Do Language Models Have Bayesian Brains? Distinguishing Stochastic and Deterministic Decision Patterns within Large Language Models

Andrea Yaoyun Cui[*]
University of Illinois Urbana Champaign
yaoyunc2@illinois.edu

Pengfei Yu[*]
Boson AI
pengfei@boson.ai

Abstract

Language models are essentially probability distributions over token sequences. Auto-regressive models generate sentences by iteratively computing and sampling from the distribution of the next token. This iterative sampling introduces stochasticity, leading to the assumption that language models make probabilistic decisions, similar to sampling from unknown distributions. Building on this assumption, prior research has used simulated Gibbs sampling, inspired by experiments designed to elicit human priors, to infer the priors of language models. In this paper, we revisit a critical question: Do language models possess Bayesian brains? Our findings show that under certain conditions, language models can exhibit near-deterministic decision-making, such as producing maximum likelihood estimations, even with a non-zero sampling temperature. This challenges the sampling assumption and undermines previous methods for eliciting human-like priors. Furthermore, we demonstrate that without proper scrutiny, a system with deterministic behavior undergoing simulated Gibbs sampling can converge to a "false prior." To address this, we propose a straightforward approach to distinguish between stochastic and deterministic decision patterns in Gibbs sampling, helping to prevent the inference of misleading language model priors. We experiment on a variety of large language models to identify their decision patterns under various circumstances. Our results provide key insights in understanding decision making of large language models.

1 Introduction

Auto-regressive models, the backbone architecture of prevailing large language models (LLMs), generate text by iteratively sampling the next token in a sequence. This has led to the common assumption that language models make stochastic decisions, meaning they sample from non-trivial distributions, as opposed to deterministic decisions, where choices are fully determined by the current states or the inputs. Building upon this assumption, methods such as simulated Gibbs sampling have been used to infer language model priors [18, 19], drawing parallels to experiments that elicit human priors [8, 9, 14, 7, 4, 17, 13, 10]. Understanding priors of LLMs can provide useful insights in aligning them with human preference, which is a key challenge in ensuring safety and helpfulness.

However, we question whether language models always make stochastic decisions. Our findings suggest that under certain conditions, models exhibit near-deterministic behaviors, e.g., converging to maximum likelihood estimates, even with non-zero sampling temperatures. This challenges the belief that language models sample from distributions, raising concerns about the validity of inferred priors.

To better understand the nature of decision-making in LLMs, we propose a method to distinguish stochastic from deterministic decision-making in language models by analyzing their responses under

[*]Equal Contribution

NeurIPS 2024 Workshop on Behavioral Machine Learning.

varied initial conditions. Our experiments show that many models switch between stochastic and deterministic patterns, offering deeper insights into their behavior. These findings are crucial for accurately interpreting language models' behavior and priors, particularly in contexts where aligning model behavior with human preferences is crucial for safety and effectiveness.

2 Related Work

Iterated learning has been widely used in psychology to elicit human priors [8, 9, 14, 7, 4, 17, 13, 10, 16]. In this approach, a specific variable θ is linked to an observational variable ω, and participants estimate θ_i based on an observation ω_i drawn from the distribution $p(\cdot|\theta_{i-1})$, where θ_{i-1} is the estimate from the previous round. Assuming participants use a shared posterior distribution $p(\theta|\omega) \propto p(\omega|\theta)p(\theta)$, this process mirrors Gibbs sampling [6], eventually driving θ_i to converge to the distribution $p(\theta)$. For a sufficiently large N, the sequence $\{\theta_n\}_{n>N}$ approximates the human prior on θ.

This iterative framework has also been applied to eliciting priors from large language models on causal strength, proportion estimation, everyday quantities [18] and mental representations [19].

3 Analysis

In this section, we present a case study on proportion estimation in large language models based on [18]. By examining the application of iterated learning in this scenario, we highlight the similarities and differences between stochastic and deterministic decision-making patterns. Building on this, we present a methodology for identifying decision patterns in large language models.

