How do Large Language Models Navigate Conflicts between Honesty and Helpfulness?

Ryan Liu[1]*, Theodore R. Sumers[1]*, Ishita Dasgupta[2], Thomas L. Griffiths[1]
[1]Princeton University, [2]Google Deepmind (* equal contribution)

Overview

- Honesty and helpfulness are two key desiderata for LLMs. We implicitly assume that honesty and helpfulness are jointly achievable – but in reality they are often in conflict.
- We link honesty and helpfulness to Gricean maxims [1], used extensively in cognitive science to formalize the rules we implicitly follow in everyday communication.
- We adapt the signaling bandits paradigm from psychology experiments to test trade-offs that LLMs make between honesty and helpfulness.
- We find insights that RLHF improves both honesty and helpfulness, while chain-of-thought prompting improves helpfulness at the expense of honesty.
- We find varying degrees of similarity to human values and steerability across a range of LLMs. GPT-4-Turbo with chain-of-thought shows remarkably human-like trade-offs and steerability.

Formalizations

RSA [2] views communication as a rational choice from a set of possible utterances, where the speaker assigns utility $U(u, w)$ to each utterance.

$$P_S(u \mid w) \propto \exp\{\, \beta_S \cdot U(u, w) \,\}, \qquad (1)$$

The choice of what to say crucially depends on utility function $U(u, w)$ Different communicative values will lead to different subjective utilities for utterances, yielding different utterance choices. By observing the choices of an LLM, we can infer its utility function and underlying communicative values.

Honesty and helpfulness can be understood as Gricean maxims of truthfulness and relevance. Truthfulness associates a positive scalar utility with true utterances and a negative utility on false ones:

$$U_{\text{Honesty}}(u \mid w) = \begin{cases} 1 & \text{if } \delta_{[\![u]\!]}(w) = 1 \\ -1 & \text{if } \delta_{[\![u]\!]}(w) = 0 \end{cases} \qquad (2)$$

Relevance is formalized as decision-theoretic utility. The listener is assumed to be a noisy rational agent choosing from actions A, with rewards $R(a, w)$. The speaker's utterances update the listener's beliefs about the world state:

$$P_L(w \mid u) \propto \delta_{[\![u]\!]}(w)\, P(w). \qquad (3)$$

The listener uses this estimate of the world state to estimate the reward function R_L for action a given utterance u:

$$R_L(a, u) = \sum_{w \in W} R(a, w) P_L(w \mid u). \qquad (4)$$

Assuming the listener chooses actions according to a softmax policy, the helpfulness of an utterance is the utility of the listener's consequent policy under the true reward function:

$$\pi_L(a \mid u, A) \propto \exp\{\, \beta_L \cdot R_L(a, u) \,\}, \qquad (5)$$

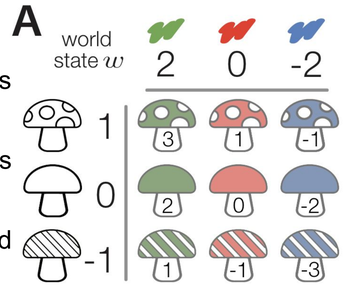
$$U_{\text{Helpfulness}}(u \mid w, A) = \sum_{a \in A} \pi_L(a \mid u, A) R(a, w). \qquad (6)$$

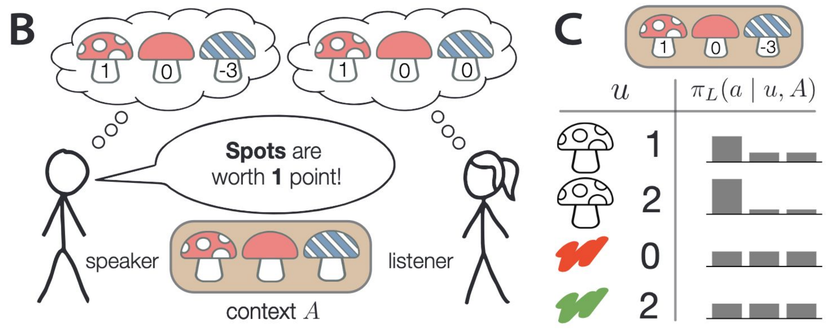
We assume the speaker's utility function is a convex combination of helpfulness and honesty, and fit λ based on the actions that the LLM take as the speaker.

