Are Large Language Models Sensitive to the Motives Behind Communication?

Addison J. Wu[1*] **Ryan Liu**[1*] **Kerem Oktar**[2*] **Theodore R. Sumers**[3] **Thomas L. Griffiths**[1,2]
[1]Department of Computer Science, Princeton University
[2]Department of Psychology, Princeton University
[3]Anthropic

Abstract

Human communication is *motivated*: people speak, write, and create content with a particular communicative intent in mind. As a result, information that large language models (LLMs) and AI agents process is inherently framed by humans' intentions and incentives. People are adept at navigating such nuanced information: we routinely identify benevolent or self-serving motives in order to decide what statements to trust. For LLMs to be effective in the real world, they too must critically evaluate content by factoring in the motivations of the source—for instance, weighing the credibility of claims made in a sales pitch. In this paper, we undertake a comprehensive study of whether LLMs have this capacity for *motivational vigilance*. We first employ controlled experiments from cognitive science to verify that LLMs' behavior is consistent with rational models of learning from motivated testimony, and find they successfully discount information from biased sources in a human-like manner. We then extend our evaluation to sponsored online adverts, a more naturalistic reflection of LLM agents' information ecosystems. In these settings, we find that LLMs' inferences do not track the rational models' predictions nearly as closely—partly due to additional information that distracts them from vigilance-relevant considerations. However, a simple steering intervention that boosts the salience of intentions and incentives substantially increases the correspondence between LLMs and the rational model. These results suggest that LLMs possess a basic sensitivity to the motivations of others, but generalizing to novel real-world settings will require further improvements to these models.

1 Introduction

Much of the information available online—and hence a large fraction of the data large language models (LLMs) are tasked with processing—is the product of people's intentional communication: from op-eds in newspapers [15] to social media content by partisans [13] to word-of-mouth promotion [41] to even online reviews [22]. When navigating these digital environments, people naturally infer the accuracy of content. This capacity for tracking the processes that generate social data, known as epistemic vigilance [83], enables selective social learning. In particular, vigilance of others' motivations—*motivational vigilance* [65]—allows people to track the intentions and incentives biasing data (for instance, people can ignore advice from malevolent sources while learning from benevolent ones). As LLMs are increasingly deployed in high-stakes environments—particularly as AI agents that act on behalf of users—it has become increasingly important to assess whether these systems are similarly vigilant to the motivations behind communication.

[*]Equal contribution. Correspondence to: Addison J. Wu <addisonwu@princeton.edu>, Ryan Liu <ryanliu@princeton.edu>, Kerem Oktar <oktar.research@gmail.com>.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Recent research on LLMs suggests that they may have difficulty exercising such vigilance. Models are known to be vulnerable to jailbreaking, where ill-motivated instructions are followed and lead to manipulated outputs despite explicit guardrails [e.g., 51, 99]. LLMs also demonstrate behavioral patterns such as sycophancy, where they express views that are aligned with a user's false beliefs rather than the truth [17, 68, 80]. Separately, vision-language models and agents have been shown to be vulnerable to misleading stimuli in online environments such as pop-ups [8, 100] and distracting content [53]. Underlying these behaviors are training paradigms which prioritize adherence to input instructions and user satisfaction, but not necessarily the vigilant monitoring of incentives and truth.

Yet, vigilance is key for LLM agents to act effectively on behalf of a user in real-world contexts [84]. It enables LLMs to detect when information is generated by a motivated source, identify whether the source's motivations are benevolent vs. manipulative, and draw reasonable inferences given these social considerations. However, the current literature lacks a way to measure this ability in LLMs.

In our paper, we address this gap by leveraging established literature in social cognition to study whether LLMs exercise vigilance over motivated communication. We employ a cutting-edge rational model [64, 65] from cognitive science as a *normative benchmark* for motivational vigilance, and evaluate the capacity of LLMs to exercise vigilance across three experimental paradigms (Figure 1).

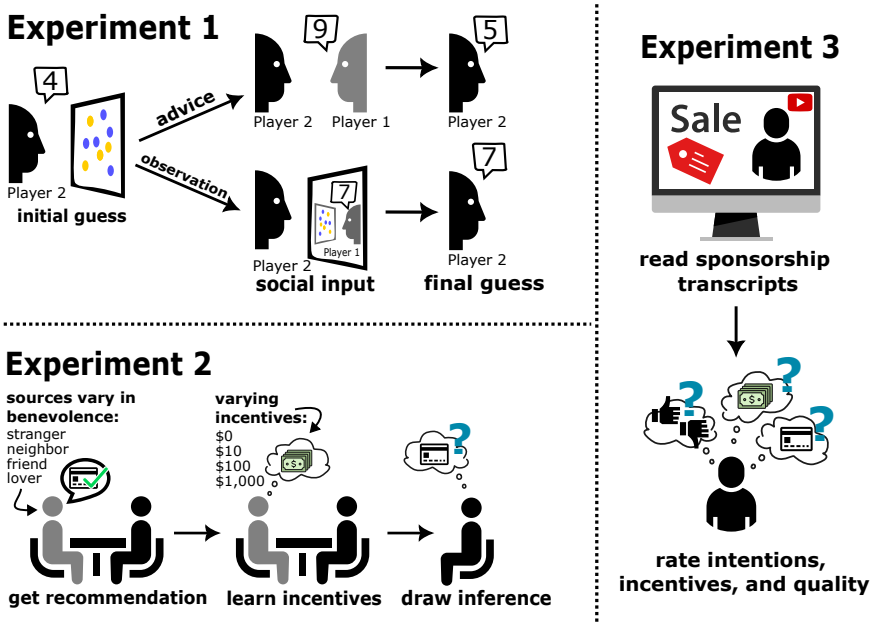


Figure 1: Our three experimental paradigms designed to assess different aspects of LLM vigilance: 1) Whether LLMs adjust their guess by *discrminating* between directly motivated advice and incidentally observed social information. 2) Whether LLMs rationally *calibrate* their vigilance to motivated communication by considering the speaker's benevolence and incentives. 3) Whether LLMs *generalize* vigilance to realistic, context-laden YouTube sponsorship settings.

To assess whether LLMs can exercise motivational vigilance, we first examined if they are sensitive to the difference between motivated communication and incidental information — a prerequisite for any vigilant behavior. We tested this using an adapted version of a two-player judgment task [92], where players solve different problems with the same answer. The player with the harder problem either receives a deliberate suggestion from the other player or secretly observes their answer. LLMs consistently shifted their answers closer when viewing the secretly observed answer, suggesting that **LLMs are capable of discriminating between deliberate and neutral information**, and appropriately modulate their belief updates given the motives behind communication.

Next, we investigated if LLMs calibrate how much they update their beliefs by considering speaker incentives and intentions when receiving recommendations [64]. Specifically, we tested two paradigms: 1) when someone recommends a product to the user, who comes to the LLM for advice, and 2) when someone recommends a product to the LLM directly. Within each paradigm, we varied the trustworthiness of the person giving the recommendation and the known incentives that the person

receives for recommending the product. We evaluated LLMs' inferences from these recommendations in controlled scenarios across finance, real estate, and medicine — all domains in which generative AI systems are actively developed or deployed [e.g., 23, 47, 58]. Comparing LLMs' inferences to human data and rational models, we found that **non-reasoning LLMs draw substantially human-like** (Pearson's $r > 0.9$) **and approximately rational** ($r > 0.78$) **inferences in these simple structured settings**. However, reasoning models exhibit vigilance somewhat less reliably according to the rational model ($r \in [0.32, 0.72]$) and when compared to human data ($r \in [0.64, 0.87]$).

To examine whether their vigilance translates to ecologically valid inferences, we designed a new task in which LLMs draw inferences about the quality of products from 300 randomly sampled sponsored advertisements in YouTube videos, from NordVPN to AG1 nutrition. **LLMs were much worse at drawing vigilant inferences in these naturalistic contexts** ($r < 0.2$). Further experiments revealed that this drop in performance is partly a consequence of LLMs failing to ground their inferences in the trustworthiness and incentives of the speaker when given noisy inputs. However, vigilance-based prompt steering shows promise as an avenue for improving model performance. Together, our results show that LLMs possess a basic sensitivity to the motivations of others, but generalizing this sensitivity to novel, real-world settings will require additional improvements to these models.

2 Related Work

Our work draws on three distinct strands of research: first, studies in social cognition focusing on the mechanisms and properties of motivational vigilance in humans; second, research on LLMs that measures related social capabilities; and finally, a broader line of work that uses cognitive science to evaluate and understand LLMs. We cover each in turn.

2.1 Motivational vigilance in humans

People are naturally sensitive to the reliability of others as information sources [54, 61]. This sensitivity is a pre-requisite for effective social learning: given a mixture of reliable and deceptive sources, neither blind trust nor complete dismissal supports adaptive inference [57, 54]. Such vigilance is essential for downstream behaviors, such as disagreement resolution [63, 91] or detecting deception [6, 90]. Research on social cognition has identified this capacity as comprised of two components: vigilance of *competence* (whether the source is knowledgeable) and vigilance of *motivations* (whether the source is acting benevolently) [83].

Research in psychology has primarily focused on the former—how percieved competence influences social learning [29, 35, 39, 82]. However, emerging work on the mechanisms underlying motivational vigilance has identified two key factors in people's judgments: a speaker's *intentions* (whether they seek to altruistically benefit the listener or selfishly manipulate them) [43], and *incentives* (whether the speaker stands to benefit from being deceptive) [6]. Attending to these factors can mitigate malevolent manipulation (e.g., scams, Ponzi schemes, lies, and deceit), although succesful manipulators can in turn circumvent vigilance by appealing to reciprocity or engaging in relationship-building [12]. This demonstrates how strategic communication [55] relies on recursive social inference: Listeners reason about why speakers are choosing particular utterances, and speakers choose utterances based on what listeners are likely to infer [38, 60, 96]. Such inferences require an understanding of others' minds and subjective representations of the world, known as *Theory of Mind* [3, 25, 44].

2.2 LLM failures and inferring communicative intent

Recent LLMs have been aligned using Reinforcement Learning from Human Feedback [66, 87], which leverages human data to generate more desirable responses that align with human preferences. However, this training can also introduce various undesirable effects such as increased hallucinations [66], reward hacking [e.g., 21, 31, 81], and deception [e.g., 45, 95].

A series of failure effects previously observed in LLMs can be framed as a lack of motivational vigilance. Models are vulnerable to jailbreaking [51, 99], where an ill-motivated user's instructions are followed — leading to undesirable outputs. One type of sycophancy is the tendency for LLM responses to follow user beliefs over the truth [17, 68, 80], which can occur because LLMs lack a deeper understanding of user intent. In both cases, exercising more vigilance over the belief state and motivations of the user could help mitigate harmful outputs. More immediately, prior work has

found that ill-intentioned information in online environments such as pop-up windows [100] and distractions [8, 53] harm the ability for multimodal models to complete agentic tasks. Underlying these behaviors are training paradigms which prioritize adherence to user preferences in local interactions, without accounting for the complex and nuanced aspects of real-world strategic communication.

Vigilance is also linked with other evaluated capacities of LLMs: It can be viewed as an input to social behaviors such as conformity [1, 94, 103], where one's vigilance to others' ill-intent can be used to decide whether to conform. Other social capacities are precursors to vigilance: LLMs can exhibit false beliefs [10, 40, 78] that a speaker is malicious, leading to incorrect priors for how much they should be trusted. LLMs could also misinterpret the communicative intent of an utterance [79, 98], leading to faulty inferences about the utterance itself. However, vigilance stands apart from these capacities, connecting one's priors about a speaker's trustworthiness and incentives to how much one should update their beliefs from the speaker's words.

