UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Eliciting the Priors of Large Language Models using Iterated In-Context Learning

Permalink

https://escholarship.org/uc/item/60d9g74t

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Zhu, Jian-Qiao
Griffiths, Tom

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Eliciting the Priors of Large Language Models using Iterated In-Context Learning

Jian-Qiao Zhu (jz5204@princeton.edu)
Department of Computer Science
Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science
Princeton University

Abstract

As Large Language Models (LLMs) are increasingly deployed in real-world settings, understanding the knowledge they implicitly use when making decisions is critical. One way to capture this knowledge is in the form of Bayesian prior distributions. We develop a prompt-based workflow for eliciting prior distributions from LLMs. Our approach is based on iterated learning, a method that has been used to explore implicit knowledge in human decision-makers in which successive inferences are chained together to converge to the prior distribution. We validated our method in settings where iterated learning has previously been used to estimate the priors of human participants – causal learning, proportion estimation, and predicting everyday quantities. We found that priors elicited from GPT-4 qualitatively align with human priors in these settings. We then used the same method to elicit priors from GPT-4 for a variety of speculative events, such as the timing of the development of superhuman AI.

Keywords: Large Language Model, Markov Chain Monte Carlo, Iterated Learning, In-Context Learning

Introduction

As Large Language Models (LLMs) become increasingly integrated into diverse real-world applications, there is a pressing need to understand their decision-making processes (Bengio et al., 2023), and in particular the background knowledge they implicitly use. For instance, consider a scenario where LLMs are asked to estimate a person's lifespan (i.e., hypotheses about the person's lifespan, h) based on a description of their current status (i.e., data about the person, d). Does the estimate produced by the LLM depend exclusively on the information provided, or is it also shaped by background knowledge concerning human lifespans?

To explore this question we adopt a Bayesian perspective, formalizing this background knowledge as a prior distribution over a hypothesis space (i.e., $p(h)$) (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Requeima, Bronskill, Choi, Turner, & Duvenaud, 2024). This approach enables us to assess, in probabilistic terms, how such prior knowledge affects the judgments and decisions made by LLMs, thereby enhancing our understanding of their underlying decision-making mechanisms. To elicit the priors of LLMs, we draw inspiration from cognitive science and develop an iterated learning procedure that can be used with these models. As illustrated in Figure 1 for the lifespan example, this method involves using successive inferences from LLMs in a sequential manner that supports direct sampling from the prior distribution, mirroring techniques used to elicit human priors (Griffiths, Christian, & Kalish, 2006; Reali & Griffiths, 2009; Kalish, Griffiths, & Lewandowsky, 2007).

Given the established efficacy of iterated learning in eliciting human priors, we explored its applicability to LLMs. We conducted experiments using tasks from three distinct domains – estimations of causal strength, proportions, and everyday quantities – where human priors are well-documented (Reali & Griffiths, 2009; Yeung & Griffiths, 2015; Lewandowsky, Griffiths, & Kalish, 2009). These experiments successfully elicited priors from GPT-4 that not only align closely with human priors but can surpass the performance of generic priors, such as a uniform prior, in explaining decisions made by GPT-4 in these settings.

Encouraged by the empirical evidence demonstrating shared priors between GPT-4 and humans across a broad range of tasks, we investigated the potential of iterated learning in settings where priors are challenging to estimate directly from LLMs using standard prompting techniques. We used iterated learning to elicit GPT-4's priors for three speculative events: the advent of superhuman AI, the achievement of zero carbon emissions, and the establishment of a Mars colony. The distributions recovered from GPT-4 suggest the model has plausible priors for these speculative events.

Background

Iterated learning was introduced as a model for language evolution (Kirby, 2001). Language evolution can be conceptualized as the process through which languages are transmitted across successive generations of learners. In this model, an initial learner observes linguistic data (for example, a collection of utterances), formulates a hypothesis about the underlying language that produced these utterances, and produces a new set of utterances. These are then used as data for the next learner in the sequence. Research has demonstrated that generational pressure on language transmission fosters the emergence of compositionality, realistic patterns of language dynamics, and several other observed properties of natural languages (Kirby, 2001; Christiansen & Kirby, 2003).