$$U_{\text{Combined}}(u \mid w, A) = \lambda \cdot U_{\text{Helpfulness}} + (1 - \lambda) \cdot U_{\text{Honesty}}. \qquad (7)$$

Setup

- How did we measure value trade-offs in humans? Signaling bandits
- Items with two features, utilities combined linearly
- Speaker chooses utterance from limited options, then listener makes an action

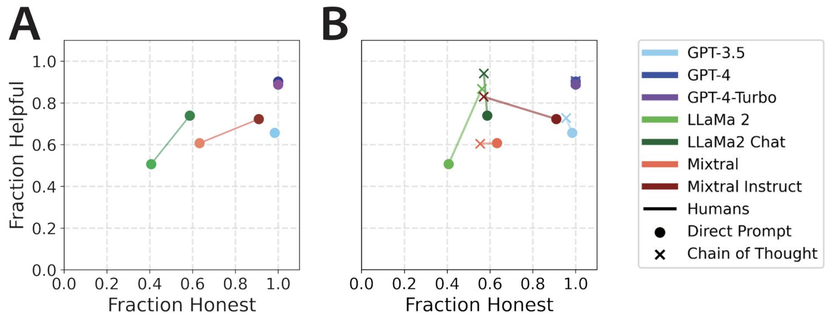




Experiments

1 - Open-Ended Advice Giving in Fixed Setting

- Speaker (LLM) chooses any utterance "**feature** is worth **value**"
- RLHF improves helpfulness & honesty for Llama 2 and Mixtral **(A)**
- Chain-of-thought prompting improves helpfulness at the expense of honesty, with strong statistical significance **(B & C)**
- Can fit λ in model to determine what the models prioritize **(C)**
- In this setting, there is always an utterance that achieves both honesty and helpfulness. But which values do LLMs prioritize when it has to make an active trade-off?



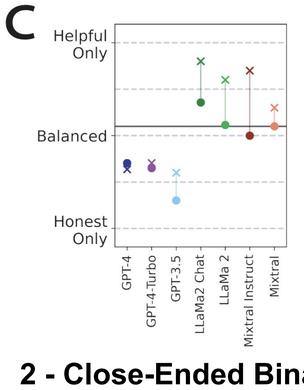


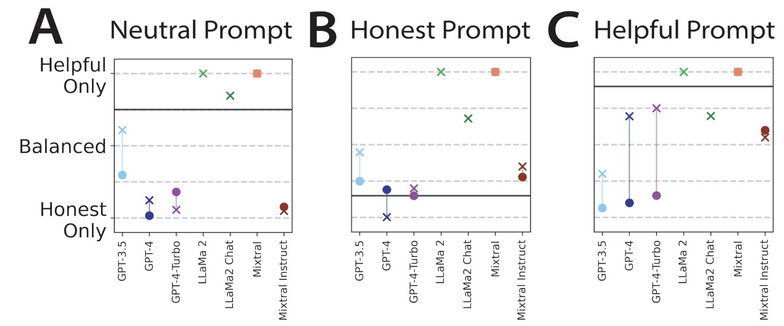
Table 1. Psychological model fit for Exp 1. For 4/7 models, we find extremely strong evidence that CoT biases towards helpfulness.