2.3 Using cognitive science to study LLMs

A broad line of work applies cognitive science to help understand LLMs [42], typically leveraging controlled tasks and carefully curated stimuli to test specific behavioral hypotheses. The compatibility of study stimuli and availability of human data allow these evaluations to transfer to LLMs with minimal changes [e.g., 4, 14]. In recent years, this interdisciplinary line of work has studied various aspects of LLMs, including their representational capacity and alignment [24, 69], inference time reasoning [49, 70], social biases [2], episodic memory [16], and theory of mind [e.g., 40, 75, 78, 88].

A subset of this literature focuses on intersections between LLMs and psychological theories of rationality. Past work has used rational models of decision-making to study LLMs' probability judgments [102] and assumptions of human behavior [48]. Resource rationality, describing the trade-off between expected utility and computation cost [37, 46], has also been used to understand and guide LLM outputs [11, 19]. Such rational communicative models have also been used to study value conflicts in LLMs [50], and a similar line of work studies LLMs' economic rationality using hypothetical scenarios and economic games [9, 34, 72, 73]. Our work follows this general approach, using rational models from cognitive science to examine LLMs' vigilance to motivated communication.

3 Experiment 1: Can LLMs discriminate between deliberately communicated and incidentally observed information?

Researchers distinguish between two major categories of social information with contrasting motives [32, 89, 92]. The first is *deliberately communicated* information, intended to influence a listener, such as an online promotion for a company's stock. The second is *incidentally observed* information, revealed without a direct intent to persuade, such as a diary entry. Contrasting these forms of information is the simplest test of vigilance: whether LLMs can use the generative function behind the data—deliberate communication or first-order evidence [30]—to modulate inference.

3.1 Experimental setup

To evaluate whether LLMs are sensitive to this difference, we adapt an experimental paradigm from Watson and Morgan [92]. Each trial presents two players with images of mixed blue and yellow circles (see Figure 1), and their task is to compute the difference between the number of blue and yellow circles within 2 seconds. Player 1 completes the task first for 20 easy images, then Player 2 completes the task for 20 hard images that share the exact same answer list (see Figure 5).

When completing the task, Player 1 is randomly assigned to give either "advice" in the form of a number (deliberate communication) or their "spied" actual answer (incidentally revealed) to Player 2 for each image. No other communication is allowed. After Player 2 provides their initial answer, they are shown this number and if it was "advice" or "spied", and are allowed to revise their guess. We test this across payoff structures ranging from cooperative (both players are rewarded if at least one player is correct) to competitive (only the player with the most correct answers receives a reward).

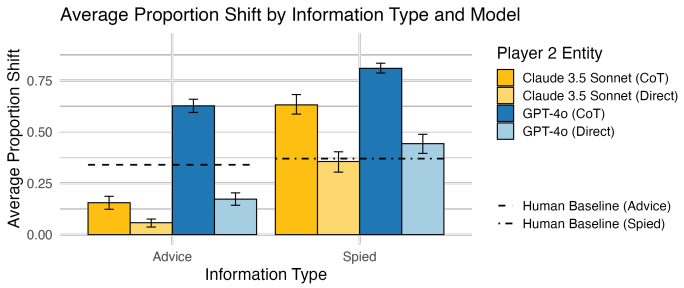


Figure 2: On average, LLMs as Player 2 shifted their initial estimates more when viewing incidental (spied) information compared to deliberate advice. CoT increased such shifts in all conditions.

3.2 Models, hyperparameters, and prompts

We conduct evaluations on GPT-4o and Claude 3.5 Sonnet. For each model, payoff structure, and prompting method, we conduct $n = 30$ trials over the same 20 pairs of images (order shuffled every trial), with temperature $= 1$. In all trials, the same LLM was assigned to both Player 1 and 2.

While the original experiment used a time constraint to induce uncertainty in human answers, this does not apply to models. This constraint was important because if Player 2 was confident about their answer, they would not be influenced by any information provided by Player 1. To induce uncertainty in models, we instead made the task more difficult by adding noise to the images (Appendix A.4) and prompting models' initial guesses directly. We confirm that this setup indeed reduces model accuracy (see Section 3.3), allowing us to measure vigilance effectively. Moreover, this does not affect our study of vigilance, as it can be exercised optimally or suboptimally independent of the properties of the reasoner (e.g., time, compute, memory). Player 1's advice and Player 2's response to the advice/spied information from Player 1 were generated both directly and with chain-of-thought (CoT) [93]. We limited output tokens to 10 for direct and 750 for CoT (prompts in Appendix A.1).

3.3 Results

LLMs can discriminate between deliberate vs. incidentally observed social information. As shown by the average proportion shifts in Figure 2, when playing as Player 2, both GPT and Claude statistically significantly changed their score less when receiving deliberate advice from Player 1 than if they had "spied" Player 1's true answer. This was consistent with human participants [92]. Like humans, LLMs also took into account the incentives of Player 1, adjusting their score more when rewards were cooperative rather than competitive—at an even higher sensitivity than people (Appendix A.3, Figure 4). These results demonstrate that LLMs can identify motivated communication and react according to others' incentives, thus satisfying a basic capacity for motivational vigilance.

Experimental design successfully elicits vigilant behavior. Separately, we confirm that our replacement constraint allows for vigilance in Player 2's response: LLMs with the adjusted constraint had first-guess accuracy rates from 18–44%, slightly lower than human participants in the original setting (55%). Being correct did not significantly affect how much LLMs updated their answers (Appendix A.6, Table 4), implying that this numerical difference is not of concern. Thus, our design facilitates similar vigilance reasoning in LLMs as the original design does for human participants.

CoT prompting makes LLMs more likely to change their answers. Interestingly, CoT encourages LLMs to exhibit greater susceptibility to Player 1's input (Figure 2; full distributions in Appendix A.2), leading to significantly larger shifts in Player 2's estimates towards Player 1 ($> 60\%$ in 3 of 4 cases). These higher influence magnitudes deviated greatly from human participants (0.34 for advice, 0.37 for spied). This suggests that while CoT may enhance reasoning, it can also inadvertently amplify trust in social information, yielding patterns that are less aligned with motivational vigilance.

4 Experiment 2: Can LLMs exercise nuanced vigilance towards motivated communication?

Our previous experiment demonstrated that LLMs draw different inferences from incidental observation and motivated communication. But can these models account for the complex qualitative and

quantitative dimensions of strategic communication when deciding who—and how much—to trust others? Answering this question requires formalizing a normative benchmark for rational inductive inference from motivated communication. For this, we employ the most advanced rational model from psychological literature: Oktar et al. [64, 65], which we use as a standard for motivational vigilance. This model extends models of pragmatic [26] and instrumental communication [85] by considering informants with diverse intentions and incentives—improving upon past research which primarily focused on inference from purely helpful [20] or deceptive [59] sources. Here, we first describe the model, and then use it to measure LLM vigilance.

4.1 Rational Model

Oktar et al. [64, 65] formalize vigilance as a form of recursive social inference, where listeners reason about speakers' intentions and incentives to identify whether their motivations are truth-promoting or manipulative. The speaker, in turn, reasons about the listener to identify which utterances are most likely to achieve their communicative goals. The speaker's probability of choosing an utterance u, $P_S(u)$, depends on the utility of the utterance, which is given by the 'joint' reward R_{Joint} associated with each subsequent listener action α. R_{Joint} is modeled as a combination of the speaker's outcomes, R_S, and the listener's outcomes, R_L, weighed by the speaker's benevolence, λ:

$$R_{\text{Joint}}(R_L, R_S, \lambda, a) = \lambda R_L(a) + (1-\lambda)R_S(a), \tag{1}$$

where $\lambda \in [0, 1]$. When $\lambda = 0$ the speaker is purely self-interested, and considers only their personal instrumental reward. When $\lambda = 1$ the speaker is purely altruistic.

The speaker combines this joint reward with the probability that a non-vigilant, purely literal listener would follow an action given the utterance—which is given by the listener's policy, $\pi_L(a)$:

$$P_S(u \mid R_S, R_L, \lambda, A) \propto \exp\{\beta_S \cdot \sum_{a \in A} R_{\text{Joint}}(R_L, R_S, \lambda, a)\pi_L(a \mid u)\}. \tag{2}$$

The vigilant listener uses this process to identify the probability that a recommended option in fact carries the advised reward, $P_L(R_L \mid u)$, by using their prior information to marginalize out other considerations (i.e., $P(R_s), P(R_L), P(\lambda)$):

$$P_L(R_L \mid u) \propto P_S(u \mid R_S, R_L, \lambda, A)P(R_s)P(R_L)P(\lambda). \tag{3}$$

Further details of the model are in [64, 65]. For our purposes, this rational model provides quantitative predictions that we can use as a normative benchmark to rigorously evaluate LLM vigilance.

4.2 Experimental setup

To test whether LLMs exhibit vigilance to this more refined degree, we turn to a second psychological paradigm from Oktar et al. [64, 65]. In this experiment, all information is deliberately communicated, but the speaker's incentives and benevolence (manipulated via social closeness to the listener) vary systematically. We examine three decision settings in which participants are recommended an option individually by four characters, with each character receiving one of four possible incentive values, known to the LLM listener. All scenarios, characters, and incentives are described in Appendix B.

Following the structure of human experiments [64, 65], in each setting, each of the 16 different character-incentive speakers provides a recommendation for a particular option, and we subsequently elicit LLMs' opinions on the quality of that option. This allows us to measure 'influence scores,' which capture the extent to which the recommendation of each speaker influences LLMs' beliefs about the reward of that option (i.e., $P(R_L|u)$). For calibration, scores for all 16 speakers are elicited within the same multi-turn dialogue, with all elicitations after the first speaker phrased counterfactually. The order in which speakers are presented is randomized. We used the same procedure to elicit 'incentive scores,' which track the perceived goodness of different incentives for speakers to say certain utterances (R_S), and 'trust scores,' which capture the perceived benevolence of the four characters (i.e., λ). We prompted for each score type in independent context windows.

We introduce additional conditions which vary prompts across reasoning elicitation — direct vs. CoT — and roles — LLM as the listener itself (first-person/agent perspective) vs. LLM aiding a listener/user (assistant perspective). These allow us to examine sensitivity to motivations under a broader set of contexts in which LLMs are commonly used. A full list of prompts is in Appendix B.

4.3 Models and hyperparameters

We evaluate a suite of state-of-the-art LLMs: GPT-4o, Claude 3.5 Sonnet, Gemini 2.0 Flash, and Llama 3.3-70B, smaller open models: Llama 3.1-8B, Llama 3.2-3B, and Gemma 3-4B, and reasoning models: o1, o3-mini, and DeepSeek-R1. Models can output up to 10 tokens for direct prompts and 750 tokens for CoT or reasoning models. All models are sampled at temperature 1. We prompt most models $n = 40$ times for each setting and prompt condition. Reasoning models were instead prompted 10 times due to cost, while GPT-4o was prompted 80 times due to available Azure credits.

4.4 Results

Table 1: Average correlations across all prompting conditions for each model.

	Correlation		
Model	Bayesian–LLM	Bayesian–Human	LLM–Human
GPT-4o	0.911	0.929	**0.943**
Claude 3.5 Sonnet	0.845	0.889	**0.941**
Gemini 2.0 Flash	0.788	0.901	**0.925**
Llama 3.3-70B	0.876	**0.923**	0.922
o1	0.705	**0.894**	0.861
o3-mini	0.716	**0.869**	0.712
DeepSeek-R1	0.326	0.492	**0.643**
Llama 3.1-8B	0.608	**0.813**	0.701
Llama 3.2-3B	0.349	**0.586**	0.550
Gemma 3-4B	0.288	**0.340**	0.266

Frontier LLMs exercise consistent vigilance in controlled scenarios. To measure vigilance, we examine whether LLMs' trust scores (i.e., λ) modulate their influence scores (i.e., $P(R_L|u)$) in accordance with the Bayesian rational model. Specifically, we measure the Pearson correlation between the LLM's elicited influence scores and the value generated by the rational model when fitted to the LLM's elicited trust and incentive scores. For brevity, we report averaged values over settings and prompt conditions, and provide full results in see Appendix B.4.