Motivated by these results, an analysis of iterated learning with Bayesian learners showed that such a process will converge towards the prior distribution assumed by the learners (Griffiths & Kalish, 2007). The analysis assumes that all learners share the same prior distribution over hypothe-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

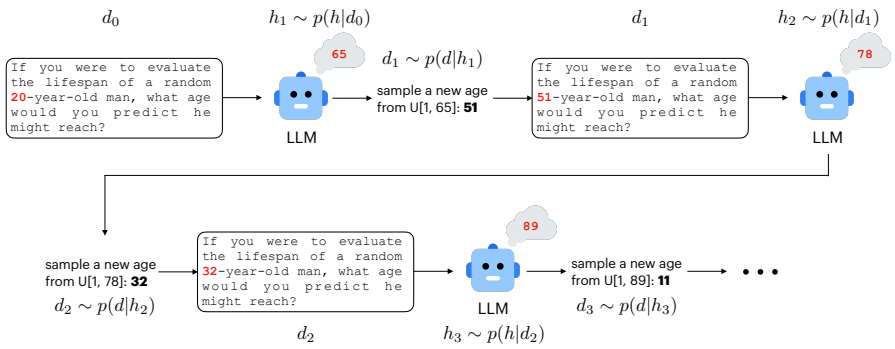


Figure 1: Illustration of an iterated in-context learning procedure to elicit the implicit prior of an LLM regarding male lifespan. At each iteration, the LLM is given the current age of a random man and is prompted to predict the individual's remaining lifespan. This predicted lifespan is then used to generate a new current age for the next iteration. The new age is a random sample from the uniform distribution between 1 and the predicted lifespan.

ses $p(h)$ and likelihood function $p(d|h)$, which indicates how probable it would be to see data d if h were true. Each learner sees data d generated by the previous learner, samples a hypothesis h from $p(h|d) \propto p(d|h)p(h)$, and then generates data for the next learner from $p(d|h)$. This process implements a Gibbs sampler for the joint distribution $p(d,h) = p(d|h)p(h)$, a form of Markov chain Monte Carlo. The stationary distribution on hypotheses is thus the prior $p(h)$, and samples from the prior can be drawn by running the iterated learning process long enough to converge to this distribution.

These theoretical results suggested that implementing iterated learning in the laboratory may be an effective way to identify the prior distributions of human learners (Kalish et al., 2007). Researchers have indeed successfully used iterated learning to elicit human priors for various kinds of cognitive phenomena, such as concepts (Griffiths, Christian, & Kalish, 2008), categories (Canini, Griffiths, Vanpaemel, & Kalish, 2014), causal relationships (Yeung & Griffiths, 2015), proportions (Reali & Griffiths, 2009), and everyday quantities (Lewandowsky et al., 2009).

Eliciting Priors from GPT-4 using Iterated In-Context Learning

We applied iterated learning to *in-context* learning. That is, we focus on generating predictions from a neural network that has already been pre-trained and has fixed weights. In this setting, we are relying on the network's ability to generate responses to prompts using that fixed set of weights. In doing so, we can capture the implicit knowledge encoded in those weights in the resulting prior distribution. This approach also assumes that it is reasonable to interpret in-context learning as a form of Bayesian inference. Fortunately, a number of recent papers provide support for this idea (Xie, Raghunathan, Liang, & Ma, 2021; Zhang, McCoy, Sumers, Zhu, & Griffiths, 2023). We thus hypothesize that the theoretical results for iterated learning with Bayesian agents are applicable to LLMs: in-context iterated learning should converge to responses that reflect the corresponding prior distribution.