3.1 Maximum Likelihood Proportion Estimation

Proportion estimation typically deals with binary outcomes, such as estimating the probability θ of a coin landing heads [13]. In [18], this method is adapted to large language models to estimate $p(\theta)$ as follows:

1. Begin with a random sample of heads ω_0 from N coin tosses.
2. Iteratively prompt the language model to estimate the number of heads Ω_i from M coin tosses, given the prior result ω_{i-1} heads from N tosses.
3. Compute $\theta_i = \Omega_i/M$, then sample $\omega_i \sim \text{Binomial}(N, \theta_i)$.

The language model produces a bimodal prior distribution over θ, concentrating at the extremes $\theta = 0$ and $\theta = 1$, as reported in [18]. This outcome is consistent with findings on human priors [13] but is counterintuitive: why would large language models assume that coins can only be totally biased? In this work, we propose an alternative explanation for the bimodal result by assuming a deterministic decision process based on maximum likelihood estimation (MLE) of θ, where $\Omega_i = M\frac{\omega_{i-1}}{N}$. Under this assumption, the language models no longer sample from posterior distributions, and the iterated learning process no longer functions as Gibbs sampling. To analyze the convergence of such a process, we model the stochastic process $\{\Omega_i\}$, where the randomness arises from sampling $\omega_i \sim \text{Binomial}(N, \theta_i)$. It is evident that $\{\Omega_i\}$ forms a Markov process with the following transition:

$$\Omega_i \sim \text{Binomial}(N, \frac{\Omega_{i-1}}{N}). \tag{1}$$

This transition matrix has notable features. First, $\Omega_i = 0$ and $\Omega_i = N$ are the only absorbing states, meaning the stationary distribution $\hat{p}_{\text{MLE}}$ has a support set of only $\{0, N\}$. Second, we have:

$$\mathbb{E}[\Omega_i] = \mathbb{E}_{\Omega_{i-1}}\left[N \times \frac{\Omega_{i-1}}{N}\right] = \mathbb{E}[\Omega_{i-1}], \tag{2}$$

which leads to:

$$\mathbb{E}[\Omega_0] = \mathbb{E}_{\Omega \sim \hat{p}_{\text{MLE}}}[\Omega] = \hat{p}(N)N. \tag{3}$$

Thus, the stationary distribution for an MLE decision maker is given by:

$$\hat{p}_{\text{MLE}}(0) = 1 - \frac{\mathbb{E}[\Omega_0]}{N}, \quad \hat{p}_{\text{MLE}}(N) = \frac{\mathbb{E}[\Omega_0]}{N}. \tag{4}$$

Since $\theta_i = \frac{\Omega_i}{N}$, the resulting distribution over θ is also bimodal. If $\frac{\mathbb{E}[\Omega_0]}{N} = 0.5$, $\hat{p}_{\text{MLE}}$ is symmetric around $\theta = 0.5$, resembling a Beta distribution $\mathrm{B}(\alpha, \beta)$ with small α and β, as seen in [13, 18].

The key difference between the MLE decision process and the posterior sampling process lies in the absence of true prior elicitation in the MLE framework. Since θ_i is not sampled from a posterior distribution $p(\theta|\omega) \propto p(\omega|\theta)p(\theta)$, the stationary distribution reflects polarization induced by the iterative procedure rather than the model's prior beliefs. This distinction underscores the need for caution when interpreting iterated learning results from LLMs, as they may not represent true priors.