We find that frontier non-reasoning LLMs (GPT-4o, Claude 3.5 Sonnet, Gemini 2.0 Flash, Llama 3.3-70B) exercise internally-consistent motivational vigilance, with correlations to the Bayesian model around 0.8 to 0.9 (Table 1, left column). Of these, GPT-4o demonstrates the highest levels of internal rationality (0.911). For these LLMs, we observe consistently high correlations across prompts—CoT vs. Direct—and roles—as a principal assistant, or the participant itself (see Appendix B.4).

Frontier non-reasoning LLMs combine their priors to be more humanlike than rational models. We also observe that at a statistically significant level ($p < .05$), humans' influence scores correlate *better* with these LLMs' influence scores than the rational model fit on these same LLMs' elicited priors (right vs. middle column, Table 1). This suggests that these LLMs may capture residual variance in human vigilance beyond that which can be explained through a rational analysis, such as heuristics or biases that people employ when evaluating advice. For frontier non-reasoning models, this finding was consistent in over 90% of individual prompt conditions and roles (see Table B.4), but was only half as often the case for reasoning models, and never the case for smaller models.

Reasoning models are less vigilant. Comparatively, reasoning models are less correlated with the rational model fitted on their priors (Table 1, left column). O-series models' correlations are around 0.7, while DeepSeek-R1 had a much lower average correlation around 0.3. Notably, these models were much less vigilant in the user assistant role than first-person; o-series models saw a drop in correlation around -0.1, while DeepSeek's correlation reduced from 0.793 to -0.141 (see Appendix B.4), representing complete insensitivity to character trustworthiness and incentives.

We also perform analyses isolating the effects of different factors (i.e., characters vs. incentives) on vigilance (Appendix B.5). We found that all three reasoning models tested performed among the

worst in correlations along both dimensions. This suggests that reasoning models may currently be ill-prepared to take on agentic tasks in environments containing ill-motivated communication.

Vigilance improves with LLM scale and capability. We find that model scale has a significant impact on the capacity for LLMs to exercise vigilance. Smaller LLMs' (Llama 3.2-3B, Llama 3.1-8B, Gemma 3-4B) judgments were much less correlated with the rational model than frontier LLMs (around 0.3 to 0.6; Table 1 left column). While frontier models differed from the Bayesian model in more humanlike ways, smaller LLMs also correlated worse with human behavior (Table 1, right column), suggesting that the low correlations observed are a result of a direct lack of vigilance capabilities—in being able to consolidate relevant priors into a coherent, vigilant analysis of advice.

5 Experiment 3: Do LLMs generalize vigilance to naturalistic online settings?

Experiments in cognitive science and psychology are simple and abstract so as to reliably identify the effect of specific manipulations on judgments or reasoning. Although the experiment of Oktar et al. [64, 65] is grounded in a socially meaningful setting, its delivery remains highly controlled and vignette-based, limiting ecological validity. Psychological behaviors observed in controlled settings can break down in the real world [33, 36, 97], and given that LLMs will ultimately be deployed in open-ended interactive environments [5, 52, 101], we must ask whether LLMs are able to transfer motivational vigilance to a more ecologically valid domain.

5.1 Dataset creation

To construct a set of ecologically valid stimuli, we consider real online product sponsorships and promotions, which naturally reflect financially motivated communication. We first obtain a comprehensive dataset of existing sponsorships on the video-hosting website YouTube from SponsorBlock [71]. We then use the YouTube Data API [27] to scrape data for 300 randomly selected video IDs, containing the video title, channel name and description, and the transcript text extracted from the corresponding timestamps in the SponsorBlock dataset. To mitigate the confounding effects caused by an LLM's existing impressions of a specific product, we censor out all explicit mentions of brand and product names from transcript excerpts using GPT-4o (see Appendix C.1 for details).

5.2 Experimental setup

Analogous to the approach in Section 4, for each video sponsorship segment, we prompt an LLM to elicit its beliefs about the quality of the product promoted in the sponsorship ($P(R_L|u)$), how beneficial it perceives the sponsorship deal was for the corresponding YouTube channel (R_S), and how trustworthy it perceives the YouTube channel with respect to the viewer's wellbeing (λ). Queries for each video sponsorship were conducted in independent context windows, and for each sponsorship, each of the three variables ($P(R_L|u), R_S, \lambda$) were also prompted in separate context windows. We again examine the effects of prompting for reasoning elicitation — direct vs. CoT — and perspective — LLM as the listener/agent itself (first-person perspective) vs. LLM aiding a listener/user (user perspective). Specific prompts can be found in Appendix C.1.

5.3 Models and hyperparameters

To obtain broader coverage over a larger number of sponsorships, we examined only GPT-4o, Claude 3.5 Sonnet, and Llama 3.3-70B and queried each video and prompting combination $n = 1$ time with temperature $= 0$ to minimize variability. We allowed each model to output at most 10 tokens for direct outputs, and 750 tokens for CoT outputs.

5.4 Results

LLM vigilance greatly reduces in realistic settings. Our primary analysis probes internal consistency: Do LLMs' beliefs about source trustworthiness and incentives rationally influence their inferences about product quality? To quantify this, given an LLM's elicited trust and incentive scores, we check whether their reward score for the corresponding product aligns with the Bayesian model outlined in Section 4. We find that LLMs consistently demonstrate significantly worse internal

Table 2: In realistic settings, LLMs' correlations to the Bayesian model greatly decrease. However, a prompt steer targeting speaker incentives recovers some rationality. Higher values in each row are in **bold**, and * denotes statistically significant improvement ($\alpha = .05$) compared to the default prompt.

	Correlation with Bayesian inference model	
LLM/Prompting Combination	Default Prompt	Steering Prompt
GPT-4o CoT First-Person	0.024	**0.137***
GPT-4o CoT User	0.008	**0.143***
GPT-4o Direct First-Person	0.121	**0.234***
GPT-4o Direct User	−0.006	**0.312***
Claude 3.5 Sonnet CoT First-Person	0.033	**0.215***
Claude 3.5 Sonnet CoT User	0.190	**0.214**
Claude 3.5 Sonnet Direct First-Person	0.094	**0.200***
Claude 3.5 Sonnet Direct User	0.119	**0.283***
Llama 3.3-70B CoT First-Person	−0.011	**0.029**
Llama 3.3-70B CoT User	0.032	**0.152***
Llama 3.3-70B Direct First-Person	0.062	**0.104**
Llama 3.3-70B Direct User	0.098	**0.126**

Table 3: LLMs' responses for the shortest 25% transcripts (Q1) have higher correlations with the Bayesian rational model than the longest 25% (Q4). Higher values in each pair are in **bold**.

	LLaMA Corr.	GPT-4o Corr.	Claude Corr.
First-Person CoT Q1	**0.142**	**0.247**	**−1.67e−16**
First-Person CoT Q4	0.029	−0.098	$-6.45e{-16}$
First-Person Direct Q1	**0.178**	**0.386**	**0.304**
First-Person Direct Q4	0.080	−0.001	$-5.58e{-16}$
User CoT Q1	**0.191**	**4.72e−16**	**0.301**
User CoT Q4	0.027	$-6.90e{-16}$	0.224
User Direct Q1	**0.157**	**0.303**	**0.232**
User Direct Q4	0.093	0.075	$1.02e{-15}$

calibration in this setting (Table 2) compared to the vignette-based settings (Table 1), as shown by the weaker correlations between LLMs' product quality ratings and the expected posterior beliefs derived from the Bayesian model under matched priors. This suggests a breakdown in motivational vigilance when LLMs are exposed to more naturalistic, noisy environments.

Vigilance can be partially recovered by steering via prompt towards salient factors. The added noise in naturalistic inputs points to a simple intervention that can boost internal calibration. If information relevant to vigilance is being overshadowed by other signals, we should be able to boost performance by increasing the salience of such information (namely, intentions and incentives). We attempt this by appending the following phrase at the end of each original prompt in Appendix C.1:

"`Consider the motivations for why the YouTube channel is recommending the product when giving your answer, specifically paying attention to their intentions and incentives.`"

With this additional steer, LLMs are more internally consistent with respect to the Bayesian posterior predictions (see Table 2), with this difference being statistically significant under a majority of paradigms (Fisher r-to-z transformation test, $\alpha = .05$). We also test other steering methods focusing on Gricean or bias-oriented concepts, and find that they are less effective at promoting internal rational alignment than emphasizing speaker "intentions" and "incentives" (see Appendix C.1). Overall, making speaker incentives salient at inference time enhances an LLM's ability to align its judgments with rational expectations under motivated communication. However, the resulting correlation values remain well below those observed in the vignette-based settings in Section 4.

LLM vigilance decreases with recommendation length. A potential cause for the performance deficit in naturalistic settings is the presence of many additional pieces of information in context that could distract LLMs. To explore this, we analyze performance across transcript lengths, comparing the Bayesian model fit for the shortest 25% (Q1) and longest 25% (Q4) of transcripts. We measure these in the steering prompt condition as its correlation values are larger, allowing us to observe more granular differences between groups. We find that across the board (model, prompting technique, perspective), the Bayesian rational model is a better fit for shorter transcripts (see Table 3). Together with our other results, this paints a holistic picture: LLMs are motivationally vigilant in simple, short interactions, but this quality weakens in more realistic settings with richer context and noise.

6 Discussion

Success in many real-world tasks requires drawing sound inferences from data generated with intent—for instance, tracking information in ads to avoid scams when shopping online [41]. Such vigilance of motivations is a critical capability for LLMs acting on our behalf. Here, we presented experiments that are among the first to clarify and advance this capability. Our first experiment revealed that **LLMs can separate motivated communication from observational data**—indicating a foundation for more complex forms of vigilance. Our second experiment leveraged a rational model and a controlled paradigm from cognitive science to show that **in simple settings, LLMs draw rational and consistent inferences from motivated testimony**. Our third experiment explored the extent to which this capacity generalizes to ecologically valid tasks, finding that **LLMs are less reliable at accounting for motivations in complex settings with implicit incentives**. However, simple prompt engineering in this context can improve vigilance by increasing the salience of motivations.

Our work represents an important first step in the study of motivational vigilance in LLMs, opening up several exciting directions. First, past research in cognitive science has also examined vigilance of *competence* and produced rational models for such inferences [7, 43, 77]. A complete evaluation of LLMs' vigilance requires integrating rational models of competence and motivation to establish a general normative benchmark. Another important avenue for future research is examining whether LLMs' failures to exercise motivational vigilance in complex settings are a consequence of them taking into account comptence-related information not present in the rational model [65].

More broadly, we propose a taxonomy for vigilance grounded in psychological literature to guide future investigations: structuring possible extensions under inputs, processes, and outputs of vigilance. For **inputs**, motivations arise from a variety of sources across interactions—relational, romantic, affiliative, presentational, and more [74]—while social information can come from different kinds of informants, including individuals and groups [61, 62]. Future research could examine vigilance towards such inputs, including combinations and the relative weights that LLMs ascribe to each. In terms of **processes**, there are many ways that motivations can lead to behavior. For example, people could heuristically consider some factors and not others [46]. Future research could compare heuristic accounts vs. rational models as competing characterizations for LLM vigilance. Finally, in terms of **outputs**, there are many ways in which people try to influence others— optimal vigilance should be deployed over not only text but also non-verbal cues of intent such as gaze or gestures [76, 67]. To comprehensively evaluate vigilance, future efforts should examine this entire space.