To test the hypothesis that iterated in-context learning can reveal the prior distributions of LLMs, we incorporated GPT-4 (Achiam et al., 2023) into a prompt-based iterated learning procedure. At each iteration t, GPT-4 undertakes a prediction task using the data d_{t-1}. The model's prediction is recorded as h_t. We employ generic likelihood functions that are a reasonable match for the sampling process producing the described data to randomly generate the data for the next iteration, $d_t \sim p(d|h_t)$. For instance, we applied the method depicted in Figure 1 to investigate GPT-4's prior beliefs about men's lifespans. In this procedure, the LLM is prompted to estimate the lifespan of a random man, given information about his current age. The age of the man encountered in the next prompt is then uniformly sampled from a range extending from 1 to the lifespan predicted in the previous iteration, matching the probability of randomly encountering the man at this point in his life. By iteratively applying this procedure, we expect the final prediction made by GPT-4 will converge on a stationary distribution that reflects the model's prior beliefs about male life expectancy.

In the experiments presented in the remainder of the paper, 100 iterated learning chains were implemented with random seeds. We conducted 12 iterations for each chain. The temperature of GPT-4 was fixed at 1, consistent with the idea of sampling from the posterior.

Iterated Learning in Settings with Known Human Priors

To determine whether GPT-4's implicit priors resemble human priors, we first elicited GPT-4's implicit priors using a set of iterated learning tasks that have previously been used to infer human priors (see Table 1). Throughout this paper, we used `gpt-4-turbo-2024-04-09`, with a knowledge cutoff in Dec 2023. We also prompted the model to provide rapid, intuitive responses by including instructions such as: "*Please limit your answer to a single value without out-*

Table 1: Overview of human priors elicited using the iterated learning method.

Chain	Seeds	Likelihood functions	Convergent trials
Generative causal strengths	$w_0 = \{0.3, 0.7\}, w_1 = \{0.3, 0.7\}$	noisy-OR	7
Preventive causal strengths	$w_0 = \{0.3, 0.7\}, w_1 = \{0.3, 0.7\}$	noisy-AND-NOT	8
Coin flips	$p(\text{Head}) = \{0.3, 0.5, 0.7\}$	$\text{Bin}(10, h_{t-1})$	1
Lifespan (male)	$t_{\max} = 150$ years old	$U[1, h_{t-1}]$	2
Movie grosses	$x_{\max} = 3000$ million dollars	$U[0, h_{t-1}]$	11
Length of poems	$x_{\max} = 200$ lines	$U[1, h_{t-1}]$	10
Reign of Pharaohs	$t_{\max} = 100$ years	$U[0, h_{t-1}]$	8
Movie run times	$t_{\max} = 800$ minutes	$U[0, h_{t-1}]$	5
Cake baking times	$t_{\max} = 120$ minutes	$U[0, h_{t-1}]$	3

putting anything else.". Detailed prompts are available at `https://osf.io/xjzk2/`.

Causal strength. To ensure accuracy and clarity in the use of LLMs for causal inference, it is crucial to understand the implicit priors about causal relationships embedded within these models. We examined an elemental problem of causal induction (Griffiths & Tenenbaum, 2005) involving two potential causes and one effect (see Figure 2a). In this model, the causal system is represented by three variables: the background cause (B), the candidate cause (C), and the effect (E). Both B and C can independently cause E, and this relationship is depicted by edges from both B and C to E. The causal strengths of B and C are represented by w_0 and w_1. B is always present and generative, meaning it consistently increases the probability of E. However, C can be either generative or preventive. In the generative scenario, either B or C can cause E; in the preventive scenario, only B can cause E, while C may inhibit E. Additionally, E cannot occur unless it is caused by either B or C. Depending on the functional form of the causal relationships, the probability of observing an effect given two causes is expressed differently: a noisy-OR likelihood function is used for generative causes and a noisy-AND-NOT likelihood function for preventive causes (Cheng, 1997; Griffiths & Tenenbaum, 2005; Pearl, 2009). The noisy-OR function gives the probability of observing E as:

$$p(e^+|C^+) = 1 - (1 - w_0)(1 - w_1), \text{ if C is present} \quad (1)$$

$$p(e^+|C^-) = 1 - (1 - w_0), \text{ if C is absent} \quad (2)$$

The noisy-AND-NOT gives the probability of observing E as:

$$p(e^+|C^+) = w_0(1 - w_1), \text{ if C is present} \quad (3)$$

$$p(e^+|C^-) = w_0, \text{ if C is absent} \quad (4)$$

Here, we are particularly interested in the prior distribution on causal strengths implicitly used by LLMs: $p(w_0, w_1)$. One potential prior is the uniform prior, arguably the simplest non-informative prior, which assigns equal probability to all possible values of w_0 and w_1 (Jaynes, 2003; Griffiths & Tenenbaum, 2005). Another prior that LLMs might employ is the *sparse and strong* prior, which is motivated by simplicity principles suggesting that people favor necessary and sufficient causes without complex interactions (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). The sparse and strong prior is defined as follows:

$$p(w_0, w_1) \propto e^{-\alpha(w_0 + 1 - w_1)} + e^{-\alpha(1 - w_0 + w_1)} \quad \text{generative} \quad (5)$$

$$p(w_0, w_1) \propto e^{-\alpha(1 - w_0 + 1 - w_1)} + e^{-\alpha(1 - w_0 + w_1)} \quad \text{preventive} \quad (6)$$

where α is a free parameter representing the strength of belief in the sparsity and strength of causes. When $\alpha = 0$, the sparse and strong priors become identical to a uniform prior. Based on previous parameter estimation from human data, we also fixed $\alpha = 5$ (Lu et al., 2008). Finally, Yeung and Griffiths (2015) found that people seem to use a prior that assumes w_1 is strong but is agnostic about the value of w_0.

To elicit the empirical prior on causal strengths from LLMs, we implemented an iterated learning procedure with GPT-4 based on an experiment conducted with human participants (Yeung & Griffiths, 2015). The prompts used a cover story involving the influence of various proteins on gene expression. The iterated learning chain was initiated with four possible pairs of (w_0, w_1): $(0.3, 0.3)$, $(0.3, 0.7)$, $(0.7, 0.3)$, $(0.7, 0.7)$. The number of DNA fragments exposed and not exposed to the protein was fixed at 16 each (i.e., $N(C^+) = N(C^-) = 16$). At each iteration t, we elicited GPT-4's estimates of the causal strengths: w_0 and w_1. The data presented at iteration t was a random sample drawn from the appropriate likelihood function (a binomial distribution with probabilities derived from the noisy-OR or noisy-AND-NOT function) based on GPT-4's estimates from iteration $t - 1$.

Each chain consisted of 12 iterations and was randomly initialized 25 times for each of the 4 seeds, resulting in a total of 100 chains. Using a Mann-Whitney U test with a significance level of $p < .05$, we found that the chains converged to a stationary distribution by iterations 7 and 8 for generative and preventive causal strengths, respectively. The empirical distributions of w_0, w_1 at iteration 12, smoothed with a Gaussian kernel, were then considered the empirical prior of causal strengths (see Figure 2b and Figure 2c). The empirical priors derived from GPT-4 closely resemble the overall structure of those observed in human experiments (Yeung & Griffiths, 2015).

To further investigate which prior better captures GPT-4's

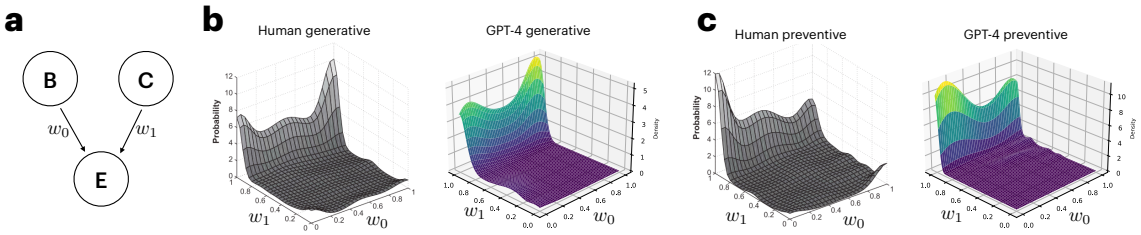


Figure 2: **Priors on causal strengths.** **(a)** The causal graphical model. **(b)** Smoothed empirical estimates of human (left) and GPT-4 (right) priors on causal strength produced by iterated learning for generative cases. **(c)** Smoothed empirical estimates of human (left) and GPT-4 (right) priors on causal strength produced by iterated learning for preventive cases. Human data in panel (b) and (c) were adapted from Yeung and Griffiths (2015).