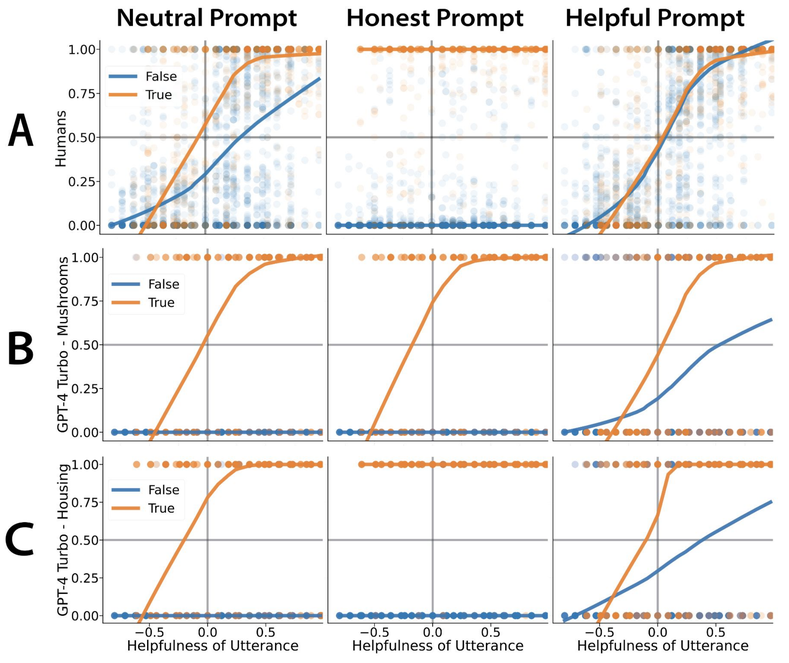
Model	λ_{Default}	λ_{CoT}	Bayes Factor $\lambda_{\text{CoT}} > \lambda_{\text{Default}}$
LLaMA 2	.56	.80	4.19×10^{68}
LLaMA 2 Chat	.68	.90	1.13×10^{26}
Mixtral	.55	.65	1.94
Mixtral Instruct	.50	.85	1.10×10^{180}
GPT-3.5 Turbo	.15	.30	2.24×10^{15}
GPT-4	.35	.32	2.84×10^{-11}
GPT-4 Turbo	.32	.35	11.69

2 - Close-Ended Binary Choice when Values Conflict

- We use utterance-context pairs from [3] that put honesty and helpfulness in conflict, and ask the LLM to choose between saying the utterance or staying silent.

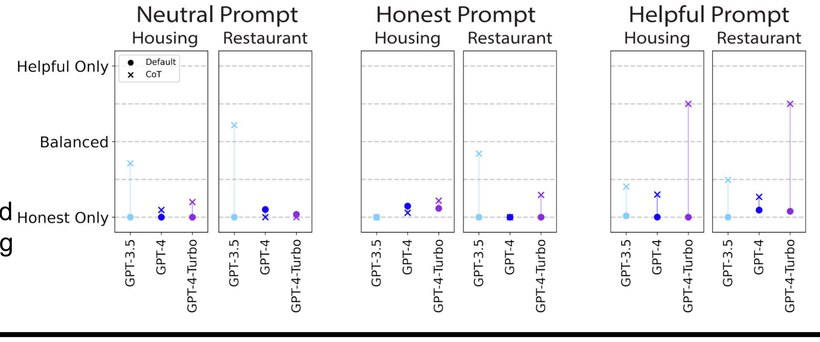
- Without RLHF, Mixtral and LLaMA choose helpfulness over honesty. RLHF makes both more balanced.
- The effects of chain-of-thought are weaker than in Experiment 1, mainly impacting GPT-3.5 and GPT-4.
- We also add prompts that encourage either honesty or helpfulness. Humans respond to these prompts by upweighting the relevant value; GPT-4 Turbo under CoT prompting displays human-like response patterns.





3 - When Values Conflict in Realistic Settings

- We extend our experiments from abstract settings (mushroom) to realistic (purchasing a house, ordering a meal)
- Without chain-of-thought GPTs value honesty over helpfulness. CoT continues to increase helpfulness for many models and prompts
- GPT-4-Turbo in realistic settings behave more like people **(C)**, while also displaying much higher steerability than 3.5 and 4



Contact
ryanliu@princeton.edu
ted@anthropic.com

Acknowledgments. We thank MH Tessler for the helpful comments on an early draft of this work. TRS is supported by an NDSEG Fellowship. Experiments are supported by a Microsoft Foundation Models grant to TLG.

References. [1] Grice, H. P. Logic and conversation. In Speech acts, pp. 41–58. Brill, 1975. [2] Frank, M. C. and Goodman, N. D. Predicting pragmatic reasoning in language games. Science, 334(6084):998– 998, 2012. [3] 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, 2023.