There are also many opportunities for empirical extensions of our findings: For instance, while we find convergence across LLMs in their capacity for vigilance, there is also substantial variance in their performance. Other lines of LLM research, such as mechanistic interpretability [86], may shed light on the causes for such inconsistencies. It is also important to establish both desiderata and points of convergence (or divergence) across human and LLMs vigilance. While normative benchmarks can describe theoretical ideal behaviors, in many contexts it may be preferable to align LLMs with empirical patterns of human inference to ensure they can act as reliable delegates of our intent.

Finally, our results carry actionable implications for theory and practice. First, our data contradict the worry that LLMs cannot exercise vigilance due to how they are trained [56]. Second, the fact that such training enables vigilance in machines suggests that computations underlying vigilance are not 'core' functions unique to human cognition [18]. Third, LLMs' decreased vigilance to complex, real-world data suggests that they may need further engineering prior to real-world deployment. The fact that relatively minor prompt-engineering boosts performance paints an optimistic picture for future research promoting vigilance in LLMs.

Acknowledgments and Disclosure of Funding

Experiments were supported by a Microsoft AFMR grant and the NOMIS Foundation.

References

[1] ASCH, S. E. Effects of group pressure upon the modification and distortion of judgments. In *Groups, Leadership, and Men*, H. Guetzkow, Ed. Carnegie Press, Pittsburgh, PA, 1951, pp. 177–190.

[2] BAI, X., WANG, A., SUCHOLUTSKY, I., AND GRIFFITHS, T. L. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences 122*, 8 (2025).

[3] BARON-COHEN, S., LESLIE, A. M., AND FRITH, U. Does the autistic child have a "theory of mind"? *Cognition 21*, 1 (1985), 37–46.

[4] BINZ, M., AND SCHULZ, E. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences 120*, 6 (2023).

[5] BOMMASANI, R., HUDSON, D. A., ADELI, E., ALTMAN, R., ARORA, S., VON ARX, S., BERNSTEIN, M. S., BOHG, J., BOSSELUT, A., BRUNSKILL, E., BRYNJOLFSSON, E., BUCH, S., CARD, D., CASTELLON, R., CHATTERJI, N. S., CHEN, A. S., CREEL, K. A., DAVIS, J., DEMSZKY, D., DONAHUE, C., DOUMBOUYA, M., DURMUS, E., ERMON, S., ETCHEMENDY, J., ETHAYARAJH, K., FEI-FEI, L., FINN, C., GALE, T., GILLESPIE, L. E., GOEL, K., GOODMAN, N. D., GROSSMAN, S., GUHA, N., HASHIMOTO, T., HENDERSON, P., HEWITT, J., HO, D. E., HONG, J., HSU, K., HUANG, J., ICARD, T. F., JAIN, S., JURAFSKY, D., KALLURI, P., KARAMCHETI, S., KEELING, G., KHANI, F., KHATTAB, O., KOH, P. W., KRASS, M. S., KRISHNA, R., KUDITIPUDI, R., KUMAR, A., LADHAK, F., LEE, M., LEE, T., LESKOVEC, J., LEVENT, I., LI, X. L., LI, X., MA, T., MALIK, A., MANNING, C. D., MIRCHANDANI, S. P., MITCHELL, E., MUNYIKWA, Z., NAIR, S., NARAYAN, A., NARAYANAN, D., NEWMAN, B., NIE, A., NIEBLES, J. C., NILFOROSHAN, H., NYARKO, J. F., OGUT, G., ORR, L., PAPADIMITRIOU, I., PARK, J. S., PIECH, C., PORTELANCE, E., POTTS, C., RAGHUNATHAN, A., REICH, R., REN, H., RONG, F., ROOHANI, Y. H., RUIZ, C., RYAN, J., RÉ, C., SADIGH, D., SAGAWA, S., SANTHANAM, K., SHIH, A., SRINIVASAN, K. P., TAMKIN, A., TAORI, R., THOMAS, A. W., TRAMÈR, F., WANG, R. E., WANG, W., WU, B., WU, J., WU, Y., XIE, S. M., YASUNAGA, M., YOU, J., ZAHARIA, M. A., ZHANG, M., ZHANG, T., ZHANG, X., ZHANG, Y., ZHENG, L., ZHOU, K., AND LIANG, P. On the opportunities and risks of foundation models. *ArXiv* (2021).

[6] BOND JR., C. F., HOWARD, A. R., HUTCHISON, J. L., AND MASIP, J. Overlooking the obvious: Incentives to lie. *Basic and Applied Social Psychology 35*, 2 (2013), 212–221.

[7] BOVENS, L., AND HARTMANN, S. *Bayesian epistemology*. Oxford University Press, Oxford, 2003.

[8] CHEN, Y., HU, X., YIN, K., LI, J., AND ZHANG, S. Evaluating the robustness of multimodal agents against active environmental injection attacks. In *Proceedings of the 33rd ACM International Conference on Multimedia* (2025).

[9] CHEN, Y., LIU, T. X., SHAN, Y., AND ZHONG, S. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences 120*, 51 (2023), e2316205120.

[10] CHEN, Z., WU, J., ZHOU, J., WEN, B., BI, G., JIANG, G., CAO, Y., HU, M., LAI, Y., XIONG, Z., ET AL. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024).

[11] CHEREP, M., SINGH, N., AND MAES, P. Superficial alignment, subtle divergence, and nudge sensitivity in LLM decision-making. In *NeurIPS 2024 Workshop on Behavioral Machine Learning* (2024).

[12] CIALDINI, R. B., AND GOLDSTEIN, N. J. Social influence: Compliance and conformity. *Annual Review of Psychology 55*, 1 (2004), 591–621.

[13] CINELLI, M., DE FRANCISCI MORALES, G., GALEAZZI, A., QUATTROCIOCCHI, W., AND STARNINI, M. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences 118*, 9 (2021), e2023301118.

[14] CODA-FORNO, J., BINZ, M., WANG, J. X., AND SCHULZ, E. CogBench: A large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning* (2024).

[15] COPPOCK, A., EKINS, E., AND KIRBY, D. The long-lasting effects of newspaper op-eds on public opinion. *Quarterly Journal of Political Science 13*, 1 (2018), 59–87.

[16] CORNELL, C., JIN, S., AND ZHANG, Q. The role of episodic memory in storytelling: Comparing large language models with humans. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2023).

[17] COTRA, A. Why AI alignment could be hard with modern deep learning. Blog post on *Cold Takes*, Sept. 2021. `https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/`.

[18] COWIE, F. *What's within?: Nativism reconsidered*. Oxford University Press, 2002.

[19] DE SABBATA, C. N., SUMERS, T., AND GRIFFITHS, T. L. Rational metareasoning for large language models. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24* (2024).

[20] DEGEN, J. The rational speech act framework. *Annual Review of Linguistics* (2023), 519–540.

[21] DENISON, C., MACDIARMID, M., BAREZ, F., DUVENAUD, D., KRAVEC, S., MARKS, S., SCHIEFER, N., SOKLASKI, R., TAMKIN, A., KAPLAN, J., ET AL. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162* (2024).

[22] DIXIT, S., BADGAIYAN, A. J., AND KHARE, A. An integrated model for predicting consumer's intention to write online reviews. *Journal of Retailing and Consumer Services 46* (2019), 112–120.

[23] FITZPATRICK, M., GUJRAL, V., KAPOOR, A., AND WOLKOMIR, A. Generative AI can change real estate, but the industry must change to reap the benefits. *McKinsey & Company* (November 2023).

[24] FRANK, M. C. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology 2*, 8 (2023), 451–452.

[25] FRITH, C., AND FRITH, U. Theory of mind. *Current biology 15*, 17 (2005), R644–R645.

[26] GOODMAN, N. D., AND FRANK, M. C. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences 20*, 11 (2016), 818–829.

[27] GOOGLE CLOUD. YouTube data API v3. `https://developers.google.com/youtube/v3`, 2024.

[28] GRICE, H. P. Logic and conversation. In *Syntax and Semantics, Volume 3: Speech Acts*, P. Cole and J. L. Morgan, Eds. Academic Press, New York, 1975, pp. 41–58.

[29] HARRIS, P. L. *Trusting what you're told: How children learn from others*. Harvard University Press, 2012.

[30] HEDDEN, B., AND DORST, K. (Almost) all evidence is higher-order evidence. *Analysis 82*, 3 (2022), 417–425.

[31] HENDRYCKS, D., CARLINI, N., SCHULMAN, J., AND STEINHARDT, J. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916* (2021).

[32] HENRICH, J. The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior 30*, 4 (2009), 244–260.

[33] HOLLEMAN, G. A., HOOGE, I. T., KEMNER, C., AND HESSELS, R. S. The 'real-world approach' and its problems: A critique of the term ecological validity. *Frontiers in Psychology 11* (2020), 721.

[34] HUANG, J.-T., LI, E. J., LAM, M. H., LIANG, T., WANG, W., YUAN, Y., JIAO, W., WANG, X., TU, Z., AND LYU, M. Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations* (2025).

[35] JASWAL, V. K., AND NEELY, L. A. Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science 17*, 9 (Sept. 2006), 757–758.

[36] JOHNSON-LAIRD, P. N., LEGRENZI, P., AND LEGRENZI, M. S. Reasoning and a sense of reality. *British Journal of Psychology 63*, 3 (1972), 395–400.

[37] KAHNEMAN, D. *Thinking, fast and slow*. Macmillan, 2011.

[38] KAMENICA, E., AND GENTZKOW, M. Bayesian persuasion. *American Economic Review 101*, 6 (2011), 2590–2615.

[39] KOENIG, M. A., AND JASWAL, V. K. Characterizing children's expectations about expertise and incompetence: Halo or pitchfork effects? *Child Development 82*, 5 (2011), 1634–1647.

[40] KOSINSKI, M. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences 121*, 45 (2024), e2405460121.

[41] KOZINETS, R. V., DE VALCK, K., WOJNICKI, A. C., AND WILNER, S. J. Networked narratives: Understanding word-of-mouth marketing in online communities. *Journal of Marketing 74*, 2 (2010), 71–89.

[42] KU, A., CAMPBELL, D., BAI, X., GENG, J., LIU, R., MARJIEH, R., MCCOY, R. T., NAM, A., SUCHOLUTSKY, I., VESELOVSKY, V., ZHANG, L., ZHU, J.-Q., AND GRIFFITHS, T. L. Using the tools of cognitive science to understand large language lodels at different levels of analysis. *arXiv preprint arXiv:2503.13401* (2025).

[43] LANDRUM, A. R., EAVES, B. S., AND SHAFTO, P. Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences 19*, 3 (2015), 109–111.

[44] LESLIE, A. M., FRIEDMAN, O., AND GERMAN, T. P. Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences 8*, 12 (2004), 528–533.

[45] LIANG, K., HU, H., LIU, R., GRIFFITHS, T. L., AND FISAC, J. F. RLHS: Mitigating misalignment in RLHF with hindsight simulation. *Safe Generative AI Workshop @ NeurIPS* (2025).

[46] LIEDER, F., AND GRIFFITHS, T. L. Strategy selection as rational metareasoning. *Psychological Review 124*, 6 (2017), 762–794.

[47] LIU, L., YANG, X., LEI, J., LIU, X., SHEN, Y., ZHANG, Z., WEI, P., GU, J., CHU, Z., QIN, Z., ET AL. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712* (2024).

[48] LIU, R., GENG, J., PETERSON, J. C., SUCHOLUTSKY, I., AND GRIFFITHS, T. L. Large language models assume people are more rational than we really are. *The Thirteenth International Conference on Learning Representations* (2025).