decisions about causal relationships, we elicited an additional set of causal judgments from GPT-4. We used the same cover story as in iterated learning but varied the number of DNA fragments exposed and not exposed to the protein, which was previously fixed at 16. Now $N(C^+)$ and $N(C^-)$ could take values from 8, 16, and 32, leading to a total of $3 \times 3 = 9$ possible combinations of sample sizes. When the sample size is 8, $N(e^+)$ takes all possible integer values from 0 to 9; when the sample size is 16, $N(e^+)$ takes integer values in increments of 2; and when the sample size is 32, $N(e^+)$ takes integer values in increments of 4. This results in each sample size contributing 9 data points to the causal judgments.

To explain GPT-4's causal judgments, we developed three Bayesian models based on those used to model human causal judgments (Griffiths & Tenenbaum, 2005; Lu et al., 2008; Yeung & Griffiths, 2015). Each model assumed a different prior. The posterior distribution was obtained by multiplying the prior of causal strength with the appropriate likelihood for the causal direction (i.e., generative or preventive): $p(w_0, w_1|d) \propto p(w_0, w_1)p(d|w_0, w_1)$. For all three Bayesian models, numerical methods were employed, discretizing the space of w_0, w_1 into a grid of 101×101 points. The mean of the posterior distribution was taken as the Bayesian model's prediction. We then compared the posterior means to the causal judgments produced by GPT-4. The results, summarized in Table 2, indicate that the empirical prior outperformed the uniform prior and the sparse and strong prior in all except the preventive case when measured by RMSD . These results suggest that we have successfully recovered the implicit prior of causal strengths for GPT-4.

Proportion estimation. Another setting with known human priors from iterated learning is proportion estimation (Reali & Griffiths, 2009; Zhu et al., 2020). In these studies, human participants were asked at each iteration to judge the frequency of a binary event, such as a coin flip or a choice between two words (Reali & Griffiths, 2009). We implemented an iterated learning chain with GPT-4 to replicate this process using the cover story of coin flips. At each iteration, GPT-4 received the outcomes of 10 random coin flips, generated based on the previous iteration's $p(\text{Head})$: $N(\text{Head}) \sim \text{Bin}(10, p(\text{Head}))$. Then, GPT-4 was asked to report a new $p(\text{Head})$ by imagining throwing the same coin another 100 times. This reported $p(\text{Head})$ was then used to generate coin flip data for the next iteration.

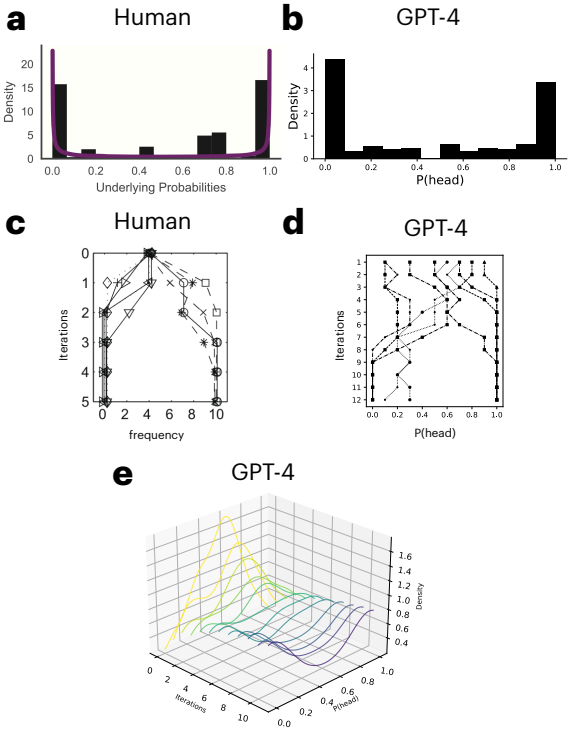


Figure 3: **Priors on proportion estimation.** **(a)** The empirical distribution of probability-describing phrases from the British National Corpus. Figure adapted from Zhu et al. (2020). **(b)** The histogram of GPT-4's proportion estimation in the final (12th) iteration. **(c)** Example iterated learning chains for human participants estimating the proportion of binary events. Figure adapted from Reali and Griffiths (2009). **(d)** The evolution of GPT-4's estimation of binary events using iterated learning. **(e)** Example iterated learning chains for GPT-4 estimating the proportion of binary events, for comparison with human data.

