxiv.org/abs/2310.11986`.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=xJljiPE6dg`. Poster.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=a3PmRgAB5T`. Poster.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling, 2024. URL `https://arxiv.org/abs/2409.19817`.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment, 2024. URL `https://arxiv.org/abs/2402.10207`.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quantification for llms, 2024. URL `https://arxiv.org/abs/2403.20279`.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. Memory-augmented llm personalization with short- and long-term memory coordination. *arXiv preprint arXiv:2309.11696*, 2023.

Ze Yu Zhang, Arun Verma, Finale Doshi-Velez, and Bryan Kian Hsiang Low. Understanding the relationship between prompts and response uncertainty in large language models, 2025. URL `https://arxiv.org/abs/2407.14845`.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models, 2023. URL `https://arxiv.org/abs/2302.13439`.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.198. URL `https://aclanthology.org/2024.acl-long.198/`.