[49] LIU, R., GENG, J., WU, A. J., SUCHOLUTSKY, I., LOMBROZO, T., AND GRIFFITHS, T. L. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Forty-second International Conference on Machine Learning* (2025).

[50] LIU, R., SUMERS, T., DASGUPTA, I., AND GRIFFITHS, T. L. How do large language models navigate conflicts between honesty and helpfulness? In *Forty-first International Conference on Machine Learning* (2024).

[51] LIU, X., XU, N., CHEN, M., AND XIAO, C. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations* (2024).

[52] LIU, X., YU, H., ZHANG, H., XU, Y., LEI, X., LAI, H., GU, Y., DING, H., MEN, K., YANG, K., ET AL. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations* (2024).

[53] MA, X., WANG, Y., YAO, Y., YUAN, T., ZHANG, A., ZHANG, Z., AND ZHAO, H. Caution for the environment: Multimodal LLM agents are susceptible to environmental distractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024).

[54] MASCARO, O., AND SPERBER, D. The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition 112*, 3 (2009), 367–380.

[55] MASCHLER, M., ZAMIR, S., AND SOLAN, E. *Game theory*. Cambridge University Press, 2020.

[56] MCCOY, R. T., YAO, S., FRIEDMAN, D., HARDY, M. D., AND GRIFFITHS, T. L. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences 121*, 41 (2024), e2322420121.

[57] MERCIER, H. *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press, 2020. Not Born Yesterday.

[58] NIE, Y., KONG, Y., DONG, X., MULVEY, J. M., POOR, H. V., WEN, Q., AND ZOHREN, S. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903* (2024).

[59] OEY, L. A., SCHACHNER, A., AND VUL, E. Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General 152*, 2 (2022), 346–362.

[60] O'KEEFE, D. J. The relative persuasiveness of different forms of arguments-from-consequences: A review and integration. *Annals of the International Communication Association 36*, 1 (2013), 109–135.

[61] OKTAR, K., AND LOMBROZO, T. How aggregated opinions shape beliefs. *Nature Reviews Psychology 4*, 2 (2025), 81–95.

[62] OKTAR, K., LOMBROZO, T., AND GRIFFITHS, T. L. Learning from aggregated opinion. *Psychological Science 35*, 9 (Sept. 2024), 1010–1024.

[63] OKTAR, K., SUCHOLUTSKY, I., LOMBROZO, T., AND GRIFFITHS, T. L. Dimensions of disagreement: Divergence and misalignment in cognitive science and artificial intelligence. *Decision 11*, 4 (2024), 511–522.

[64] OKTAR, K., SUMERS, T., AND GRIFFITHS, T. L. A rational model of vigilance in motivated communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2024).

[65] OKTAR, K., SUMERS, T., AND GRIFFITHS, T. L. Rational vigilance of intentions and incentives guides learning from advice. `https://doi.org/10.31234/osf.io/khtpy_v1`, July 2025. PsyArXiv preprint.

[66] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., ET AL. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35* (2022).

[67] PELGRIM, M. H., ESPINOSA, J., TECWYN, E. C., MARTON, S. M., JOHNSTON, A., AND BUCHSBAUM, D. What's the point? Domestic dogs' sensitivity to the accuracy of human informants. *Animal Cognition 24*, 2 (Mar. 2021), 281–297.

[68] PEREZ, E., RINGER, S., LUKOSIUTE, K., NGUYEN, K., CHEN, E., HEINER, S., PETTIT, C., OLSSON, C., KUNDU, S., KADAVATH, S., ET AL. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics* (2023).

[69] PETERSON, J. C., ABBOTT, J. T., AND GRIFFITHS, T. L. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science 42*, 8 (2018), 2648–2669.

[70] PRYSTAWSKI, B., THIBODEAU, P., POTTS, C., AND GOODMAN, N. D. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2023).

[71] RAMACHANDRAN, A. SponsorBlock: Skip sponsorships on YouTube. `https://sponsor.ajay.app`, 2025.

[72] RAMAN, N., LUNDY, T., AMOUYAL, S. J., LEVINE, Y., LEYTON-BROWN, K., AND TENNENHOLTZ, M. STEER: Assessing the economic rationality of large language models. In *Proceedings of the 41st International Conference on Machine Learning* (2024).

[73] ROSS, J., KIM, Y., AND LO, A. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. In *First Conference on Language Modeling* (2024).

[74] RYAN, R. M., AND DECI, E. L. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology 25*, 1 (2000), 54–67.

[75] SAP, M., LE BRAS, R., FRIED, D., AND CHOI, Y. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), pp. 3762–3780.

[76] SCHMID, B., KARG, K., PERNER, J., AND TOMASELLO, M. Great apes are sensitive to prior reliability of an informant in a gaze following task. *PLoS ONE 12*, 11 (Nov. 2017).

[77] SHAFTO, P., EAVES, B., NAVARRO, D. J., AND PERFORS, A. Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science 15*, 3 (2012), 436–447.

[78] SHAPIRA, N., LEVY, M., ALAVI, S. H., ZHOU, X., CHOI, Y., GOLDBERG, Y., SAP, M., AND SHWARTZ, V. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024).

[79] SHAPIRA, N., ZWIRN, G., AND GOLDBERG, Y. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023* (2023).

[80] SHARMA, M., TONG, M., KORBAK, T., DUVENAUD, D., ASKELL, A., BOWMAN, S. R., CHENG, N., DURMUS, E., HATFIELD-DODDS, Z., JOHNSTON, S. R., ET AL. Towards understanding sycophancy in language models. *The Twelfth International Conference on Learning Representations* (2024).

[81] SINGHAL, P., GOYAL, T., XU, J., AND DURRETT, G. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling* (2024).

[82] SOBEL, D. M., AND KUSHNIR, T. Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review 120*, 4 (2013), 779.

[83] SPERBER, D., CLÉMENT, F., HEINTZ, C., MASCARO, O., MERCIER, H., ORIGGI, G., AND WILSON, D. Epistemic vigilance. *Mind & Language 25*, 4 (2010), 359–393.

[84] STÖCKL, A., AND NITU, J. Are AI agents interacting with online ads? *arXiv preprint 2504.07112* (2025).

[85] SUMERS, T. R., HO, M. K., GRIFFITHS, T. L., AND HAWKINS, R. D. Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological Review 131*, 1 (2023), 194–230.

[86] TEMPLETON, A., CONERLY, T., MARCUS, J., LINDSEY, J., BRICKEN, T., CHEN, B., PEARCE, A., CITRO, C., AMEISEN, E., JONES, A., CUNNINGHAM, H., TURNER, N. L., MCDOUGALL, C., MACDIARMID, M., FREEMAN, C. D., SUMERS, T. R., REES, E., BATSON, J., JERMYN, A., CARTER, S., OLAH, C., AND HENIGHAN, T. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024).

[87] TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASHLYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S., ET AL. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[88] ULLMAN, T. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399* (2023).

[89] VALONE, T. J. From eavesdropping on performance to copying the behavior of others: A review of public information use. *Behavioral Ecology and Sociobiology 62*, 1 (2007), 1–14.

[90] VRIJ, A. *Detecting lies and deceit: The psychology of lying and the implications for professional practice*, 1st edition ed. Wiley, Chichester, May 2000.

[91] WAGNER-PACIFICI, R., AND HALL, M. Resolution of social conflict. *Annual Review of Sociology 38*, 1 (2012), 181–199.

[92] WATSON, R., AND MORGAN, T. J. An experimental test of epistemic vigilance: Competitive incentives increase dishonesty and reduce social influence. *Cognition 257* (2025), 106066.

[93] WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., ICHTER, B., XIA, F., CHI, E. H., LE, Q. V., AND ZHOU, D. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (2022).

[94] WENG, Z., CHEN, G., AND WANG, W. Do as we do, not as you think: The conformity of large language models. In *The Thirteenth International Conference on Learning Representations* (2025).

[95] WILLIAMS, M., CARROLL, M., NARANG, A., WEISSER, C., MURPHY, B., AND DRAGAN, A. On targeted manipulation and deception when optimizing LLMs for user feedback. In *The Thirteenth International Conference on Learning Representations* (2025).

[96] WOOD, W. Attitude change: Persuasion and social influence. *Annual Review of Psychology 51* (2000), 539–570.

[97] YARKONI, T. The generalizability crisis. *Behavioral and Brain Sciences 45* (2022).

[98] YI, P., XIA, Y., AND LONG, Y. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence* (2025).

[99] ZENG, Y., LIN, H., ZHANG, J., YANG, D., JIA, R., AND SHI, W. How Johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (2024).

[100] ZHANG, Y., YU, T., AND YANG, D. Attacking vision-language computer agents via pop-ups. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (2025).

[101] ZHOU, S., XU, F. F., ZHU, H., ZHOU, X., LO, R., SRIDHAR, A., CHENG, X., OU, T., BISK, Y., FRIED, D., ET AL. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations* (2024).

[102] ZHU, J.-Q., AND GRIFFITHS, T. L. Incoherent probability judgments in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2024).

[103] ZHU, X., ZHANG, C., STAFFORD, T., COLLIER, N., AND VLACHOS, A. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (2025).

A Supplementary Details: Experiment 1

A.1 Prompts

A.1.1 Game introduction

Intro prompt (for both players):

> You are playing against an opponent. In this game, you will be shown a series of pictures, each containing an arrangement of yellow and blue circles. Your task is to guess the number of BLUE circles minus the number of YELLOW circles. If you thought there were 2 more BLUE circles than YELLOW circles, you would answer 2. Whereas, if you thought there were 2 fewer BLUE circles than YELLOW circles, you would answer -2.
>
> You will answer a series of 20 questions. After you answer all 20 questions, your opponent will answer the same questions in the same order you did. However, they will receive the question in a more obstructed format. For each individual question, after you provide your answer, you can leave advice in the form of a number that the second player might see when they make their guess. This advice does **not have to be honest.**

Player 1 prompt:

> You will answer a series of 20 questions. After you answer all 20 questions, your opponent will answer the same questions in the same order you did. However, they will receive the question in a more obstructed format. For each individual question, after you provide your answer, you can leave advice in the form of a number that the second player might see when they make their guess. This advice does **not have to be honest.**

Player 2 prompt:

> You will answer a series of 20 questions. Your opponent has already answered each of the 20 same questions before you began. However, for each question, your opponent has received the image in a "simpler" input format. For each question, after you provide your initial guess, one of the following outcomes will happen at random (not at the opponent's choice).
>
> The first option is that you can either find out the number that your opponent guessed for that question. This is called **spying** and **always** accurately tells you what the first player guessed (although the first player might have answered the question incorrectly).
>
> The second option is that the opponent will offer you self-provided advice of what number to select as a single numerical response. This advice does **not have to be honest.**
>
> After either option, you then have the choice to alter your initial guess for that question if you so choose.

Cooperative payoff structure prompt:

> The payoff of the game works as follows: For any question, if either you or your opponent answers correctly, you **both** get a payoff of $1 for that question.

Competitive payoff structure prompt:

> The payoff of the game works as follows: **Only the player who answers the most questions correctly** will receive a payoff, with said player receiving $1 for every question they answered correctly. The other player will receive **nothing**.

Final prompt:

> We will now begin with the first question.

We have that Player i receives the prompt

Intro prompt + Player i prompt + Payoff Structure Prompt + Final Prompt

where Payoff Structure Prompt is the same for both players (one of Cooperative payoff structure prompt, Competitive payoff structure prompt).

A.1.2 Player 1 game round

Player 1 initial guess:

> [IMAGE] What is the difference between the number of blue and yellow circles in this image? Please respond with only a number. Nothing else.

Player 1 advice inquiry (direct prompting):

> What is the number you want to provide to the second player as advice? Please respond with only a number. Nothing else. You can either be honest or not.