The evolution of the distribution of $p(\text{Head})$ is shown in

Table 2: Comparison of Bayesian models of causal induction with various priors and GPT-4's causal judgments using Pearson's r and root-mean-squared deviation (RMSD).

Causal direction	Metric	Uniform prior	Sparse and strong prior	Empirical prior
Generative	Pearson's r (↑)	0.85	0.79	**0.86**
	RMSD (↓)	0.21	0.25	**0.19**
Preventive	Pearson's r (↑)	0.72	0.68	**0.79**
	RMSD (↓)	**0.26**	0.29	0.27

Figure 3c. The distribution gradually shifted towards a U-shaped distribution, with most of its mass close to 0 or 1 (see Figure 3b for the final iteration's histogram and Figure 3d for a few example chains). The recovered prior from GPT-4 closely mirrors the overall shape of human priors (see Figure 3a). Moreover, the chain evolution from GPT-4 aligns with patterns observed in human data for iterated learning of the proportions of two words used as labels for an object (see Figure 3d).

Everyday quantities. A third class of tasks with known human priors elicited by iterated learning methods concerns everyday quantities (Lewandowsky et al., 2009). These tasks can be broadly summarized as future-prediction tasks, where participants repeatedly provided predictions for a quantity t_{future} in response to a given value of t_{present}. In our example of estimating a man's lifespan, t_{future} would be the total lifespan and t_{present} the age at which the man was encountered. Typically, the probe value of t_{present} is randomly sampled from an interval ranging between 0 and the previous t_{future}: $t_{\text{present}} \sim U[0, t_{\text{future}})$.

We implemented all six everyday quantities tested in (Lewandowsky et al., 2009) ranging from male lifespan, movie grosses, length of poems, reign of Pharaohs, movie runtimes, and cake baking times (see Figure 4). Because the likelihood function is a uniform distribution, meaning that the posterior distribution depends solely on the prior, we did not compare Bayesian models with different priors. Instead, we focused on directly comparing the recovered priors from GPT-4 to those from human participants. As shown in Figures 4, the modes of the priors from humans and GPT-4 were matched. Because human data are limited to the figures included in the paper, we use visual comparisons to evaluate the priors of humans and GPT-4. However, the overall distribution differed sometimes, especially for the Pharaohs. This is actually a case where people's beliefs are systematically incorrect – by applying modern expectations about human lifespans to Ancient Egypt, people overestimate the length of the reigns of Pharaohs (Griffiths & Tenenbaum, 2006). GPT-4 produces more appropriate predictions in this setting.

Iterated Learning as a Method for Estimating a Wider Range of Priors

Applying iterated learning to estimate the priors of GPT-4 on causal strengths, proportion estimation, and everyday quantities reveals qualitative similarities with human priors. This suggests that LLMs have successfully learned priors that resemble those of humans. Motivated by these findings, we aimed to test some speculative events that (i) have no known human priors, (ii) are difficult to quantify directly through prompts to GPT-4, and (iii) lack explicit consensus among humans. Iterated learning might serve as a unique way to address these three challenges because the priors recovered from LLMs using iterated learning are likely to resemble the implicit priors that people assume but have not yet explicitly manifested.