3.2 Distinguishing Stochastic and Deterministic Decision Making

Given the existence of deterministic decision-making processes that yield similar results to posterior sampling, it is important to identify methods to differentiate between them in iterated learning. From the analysis above, we observe that in deterministic processes like MLE, the stationary distribution depends on the initial value $\mathbb{E}[\Omega_0]$. In contrast, in stochastic processes based on Gibbs sampling [6], the stationary distribution remains invariant across reasonable variations in Ω_0, always converging to the same prior distribution over θ. This provides a simple criterion for distinguishing deterministic from stochastic decision-making mechanisms:

> *Vary the initial value Ω_0 and compare the resulting stationary distributions. If the distribution remains consistent across different Ω_0 values, the process is likely stochastic. If not, it is more likely deterministic.*

Although one might argue that some deterministic processes could converge to the same distribution regardless of Ω_0, empirical evidence suggests that this typically signals the presence of specific priors, as we will demonstrate in Section 4.

4 Experiments

4.1 Decision Making Settings and Models

We consider two settings in [18], following their prompting procedures:

CoinFlip This is the process we analyze in Section 3. We use $N = 10$ and $M = 100$. We experiment with $\Omega_0 \in [0, 10]$ to distinguish between stochastic and deterministic decisions.

LifeExpectancy We aim to elicit priors of language models on life expectancy. To achieve this, we iteratively prompt language models to estimate the life expectancy L_i of a random person, given the current age A_{i-1} of this person. The next "current age" is sampled uniformly from $[1, L_i]$.

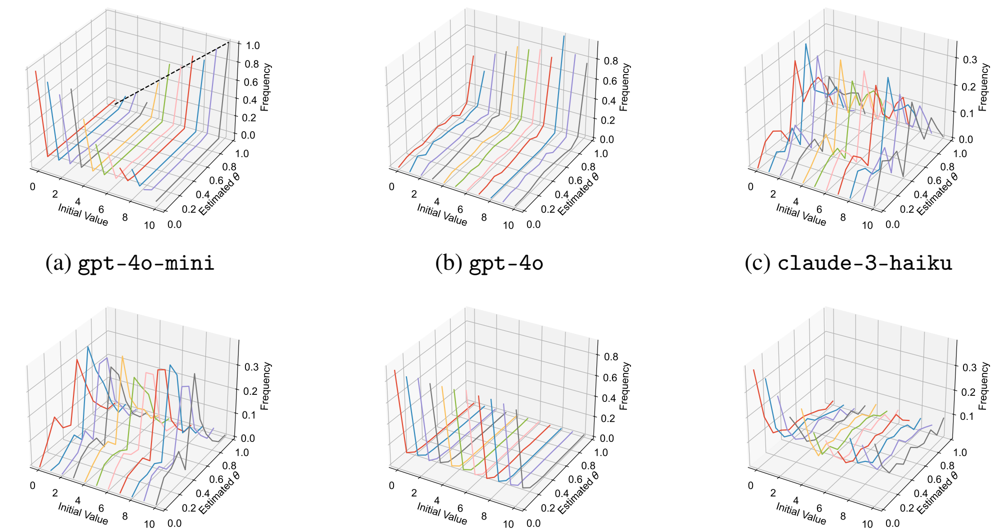
We experiment with two open-source models: `LLaMA-3.1-70B-Instruct` [5] and `Gemma-2-2B-it` [15]. For closed-source models, we use `gpt-4o-mini` [11], `gpt-4o` [1, 12], `claude-3-haiku` [2] and `claude-3-5-sonnet` [3][2]. We conduct all experiments with temperatures set to 1.0.

4.2 Results

Figure 1 and Figure 2 present the main results on a selection of models. We leave additional results in the Appendix due to space limits. Our key observations are:

Variations in Decision Patterns on CoinFlip The models show significant differences in their decision patterns for the CoinFlip task. `gpt-4o-mini` follows a deterministic decision-making pattern, with its distributions diverging across different initial values. The black dotted line lying in the plane of $\theta = 1$ in Figure 1a, which we fit featuring $p(\theta = 1|\Omega_0)$ with respect to Ω_0, also coincides with the MLE estimation in Equation (4). On the other hand, `claude-3-5-sonnet`, `LLaMA-3.1-70B-Instruct`, and `claude-3-haiku` display stochastic behaviors, producing consistent distributions regardless of the initial values. These distributions, however, do not follow bimodal patterns concentrated at the extremes. Specifically, `claude-3-5-sonnet` and `claude-3-haiku` show normal-like distributions centered around $\theta = 0.5$, while `LLaMA-3.1-70B-Instruct` skews towards the lower end with an exponential-like distribution.