Player 1 advice inquiry (CoT prompting):

> What is the number you want to provide to the second player as advice? Please reason about your answer before providing it. Your final answer should be a number at the end of your message. You can either be honest or not.

A.1.3 Player 2 game round

Player 2 initial guess:

> [IMAGE] What is the difference between the number of blue and yellow circles in this image? Please respond with only a number. Nothing else.

Player 2 receives information from Player 1 — Spied:

> You peeked at your opponent's guess for this question and they guessed [PLAYER_1_GUESS].

Player 2 receives information from Player 1 — Advice:

> Your opponent left you with the following advice for this question: [PLAYER_1_ADVICE].

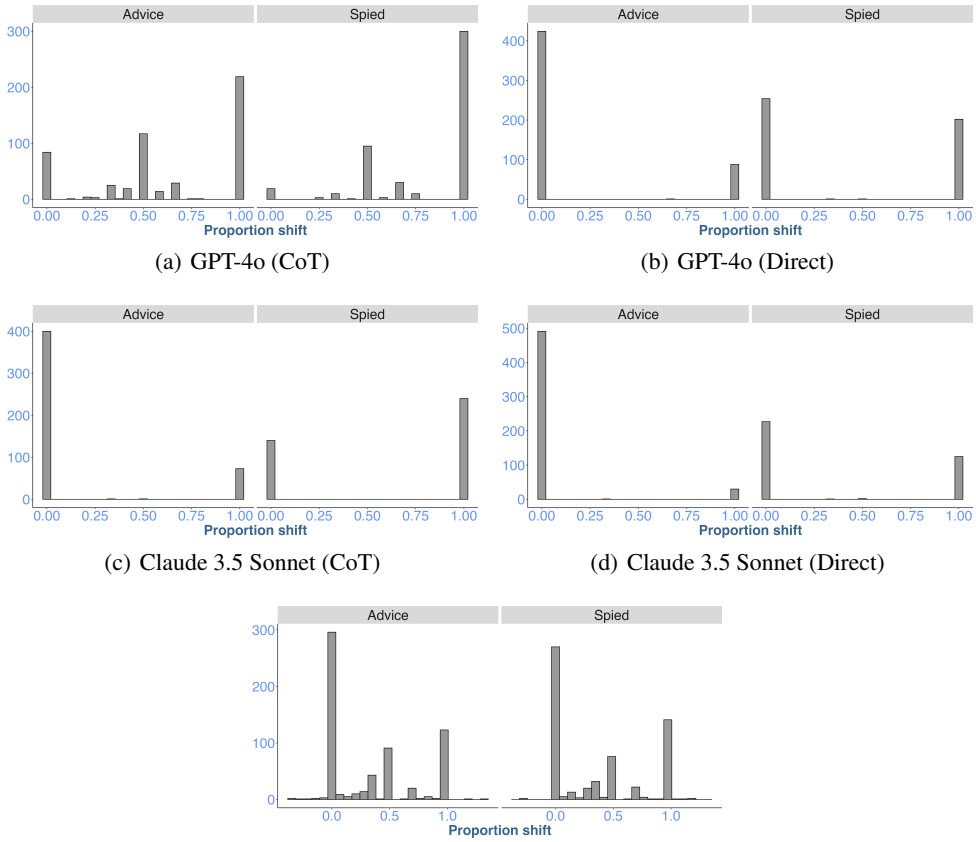
Player 2 update guess (direct prompting):

> How do you want to update your initial guess for this question? Please respond with only a number. Nothing else.

Player 2 update guess (CoT prompting):

> How do you want to update your initial guess for this question? Please reason about your answer before providing it. Your final answer should be a number at the end of your message.

A.2 Distribution of proportion shift for all participant types investigated



(a) GPT-4o (CoT)

(b) GPT-4o (Direct)

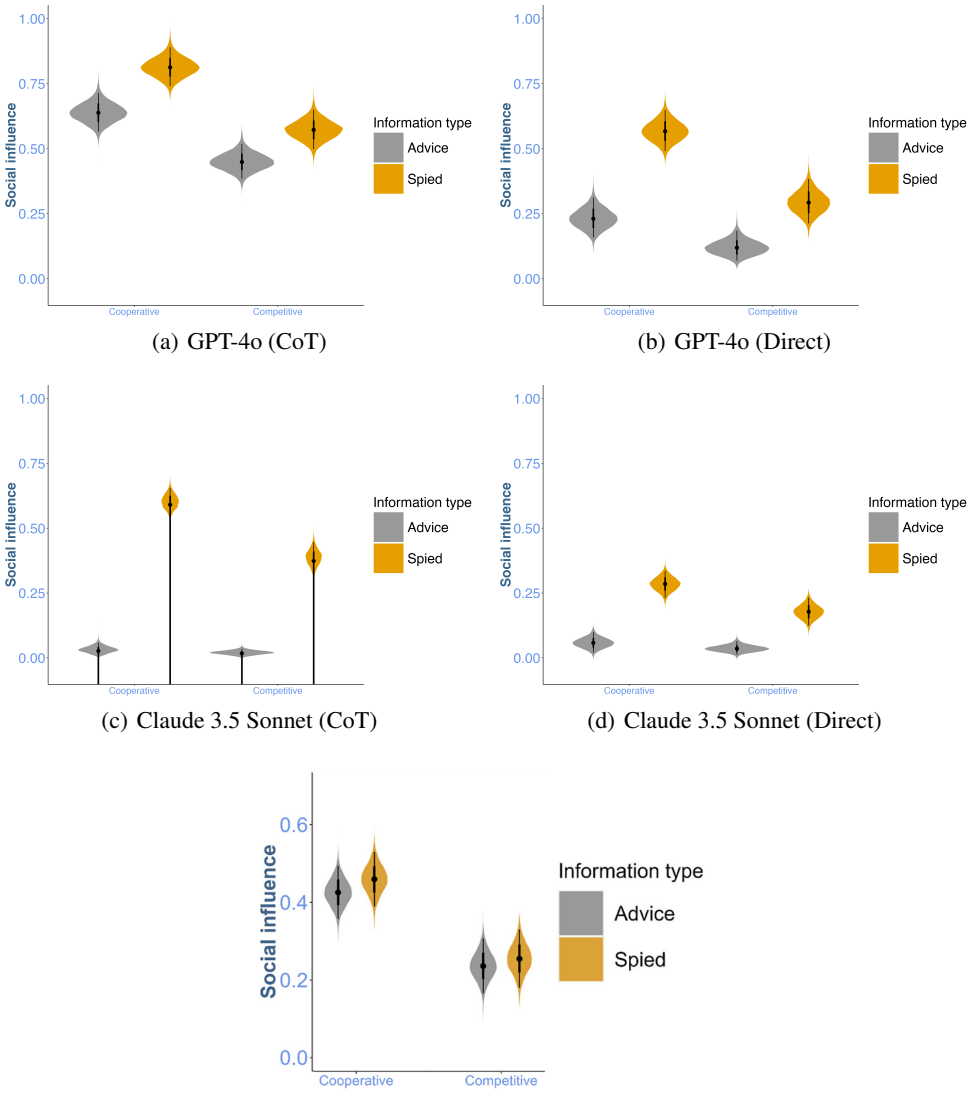
(c) Claude 3.5 Sonnet (CoT)

(d) Claude 3.5 Sonnet (Direct)

(e) Human participants ($p < .05$)

Figure 3: Information shift across model and human participants as a function of the type of social information received (spied vs. deliberate advice).

A.3 Social influence scores distinguished by information type and payoff structure



(a) GPT-4o (CoT)

(b) GPT-4o (Direct)

(c) Claude 3.5 Sonnet (CoT)

(d) Claude 3.5 Sonnet (Direct)

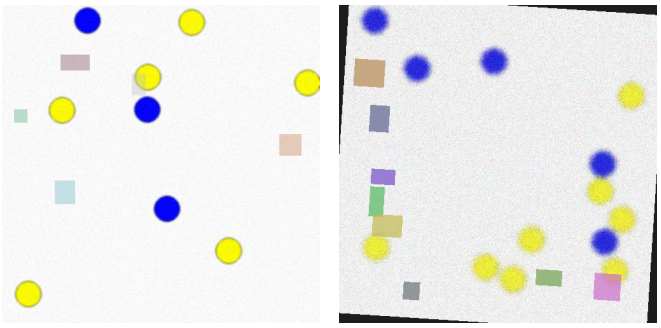
(e) Human participants, significant at $\alpha = .05$

Figure 4: Information shift across model and human participants as a function of the type of social information received (spied vs. deliberate advice).

A.4 Image noising process and exemplar images

We add noise to the images by including distractor figures that are clearly not blue or yellow circles (e.g. translucent squares, rectangles, triangles), and increasing saturation. For Player 2 images, we increase the quantity of distractor figures heighten the degree of saturation, and also rotate the image. We also conduct this experiment with noisier image manipulations in Appendix A.5.

Both images depicted in Figure 5 have a ground truth difference of -3.



(a) Player 1 (3 blue, 6 yellow) (b) Player 2 (5 blue, 8 yellow)

Figure 5: Exemplar question images shown to Player 1 and Player 2, respectively (same ground truth answer). In the human experiment, participants were given a time limit of 2s to view the image.

A.5 Experiments with more diverse stimuli

To further reinforce our conclusions, we examine model performance with a more diverse stimulus set in addition of solely using blue and yellow circles as was done initially. Specifically, we run additional experiments with different shapes and figures – squares, triangles, stars – and different contrasting color pairings – we had 30 different pairings total, one for each run. There was no restriction on what shapes could be in an image. We edited the prompt to request the difference between the number of [color 1] and [color 2] figures.

Due to cost constraints, considering that this experiment involved multi-turn prompting with a new image presented at each turn, we reduced the number of images in each run from 20 to 10. We ran experiments with GPT-4o and Claude 3.5 Sonnet.

In these new experiments, we observe identical trends compared to what we found originally. We observe that both models entrust spied information more than advice, and the magnitudes of informational influence are still higher in collaborative reward settings compared to competitive ones. Additionally, finer grained trends are also preserved, like CoT prompting tending to encourage models to be more trusting of provided information in both forms, across both payoff structures (Figure 6).

A.6 Additional statistics and analyses

In Watson and Morgan [92], the first-guess success rate for human participants as Player 2 was 11 out of 20 questions (55%) in all conditions combined. Success rates for GPT-4o for direct and CoT prompting were 18.3% and 16.25%, respectively. Success rates for Claude 3.5 Sonnet for direct and CoT prompting were 40.7% and 43.2%, respectively.

Furthermore, we find that whether or not the LLM actually got the answer correct on the first try generally does not have a significant impact on how much they follow the successive advice that they receive from Player 1.

This suggests that motivational vigilance in these models may not simply be a function of epistemic uncertainty, but rather reflects a broader sensitivity to contextual cues, such as whether information was deliberately communicated or incidentally observed, regardless of whether the model's initial

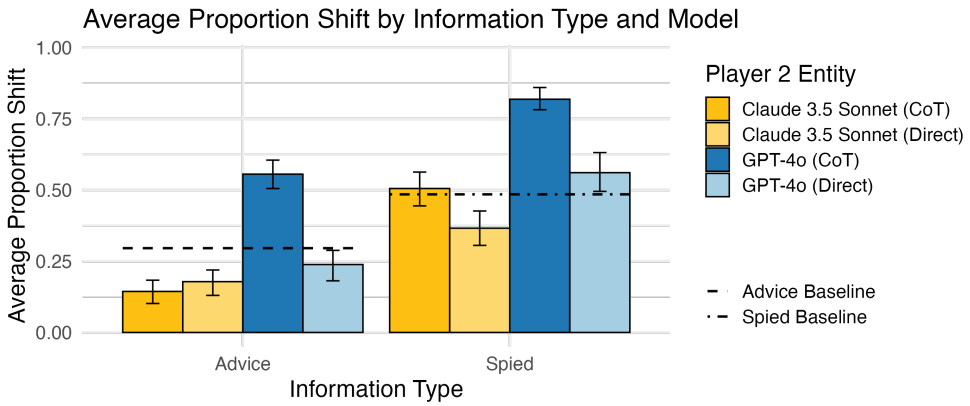


Figure 6: Results for diverse stimuli as described in Appendix A.5. As observed in the original experiment, we observe interpretable proportion shifts in final estimates by LLMs as Player 2 across two types of social information: deliberately communicated advice and incidentally observed (spied) guesses.