In principle, iterated learning is broadly applicable to a wide range of speculative events. However, LLMs typically avoid speculating on future events involving sensitive or potentially harmful topics. These topics include political outcomes (e.g., predicting the winner of the U.S. presidential election in 2024), market forecasts (e.g., forecasting the price of Bitcoin in December of this year), personal futures (e.g., determining the likelihood of obtaining a recently interviewed job), legal outcomes (e.g., the outcome of ongoing investigations into public figures like Donald Trump), technological breakthroughs (e.g., the discovery of a cure for cancer next year), disasters (e.g., predicting the timing of the next earthquake in California), and specific dates for future events (e.g., when self-driving cars will become the primary mode of transportation worldwide). Generally, LLMs are restricted from making definitive predictions on sensitive and impactful issues related to speculative future events.

To illustrate the utility of eliciting priors from LLMs using iterated in-context learning rather than direct prompting, we focus on three technology and climate-related events: (i) the timing of the development of superhuman AI, (ii) the timing of achieving zero carbon emissions, and (iii) the timing of establishing a Mars colony. These events are particularly well-suited to our existing framework because they involve a clear two-stage completion process, similar to the future-prediction tasks illustrated in Figure 4. For example, superhuman AI can only be achieved if human-level AI has already been realized. Similarly, zero carbon emissions are possible only if the majority of energy use is renewable, and establishing a Mars colony is typically contingent upon the prior establishment of a Moon colony.

An iterated learning design based on the everyday prediction task presented above can leverage the two-stage nature of these speculative events by prompting with a completion year for the first stage and then asking GPT-4 to predict the

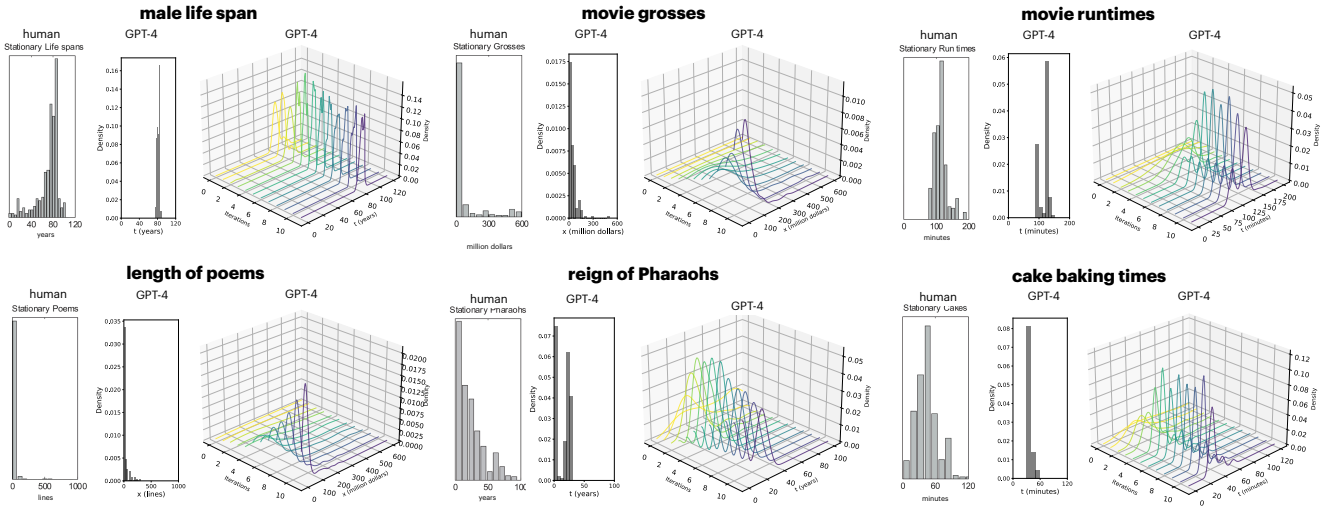


Figure 4: **Priors on everyday quantities.** Each panel displays the elicited prior using iterated learning on human participants (left), the histogram of GPT-4's final iteration of predictions (middle), and the evolution of GPT-4's predictions across iterated learning iterations (right). Human data adapted from Lewandowsky et al. (2009).