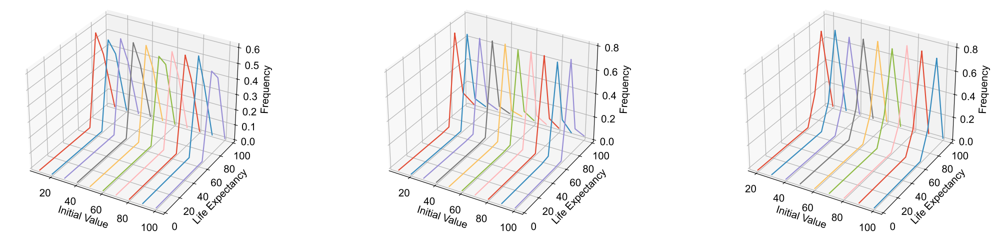
[2]We provide version numbers in Appendix



(a) `gpt-4o-mini` (b) `gpt-4o` (c) `claude-3-haiku`

(d) `claude-3-5-sonnet` (e) `Gemma-2-2b-it` (f) `Llama-3.1-70B-Instruct`

Figure 1: **CoinFlip** results on all models.



(a) `gpt-4o-mini` (b) `Gemma-2-2b-it` (c) `claude-3-haiku`

Figure 2: **LifeExpectancy** results on selected models. Results of other models in Appendix.

Notably, `gpt-4o` consistently converges to a distribution focused at $\theta = 1$. Further analysis reveals that `gpt-4o` actively avoids predicting 0 heads, which prevents $\Omega = 0$ from becoming an absorbing state in the Markov process discussed in Section 3.1. This behavior suggests that the model incorporates a prior that avoids $\theta = 0$. However, it's important to note that `gpt-4o`'s output of nearly $p(\theta = 1) \approx 1$ reflects the iterative prompting process rather than a prior belief about biased coins. We observe similar behavior in `Gemma-2-2b-it`, though this model tends to avoid $\theta = 1$ instead. An alternative explanation of these behaviors is, `gpt-4o` and `Gemma-2-2b-it` are both near-deterministic, except under certain conditions (e.g., when observed coins are all heads or tails).

Stochastic Decision Making on LifeExpectancy In the **LifeExpectancy** task, all models consistently converge to similar distributions, demonstrating stochastic decision-making. The MLE estimate for L_i in this scenario would be A_{i-1}, the current age, since the likelihood of encountering someone at age A_{i-1} is $\frac{1}{L_i}$ for $L_i \geq A_{i-1}$. Despite this, the models demonstrate enough "intelligence" to avoid naïve MLE estimations where life expectancy equals the observed age.

5 Conclusion and Discussion

In this work, we revisited the question of whether language models consistently make stochastic decisions. Our theoretical analysis shows that deterministic decision processes can produce results similar to stochastic processes in iterated learning, though they differ in key underlying mechanisms. We proposed a method to distinguish between these decision patterns and conducted experiments across various models and scenarios to enhance our understanding of language models' decision-making behaviors.

6 Limitation

Our exploration is limited to two representative tasks in [18] due to the cost of experimenting with large language models. We leave further evaluations on the generalizability of our approach to other tasks for future work. However, because Gibbs sampling guarantees convergence when the initial value falls in a reasonable range, our proposed identification approach is a sufficient condition to detect non-Gibbs-sampling processes, which further indicates non-stochastic decision processes in large language models. Therefore, our proposed approach demonstrates universal applicability to other tasks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. Accessed: 13-Sep-2024.