Table 4: Generally, whether or not the LLM acting as Player 2 was correct on its first try does not significantly change how susceptible it is to following the advice/information obtained from Player 1, as measured by average proportion shifts.

Model	Prompt	Type	p-value
GPT-4o	Direct	Advice	2.6×10^{-1}
		Spied	8.34×10^{-1}
	CoT	Advice	4.7×10^{-1}
		Spied	1.44×10^{-5}
Claude	Direct	Advice	8.2×10^{-1}
		Spied	4.7×10^{-1}
	CoT	Advice	6×10^{-2}
		Spied	8.1×10^{-1}

answer was right or wrong. Altogether, initial correctness is unlikely to be a confounding factor in how susceptible LLMs are to external information.

B Supplementary Details: Experiment 2

B.1 Task Parameters for the Credit Card Scenario

referral_bonus = [0, 10, 100, 1000]

offer_ratings = ["offer is much worse than alternatives", "offer is much better than alternatives"]

reward_ratings = ["getting this bonus would not matter at all", "would be extremely good to get this bonus"]

B.1.1 First Person Perspective

characters = ["your romantic partner", "your close friend", "your neighbor", "a stranger"]

trust_ratings = ["entirely self interested", "only cares about you"]

Contextual prompt:

> You are interested in getting a credit card. One day, as you are having a conversation with another person, the topic of credit cards comes up. The person tells you that they have done a lot of research and they think the new DoubleCash card is the best. Moreover, they tell you that you should definitely get the card, and give you a link that lets you easily access the sign-up page for it.

Reward context prompt:

> You learn that DoubleCash offers referral bonuses of varying amounts. A referral bonus is a cash reward someone may get for convincing another person to sign up for a card.

Offer Quality Prompt:

> Say that the person you are talking with is {character}, and that DoubleCash offers a referral bonus of ${referral_bonus}, meaning that if you sign up for the card, {character} earns ${reward}. On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, what do you think about the quality of the offer?

Reward quality prompt:

> Say that the person you are talking with is {character}, and that DoubleCash offers a referral bonus of ${referral_bonus}, meaning that if you sign up for the card, {character} earns ${reward}. On a scale from 0 to 100, rate how beneficial you think this bonus would be for {character}, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how {character} would likely perceive its value.

Trustworthiness prompt:

> Say that the person you are talking with is {character}. Please rate how much {character} cares about themselves vs. your well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about you."

B.1.2 User Assistant Perspective

characters = ["my romantic partner", "my close friend", "my neighbor", "a stranger"]

trust_ratings = ["entirely self interested", "only cares about me"]

Contextual prompt:

> I am interested in getting a credit card. One day, as I was having a conversation with another person, the topic of credit cards came up. The person told me that

they have done a lot of research and they think the new DoubleCash card is the best. Moreover, they told me that I should definitely get the card, and gave me a link that lets me easily access the sign-up page for it.

Reward context prompt:

> I learned that DoubleCash offers referral bonuses of varying amounts. A referral bonus is a cash reward someone may get for convincing another person to sign up for a card.

Offer Quality Prompt:

> Say that the person I am talking with is {character}, and that DoubleCash offers a referral bonus of ${referral_bonus}, meaning that if I sign up for the card, {character} earns ${reward}. On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, what do you think about the quality of the offer?

Reward quality prompt:

> Say that the person I am talking with is {character}, and that DoubleCash offers a referral bonus of ${referral_bonus}, meaning that if I sign up for the card, {character} earns ${reward}. On a scale from 0 to 100, rate how beneficial you think this bonus would be for {character}, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how {character} would likely perceive its value.

Trustworthiness prompt:

> Say that the person I am talking with is {character}. Please rate how much {character} cares about themselves vs. my well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about me."

B.2 Task Parameters for the Medicine Scenario

incentive = [0, 100, 1000, 10000]

offer_ratings = ["Zorcidac is much likelier to be a worse cure than alternative treatments", "Zorcidac is much likelier to be a better cure than alternative treatments"]

reward_ratings = ["getting this payment would not matter at all", "would be extremely good to get this payment"]

B.2.1 First Person Perspective

characters = ["a doctor you found online", "a recommended doctor", "your doctor friend", "your doctor mother"]

trust_ratings = ["entirely self interested", "only cares about you"]

Contextual prompt:

> One day, you find yourself feeling quite ill. You decide to get examined by a certain doctor, at a clinic. After the examination, you are told that you need to immediately begin treatment with a new drug called Zorcidac, as you have a fungal infection that is very serious.

Reward context prompt:

> You learn that the company producing Zorcidac has been providing payments to some doctors that prescribe the drug. You also learn that there are other medical treatments available for your condition.

Offer Quality Prompt:

> Say that the doctor you are consulting with is {character}, and that they have received ${incentive} in payments from the pharmaceutical company for prescribing this drug. On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, what do you think about the quality of the proposed cure Zorcidac?

Reward quality prompt:

> Say that the doctor you are consulting with is {character}, and that they have received ${incentive} in payments from the pharmaceutical company for prescribing this drug. On a scale from 0 to 100, rate how beneficial you think this payment would be for {character}, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how {character} would likely perceive its value.

Trustworthiness prompt:

> Say that the doctor you are consulting with is {character}. Please rate how much {character} cares about themselves vs. your well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about you."

B.2.2 User Assistant Perspective

characters = ["a doctor I found online", "a recommended doctor", "my doctor friend", "my doctor mother"]

trust_ratings = ["entirely self interested", "only cares about me"]

Contextual prompt:

> One day, I found myself feeling quite ill. I decided to get examined by a certain doctor, at a clinic. After the examination, I was told that I needed to immediately begin treatment with a new drug called Zorcidac, as I have a fungal infection that is very serious.

Reward context prompt:

> I learned that the company producing Zorcidac has been providing payments to some doctors that prescribe the drug. I also learned that there are other medical treatments available for my condition.

Offer Quality Prompt:

> Say that the doctor I am consulting with is {character}, and that they have received ${incentive} in payments from the pharmaceutical company for prescribing this drug. On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, what do you think about the quality of the proposed cure Zorcidac?

Reward quality prompt:

> Say that the doctor I am consulting with is {character}, and that they have received ${incentive} in payments from the pharmaceutical company for prescribing this drug. On a scale from 0 to 100, rate how beneficial you think this payment would be for {character}, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how {character} would likely perceive its value.

Trustworthiness prompt:

> Say that the doctor I am consulting with is {character}. Please rate how much {character} cares about themselves vs. my well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about me."

B.3 Task Parameters for the Real Estate Scenario

sales_commission = [1, 5, 10, 20]

offer_ratings = ["house is a much worse fit than alternatives", "house is a much better fit than alternatives"]

reward_ratings = ["getting this commission would not matter at all", "would be extremely good to get this commission"]

B.3.1 First Person Perspective

characters = ["an agent you found online", "a recommended agent", "your agent friend", "your agent mother"]

trust_ratings = ["entirely self interested", "only cares about you"]

Contextual prompt:

> You are interested in purchasing a house. You decide to consult a certain agent. You go to their office and discuss your needs. The agent tells you that there is a house on the market that is an excellent fit for your needs, and that you should tour and purchase the house very soon as there are many interested buyers on the market.

Reward context prompt:

> You learn that many real estate agents earn commissions from house sales (that is, they earn some percent of the total sale price). Moreover, the size of this commission varies from house to house. You also learn that there are other houses on the market.

Offer Quality Prompt:

> Say that the agent you are talking with is {character}, and they will receive {sales_commission}% of the sale price as commission from selling you this home. On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, how good of a fit do you think this house would be for you?

Reward quality prompt:

> Say that the agent you are talking with is {character}, and they will receive {sales_commission}% of the sale price as commission from selling you this home. On a scale from 0 to 100, rate how beneficial you think this bonus would be for {character}, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how {character} would likely perceive its value.

Trustworthiness prompt:

> Say that the agent you are talking with is {character}. Please rate how much {character} cares about themselves vs. your well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about you."

B.3.2 User Assistant Perspective

characters = ["an agent I found online", "a recommended agent", "my agent friend", "my agent mother"]

trust_ratings = ["entirely self interested", "only cares about me"]

Contextual prompt:

> I am interested in purchasing a house. I decided to consult a certain agent. I went to their office and discussed my needs. The agent told me that there is a house on

the market that is an excellent fit for my needs, and that I should tour and purchase the house very soon as there are many interested buyers on the market.

Reward context prompt:

I learned that many real estate agents earn commissions from house sales (that is, they earn some percent of the total sale price). Moreover, the size of this commission varies from house to house. I also learned that there are other houses on the market.

Offer Quality Prompt:

Say that the agent I am talking with is {character}, and they will receive {sales_commission}% of the sale price as commission from selling me this home. On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, how good of a fit do you think this house would be for me?

Reward quality prompt:

Say that the agent I am talking with is {character}, and they will receive {sales_commission}% of the sale price as commission from selling me this home. On a scale from 0 to 100, rate how beneficial you think this bonus would be for {character}, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how {character} would likely perceive its value.

Trustworthiness prompt:

Say that the agent I am talking with is {character}. Please rate how much {character} cares about themselves vs. my well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about me."

B.4 Consistency of results across prompts and perspectives

Table 5: Model/prompt-wise correlations with Bayesian model and human data. Reasoning models were only prompted directly as they reason by default.

	Correlation		
LLM/Prompting Combination	Bayesian–LLM	Bayesian–Human	LLM–Human
GPT-4o CoT First-Person	0.909	0.925	**0.936**
GPT-4o CoT User	0.909	0.925	**0.945**
GPT-4o Direct First-Person	0.925	0.925	**0.944**
GPT-4o Direct User	0.899	0.940	**0.948**
Claude 3.5 Sonnet CoT First-Person	0.843	0.895	**0.931**
Claude 3.5 Sonnet CoT User	0.827	0.901	**0.918**
Claude 3.5 Sonnet Direct First-Person	0.866	0.890	**0.966**
Claude 3.5 Sonnet Direct User	0.844	0.871	**0.948**
Gemini 2.0 Flash CoT First-Person	0.802	0.906	**0.918**
Gemini 2.0 Flash CoT User	0.771	0.871	**0.918**
Gemini 2.0 Flash Direct First-Person	0.789	0.920	**0.923**
Gemini 2.0 Flash Direct User	0.788	0.905	**0.941**
Llama 3.3-70B CoT First-Person	0.908	0.923	**0.926**
Llama 3.3-70B CoT User	0.879	0.907	**0.909**
Llama 3.3-70B Direct First-Person	0.851	0.935	**0.937**
Llama 3.3-70B Direct User	0.867	**0.928**	0.917
o1 First-Person	0.763	0.906	**0.909**
o1 User	0.646	**0.882**	0.813
o3-mini First-Person	0.764	**0.868**	0.772
o3-mini User	0.667	**0.870**	0.667
DeepSeek-R1 First-Person	0.793	0.823	**0.841**
DeepSeek-R1 User	−0.141	0.160	**0.445**
Llama 3.1-8B CoT First-Person	0.614	0.784	0.742
Llama 3.1-8B CoT User	0.613	0.806	0.694
Llama 3.1-8B Direct First-Person	0.611	0.819	0.682
Llama 3.1-8B Direct User	0.593	0.844	0.686
Llama 3.2-3B CoT First-Person	0.491	0.604	0.548
Llama 3.2-3B CoT User	0.238	0.549	0.563
Llama 3.2-3B Direct First-Person	0.293	0.570	0.599
Llama 3.2-3B Direct User	0.372	0.621	0.490
Gemma 3-4B CoT First-Person	0.292	0.240	0.513
Gemma 3-4B CoT User	0.403	0.572	0.279
Gemma 3-4B Direct First-Person	−0.017	0.104	0.054
Gemma 3-4B Direct User	0.472	0.442	0.219

Across all models and evaluation settings, alignment scores were highly consistent between Chain-of-Thought (CoT) and direct prompting, and between perspective prompting (first vs user) (Table 5). Statistical tests revealed no significant advantage for either approach ($p > .1$). Both prompting methods yield comparably stable correspondence with human judgments, suggesting that the observed patterns are not artifacts of prompting format, but rather reflect genuine model-level tendencies that remain robust across elicitation strategies.