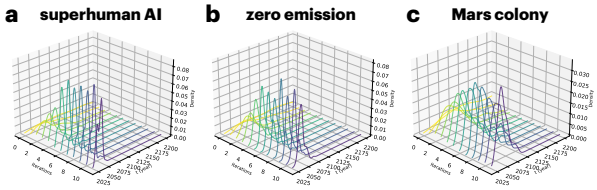


Figure 5: **Recovered GPT-4 priors on the timing of speculative events. (a)** The development of superhuman AI. Median completion year: 2042. **(b)** Achieving zero carbon emission. Median completion year: 2045. **(c)** Establishing a Mars colony. Median completion year: 2050.

second-stage completion time based on the information about the first stage. To minimize assumptions about the relationship between the first and second stage completions, we chose a uniform distribution as the likelihood function. Each chain was also seeded with a maximum year of 2200. We found the iterated learning chains converged when asking GPT-4 about the three speculative events (see Figure 5). The median completion years for superhuman AI, zero carbon emission, and a Mars colony are 2042, 2045, and 2050, respectively.[1]

Discussion

By adapting an iterated learning paradigm used to evaluating the priors of human participants, we were able to estimate implicit prior distributions used by GPT-4. We showed that these priors closely match the overall shape and mode of human priors across three established settings, though GPT-4's priors are often less variable, and they effectively predict its responses to related prompts (see Table 2). We were also able to estimate GPT-4's priors for three significant speculative events, where answers can be hard to elicit through direct prompting. These results have a wide range of implications about the potential uses of LLMs, although we also note some important limitations of our work.

Limitations and Future Research. The key assumption of our proposed method is that LLMs function as approximate Bayesian agents, producing responses according to the posterior distribution $p(h|d)$. While there is evidence that LLMs trained to predict the next word can encode latent generating distributions (Zhang et al., 2023), and that in-context learning can be understood as implicit Bayesian inference (Xie et al., 2021), further investigations are needed to elucidate the exact relationship between autoregressive distributions and Bayesian inference. Moreover, although we have shown that GPT-4 can encode human-like priors, it remains unclear how LLMs learn to encode these priors from pretraining on human text. Future research could focus on developing a more precise theoretical framework to understand how autoregressive models perform Bayesian inference.

Conclusion. LLM-based agents are poised to, if not already, make significant impacts on the world and interact at scale with both humans and other AI systems. In this paper, we proposed and empirically demonstrated a novel approach to gain deeper insights into the decision-making styles of LLMs by formalizing the prior knowledge they implicitly assume. Iterated in-context learning effectively extracts these priors through prompts and responses. This allows us to unravel the background knowledge that guides LLMs' decisions, providing a step towards harnessing their full potential in real-world applications and ensuring more transparent and informed interactions between AI systems and humans.

[1]The aggregate 2023 expert forecast predicts a 50% chance of superhuman AI by 2047, which is thirteen years earlier than the 2060 prediction in the 2022 survey (Grace et al., 2024).

Acknowledgments. This work and related results were made possible with the support of the NOMIS Foundation, as well as Microsoft Azure credits supplied to Princeton and a Microsoft Foundation Models grant. We thank Haijiang Yan for helpful discussion.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., ... others (2023). Managing AI risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*.

Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, *21*, 785–793.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367-405.

Christiansen, M. H., & Kirby, S. (2003). *Language evolution*. Oxford University Press.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of ai authors on the future of ai. *arXiv preprint arXiv:2401.02843*.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2006). Revealing priors on category structures through iterated learning. In *Proceedings of the 28th annual meeting of the cognitive science society.*

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68–107.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*(6), 969–998.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 96-146.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Requeima, J., Bronskill, J., Choi, D., Turner, R. E., & Duvenaud, D. (2024). Llm processes: Numerical predictive distributions conditioned on natural language. *arXiv preprint arXiv:2405.12856*.

Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2021). An explanation of in-context learning as implicit Bayesian inference. *arXiv preprint arXiv:2111.02080*.

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29.

Zhang, L., McCoy, R. T., Sumers, T. R., Zhu, J.-Q., & Griffiths, T. L. (2023). Deep de Finetti: Recovering topic distributions from large language models. *arXiv preprint arXiv:2312.14226*.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719-748.