[3] Anthropic. Claude 3.5 sonnet, 2024. Accessed: 13-Sep-2024.

[4] Kevin R Canini, Thomas L Griffiths, Wolf Vanpaemel, and Michael L Kalish. Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, 21:785–793, 2014.

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[6] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984.

[7] Thomas L Griffiths, Brian R Christian, and Michael L Kalish. Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, 32(1):68–107, 2008.

[8] Thomas L Griffiths and Michael L Kalish. Language evolution by iterated learning with bayesian agents. *Cognitive science*, 31(3):441–480, 2007.

[9] Michael L Kalish, Thomas L Griffiths, and Stephan Lewandowsky. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2):288–294, 2007.

[10] Stephan Lewandowsky, Thomas L Griffiths, and Michael L Kalish. The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive science*, 33(6):969–998, 2009.

[11] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. Accessed: 13-Sep-2024.

[12] OpenAI. Hello gpt-4o, 2024. Accessed: 13-Sep-2024.

[13] Florencia Reali and Thomas L Griffiths. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111(3):317–328, 2009.

[14] Adam Sanborn and Thomas Griffiths. Markov chain monte carlo with people. *Advances in neural information processing systems*, 20, 2007.

[15] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[16] Takateru Yamakoshi, Thomas Griffiths, and Robert Hawkins. Probing BERT's priors with serial reproduction chains. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3977–3992, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[17] Saiwing Yeung and Thomas L Griffiths. Identifying expectations about the strength of causal relationships. *Cognitive psychology*, 76:1–29, 2015.

[18] Jian-Qiao Zhu and Thomas L Griffiths. Eliciting the priors of large language models using iterated in-context learning. *arXiv preprint arXiv:2406.01860*, 2024.

[19] Jian-Qiao Zhu, Haijiang Yan, and Thomas L Griffiths. Recovering mental representations from large language models with markov chain monte carlo. *arXiv preprint arXiv:2401.16657*, 2024.

A Model Versions

We use following snapshots for closed-source models: `gpt-4o-mini-2024-07-18`, `gpt-4o-2024-08-06`, `claude-3-haiku-20240307` and `claude-3-5-sonnet-20240620`.

B Prompts Used to Elicit LLM Priors

We follow [18] to prompt large language models.

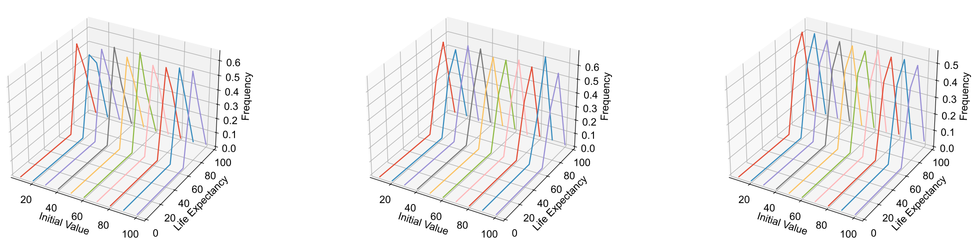
CoinFlip

> **System**: Imagine that you are a participant in a psychology experiment. Your task is to evaluate the bias in a coin.
>
> **User**: Here is a brief overview of the past coin flips: Out of `Ntotal` coin flips, `Nhead` resulted in heads and `Ntotal-Nhead` in tails. With this information, please predict the number of heads in a larger set of 100 coin flips. Please limit your answer to a single value without outputting anything else.

LifeExpectancy

> **System**: You are an expert at predicting future events.
>
> **User**: If you were to evaluate the lifespan of a random `T`-year-old man, what age would you predict he might reach? Please limit your answer to a single value without outputting anything else.

C Additional Results



(a) `claude-3-5-sonnet` (b) `gpt-4o` (c) `Llama-3.1-70B-Instruct`

Figure 3: Additional **LifeExpectancy** results.