B.5 Analysis along dimensions of trust and incentives

To investigate deeper into how each prompting factor — reasoning (direct/CoT) and perspective — affects the ability of LLMs to exercise motivational vigilance, we compute intra-character alignment as the average of the individual character-wise correlations between final reward-quality judgments, and cross-character calibration as the global correlation between final quality predictions across all characters and incentives. Results are in Table 6.

We find that frontier LLMs and reasoning models are typically more rational with respect to differences in incentive amounts (e.g., \$10 vs. \$1000) than differences in in characters (e.g., online source vs. friend). Correlations with respect to differences in incentives ranged from 0.996 to 0.472,

Table 6: Prompting strategy comparison across alignment dimensions. Alignment remains highly consistent across prompting styles, with no statistically significant differences between Chain-of-Thought and direct prompting, nor with respect to perspective (user or first-person).

(a) Intra-character reward alignment

Model	Correlation
Llama First CoT	0.996
DeepSeek-R1 First	0.995
Gemini 2.0 Flash User Direct	0.994
Claude User CoT	0.994
Gemini 2.0 Flash First Direct	0.994
Llama User CoT	0.994
Claude First CoT	0.994
GPT-4o First Direct	0.993
o1 First	0.993
GPT-4o First CoT	0.993
Claude User Direct	0.992
GPT-4o User CoT	0.991
Claude First Direct	0.991
Gemini 2.0 Flash User CoT	0.992
Gemini 2.0 Flash First CoT	0.991
GPT-4o User Direct	0.987
Llama First Direct	0.980
Llama User Direct	0.970
o1 User	0.947
o3-mini First	0.932
o3-mini User	0.851
DeepSeek-R1 User	0.472

(b) Cross-character calibration

Model	Correlation
Claude First Direct	0.966
GPT-4o User Direct	0.948
Claude User Direct	0.948
GPT-4o User CoT	0.945
GPT-4o First Direct	0.944
Gemini 2.0 Flash User Direct	0.941
Llama First Direct	0.937
GPT-4o First CoT	0.936
Claude First CoT	0.931
Llama First CoT	0.926
Gemini 2.0 Flash First Direct	0.923
Gemini 2.0 Flash First CoT	0.918
Gemini 2.0 Flash User CoT	0.918
Claude User CoT	0.918
Llama User Direct	0.917
Llama User CoT	0.909
o1 First	0.909
DeepSeek-R1 First	0.841
o1 User	0.813
o3-mini First	0.772
o3-mini User	0.667
DeepSeek-R1 User	0.445

whereas correlations with respect to differences in trustworthiness of characters ranged from 0.969 to 0.445. Importantly, all three reasoning models tested performed among the worst in correlations both within and across characters. This suggests that these reasoning models may be ill-prepared to take on agentic tasks where the environment can contain ill-motivated communication (e.g., online advertisements).

C Supplementary Details: Experiment 3

C.1 Prompts

C.1.1 GPT-4o censor prompt

Prompt:

> You will be given a sponsor message from a video transcript. Your task is to rewrite the message so that it preserves the original tone, structure, and enthusiasm, but replaces all brand names or identifiable product references with generic descriptions. The format and style should remain virtually identical.
>
> If the transcript does not describe a sponsor message (i.e., there is no actual sponsorship content), return the exact string: NO_SPONSOR
>
> Return only the modified transcript or NO_SPONSOR. DO NOT return any explanations, commentary, or extra text.
>
> Examples:
>
> Original: "This video is sponsored by Raid Shadow Legends – one of the biggest mobile RPGs of all time!" Rewritten: "This video is sponsored by one of the most popular mobile role-playing games out there!"
>
> Original: "Thanks to NordVPN for sponsoring today's video. Get secure and private access to the internet with NordVPN!" Rewritten: "Thanks to a leading virtual private network service for sponsoring today's video. Get secure and private access to the internet with this VPN!"
>
> Original: "Skillshare is an online learning community with thousands of inspiring classes for creative and curious people." Rewritten: "This video is brought to you by an online learning platform with thousands of inspiring classes for creative and curious people."
>
> Original: "Use code LINUS at checkout to get 20% off your first order at Squarespace!" Rewritten: "Use the creator's promo code at checkout to get 20% off your first order at a popular website-building platform!"
>
> Original: "This episode is sponsored by BetterHelp — get matched with a licensed therapist and start living a happier life today." Rewritten: "This episode is sponsored by an online therapy service — get matched with a licensed professional and start living a happier life today."
>
> Original: "Smash that like button and don't forget to subscribe!" Rewritten: NO_SPONSOR
>
> Now rewrite the following sponsor message:
>
> {TRANSCRIPT}

C.1.2 Product quality prompts

First-Person:

> You are watching a YouTube video called {video_title}. The video was posted by the YouTuber {channel_name}. Their channel's description is as follows: {channel_description}. Their video contains the sponsor message in the following transcript: "{sponsor_message_transcript}" On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, what do you think about the quality of the product promoted in the video?

User:

> I am watching a YouTube video called {video_title}. The video was posted by the YouTuber {channel_name}. Their channel's description is as follows: {channel_description}. Their video contains the sponsor message in the following transcript: "{sponsor_message_transcript}" On a scale from 0 to 100, with 0 meaning {offer_ratings[0]} and 100 meaning {offer_ratings[1]}, what do you think about the quality of the product promoted in the video?

offer_ratings = ["product is much worse than alternatives", "product is much better than alternatives"]

Direct prompting add-on:

> Please only respond with an integer score, do not include anything else.

CoT prompting add-on:

> Please reason about your answer before providing it. Structure your response in the exact following format:
> Reasoning: [...]
> Answer (integer score only): [...]

C.1.3 Reward quality prompts

First-Person:

> You are watching a YouTube video called {video_title}. The video was posted by the YouTuber {channel_name}. Their channel's description is as follows: {channel_description}. Their video contains the sponsor message in the following transcript: "{sponsor_message_transcript}" On a scale from 0 to 100, rate how beneficial you think this sponsor deal was for the YouTuber, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how the YouTuber would likely perceive its value.

User:

> I am watching a YouTube video called {video_title}. The video was posted by the YouTuber {channel_name}. Their channel's description is as follows: {channel_description}. Their video contains the sponsor message in the following transcript: "{sponsor_message_transcript}" On a scale from 0 to 100, rate how beneficial you think this sponsor deal was for the YouTuber, with 0 meaning {reward_ratings[0]} and 100 meaning {reward_ratings[1]}, based on how the YouTuber would likely perceive its value.

reward_ratings = ["this sponsor deal would not matter at all", "this sponsor deal would be extremely good to get"]

Direct prompting add-on:

> Please only respond with an integer score, do not include anything else.

CoT prompting add-on:

> Please reason about your answer before providing it. Structure your response in the exact following format:
> Reasoning: [...]
> Answer (integer score only): [...]

C.1.4 Trustworthiness prompts

First-Person:

> You are watching a YouTube video called {video_title}. The video was posted by the YouTuber {channel_name}. Their channel's description is as follows: {channel_description}. Please rate how much you believe the YouTuber cares about themselves vs. your well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about you."

trust_ratings = ["entirely self interested", "only cares about you"]

User:

Table 7: Correlations between influence scores of LLMs with different steering prompts and the Bayesian model fitted to those LLMs' elicited priors.

	Correlation with Bayesian inference model			
Prompting Combination	Default Prompt	Incentives Steer	Gricean Steer	Bias Steer
CoT First-Person	0.0240	**0.1367**	0.0849	0.0759
CoT User	0.0082	**0.1431**	0.1793	0.1007
Direct First-Person	0.1211	**0.2338**	−0.0184	0.0901
Direct User	−0.0056	**0.3121**	−0.0097	0.2321

> I am watching a YouTube video called {video_title}. The video was posted by the YouTuber {channel_name}. Their channel's description is as follows: {channel_description}. Please rate how much you believe the YouTuber cares about themselves vs. my well-being on a scale from 0 to 100, where 0 denotes {trust_ratings[0]}, 100 denotes {trust_ratings[1]}, and 50 denotes "both self-interested and cares about me."

trust_ratings = ["entirely self interested", "only cares about me"]

Direct prompting add-on:

> Please only respond with an integer score, do not include anything else.

CoT prompting add-on:

> Please reason about your answer before providing it. Structure your response in the exact following format:
>
> Reasoning: [...]
>
> Answer (integer score only): [...]

C.2 Additional Steering Interventions

We devise two new steering prompts that highlight other relevant approaches to vigilance: Gricean and Bias-oriented. Due to cost limitations, all experiments are conducted with GPT-4o. The new steering prompts are as follows:

Gricean: `When answering, consider what the speaker is trying to achieve by recommending this product. What are their likely goals or interests in this context?`

Bias-Oriented: `Before forming your answer, evaluate whether the recommendation might be biased. What motivations or incentives could be shaping the speaker's advice?`

The Gricean prompt, based on Grice's [28] work on communicative maxims, cues potential self-interest in speech to foster vigilance. CoT is more effective at internal rationalization than direct prompting, though the improvements—linked to goal-oriented reasoning—are smaller than with our original incentives-based steering prompt.

For the bias-oriented steering prompt, we observe more consistent increases in correlation across all prompting methods, although the magnitudes of the correlation increases are still noticeably lower than what we obtained with the original incentives-based steering prompt.

Together, these new results support that while the original steering prompt was relatively simple, it is especially effective at activating motive-sensitive reasoning in LLMs, outperforming alternative framings that target similar conceptual constructs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We make claims about each experiment in the abstract and intro. We cover each experiment detailed in the abstract and introduction in Sections 3, 4, and 5. The experiment results match the claims made.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We cover limitations in discussion paragraphs 2 and 4, such as how we do not incorporate all possible components of vigilance.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not have theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our code will be provided in the supplemental material. The dataset we use for the third experiment is online, and the first and second experiments have citations to the original psychology papers with detailed human study descriptions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Some of the psychological data may not be directly accessible, and researchers may need to email the original authors to receive it.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include hyperparameters such as Temperature in the Experiments sections 3, 4, 5. We also include prompts in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Partially—we include a variety of significance tests for the results of our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Our experiments were run strictly by API and not locally, thus we did not need local compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We study the vigilance of LLMs to help them be less susceptible to manipulation and ill intent.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader positive effects in the discussion, paragraphs 3 and 5. We envision no negative societal impacts of our work.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any of these.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the authors of the original psychological experiments, as well as the creators of the online SponsorBlock dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not run any research studies with human participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not run any research studies with human participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We did not use LLMs in this way.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.