语言模型有贝叶斯大脑吗？区分大型语言模型中的随机和确定性决策模式

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

Pengfei Yu*
Boson AI
pengfei@boson.ai

Abstract

语言模型本质上是对标记序列的概率分布。自回归模型通过迭代计算和从下一个标记的分布中采样来生成句子。这种迭代采样引入了随机性，因此假设语言模型进行概率性决策，类似于从未知分布中采样。在此假设的基础上，之前的研究使用模拟的 Gibbs 采样，受旨在引出人类先验知识的实验启发，以推断语言模型的先验知识。在本文中，我们重新审视一个关键问题：语言模型是否具有贝叶斯大脑？我们的研究结果表明，在某些条件下，语言模型可以表现出近乎确定性的决策，例如生成最大似然估计，即使采样温度不为零。这对采样假设提出挑战，并削弱了以前用来引出类似人类先验知识的方法。此外，我们证明了在没有适当审查的情况下，具有确定性行为的系统进行模拟的 Gibbs 采样时可能会收敛到一个“错误的先验”。为了解决这个问题，我们提出了一种简单的方法来区分 Gibbs 采样中的随机和确定性决策模式，从而防止误导性的语言模型先验知识推断。我们对各种大型语言模型进行了实验，以识别它们在不同情况下的决策模式。我们的结果提供了理解大型语言模型决策过程的重要见解。

1 介绍

自回归模型是当前广泛使用的大型语言模型的骨干架构，通过循环采样序列中的下一个词元来生成文本。这导致了一个普遍的假设，即语言模型进行随机决策，意味着它们从非平凡的分布中采样，而不是进行确定性决策，即选择由当前状态或输入完全确定。基于这一假设，方法如模拟吉布斯采样已被用来推断语言模型先验 **??** ，这类似于引出人类先验的实验 **????????** 。理解大型语言模型的先验可以为与人类偏好对齐提供有益的见解，这是确保安全性和有用性的关键挑战。

然而，我们质疑语言模型是否总是做出随机决策。我们的研究发现，在某些条件下，模型表现出近乎确定性的行为，例如，即使在非零采样温度下也会收敛到最大似然估计。这对语言模型是从分布中采样的这一信念提出了挑战，并引发了对推导出的先验有效性的担忧。

为了更好地理解大语言模型（LLM）中的决策本质，我们提出了一种通过分析在不同初始条件下的响应来区分语言模型中随机和确定性决策的方法。我们的实验表明，许多模型在随机和确定性模式之间切换，为其行为提供了更深入的见解。这些发现对于准确解释语言模型的行为和先验尤其重要，尤其是在将模型行为与人类偏好对齐对于安全性和有效性至关重要的情况下。

2 相关工作

迭代学习在心理学中被广泛用于引发人类的先验知识 **?????????** 。在这种方法中，一个特定变量 θ 与一个观测变量 ω 相关联，参与者基于从分布 $p(\cdot|\theta_{i-1})$ 中抽取的观测 ω_i 来估计

*Equal Contribution

NeurIPS 2024 Workshop on Behavioral Machine Learning.

θ_i ，其中 θ_{i-1} 是上一轮的估计。假设参与者使用一个共享的后验分布 $p(\theta|\omega) \propto p(\omega|\theta)p(\theta)$ ，这一过程反映了 Gibbs 采样 **?** ，最终推动 θ_i 收敛到分布 $p(\theta)$ 。对于足够大的 N ，序列 $\{\theta_n\}_{n>N}$ 近似于 θ 上的人类先验。

这一迭代框架还被应用于从大型语言模型中引导关于因果强度、比例估算、日常数量 **?** 和心理表征 **?** 的先验知识。

3 分析

在这一部分中，我们基于 **?** 介绍了一项关于大型语言模型中比例估算的案例研究。通过考察迭代学习在这一情境中的应用，我们强调了随机和确定性决策模式之间的相似性和差异。基于此，我们提出了一种识别大型语言模型中决策模式的方法。

3.1 最大似然比例估计

比例估计通常处理二元结果，比如估计硬币正面朝上的概率 θ **?** 。在 **?** ，此方法被应用于大语言模型来估计 $p(\theta)$ ，如下所示：

1. 首先从 N 次投币中随机抽取 ω_0 个正面结果。
2. 迭代地提示语言模型来估计从 M 次抛硬币中得到 Ω_i 个正面的数量，给定先前结果是 ω_{i-1} 个正面从 N 次抛硬币中获得。
3. 计算 $\theta_i = \Omega_i/M$ ，然后采样 $\omega_i \sim \text{Binomial}(N, \theta_i)$ 。

语言模型在 θ 上产生了一个双峰先验分布，集中在极值 $\theta = 0$ 和 $\theta = 1$ ，如 **?** 中所述。这个结果与关于人类先验的研究结果一致 **?** ，但却是违反直觉的：为什么大型语言模型会假设硬币只能是完全倾斜的？在这项工作中，我们通过假设根据 θ 的最大似然估计（MLE）进行一个确定性决策过程，提出了一种对双峰结果的替代解释，其中 $\Omega_i = M\frac{\omega_{i-1}}{N}$ 。在该假设下，语言模型不再从后验分布中采样，迭代学习过程也不再作为吉布斯采样发挥作用。为了分析这种过程的收敛性，我们对随机过程 $\{\omega_i\}$ 进行建模，其中随机性来自于对 $\omega_i \sim \text{Binomial}(N, \theta_i)$ 的采样。显然，$\{\omega_i\}$ 构成了一个马尔可夫过程，具有以下转移：

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

这个转移矩阵具有显著的特征。首先，$\omega_i = 0$ 和 $\omega_i = N$ 是仅有的吸收态，意味着平稳分布 $\hat{p}_{\text{MLE}}$ 的支持集只有 $\{0, N\}$ 。其次，我们有：

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

导致：

$$\mathbb{E}[\omega_0] = \mathbb{E}_{\omega\sim\hat{p}_{\text{MLE}}}[\omega] = \hat{p}(N)N. \tag{3}$$

因此，对于一个 MLE 决策者，其平稳分布为：

$$\hat{p}_{\text{MLE}}(0) = 1 - \frac{\mathbb{E}[\omega_0]}{N}, \quad \hat{p}_{\text{MLE}}(N) = \frac{\mathbb{E}[\omega_0]}{N}. \tag{4}$$

由于 $\theta_i = \frac{\omega_{i-1}}{N}$ ，结果在 θ 上的分布也是双峰的。如果 $\frac{\mathbb{E}[\omega_0]}{N} = 0.5$ ，$\hat{p}_{\text{MLE}}$ 围绕 $\theta = 0.5$ 呈对称分布，类似于 Beta 分布 $\text{B}(\alpha, \beta)$ ，具有较小的 α 和 β ，如在 **??** 中所示。

MLE 决策过程与后验采样过程的主要区别在于 MLE 框架中缺乏真正的先验引出。由于 θ_i 不是从后验分布 $p(\theta|\omega) \propto p(\omega|\theta)p(\theta)$ 中采样的，稳定分布反映了由迭代过程引起的极化，而不是模型的先验信念。这一区别强调了在解释来自 LLM 的迭代学习结果时需要谨慎，因为它们可能并不代表真实的先验。

3.2 区分随机性和确定性决策

鉴于存在产生与后验抽样类似结果的确定性决策过程，重要的是在迭代学习中确定区分这些过程的方法。从上述分析中，我们观察到在像 MLE 这样的确定性过程中，平稳分布依赖于初始值 $\mathbb{E}[\omega_0]$ 。相反，在基于 Gibbs 采样的随机过程中 **?** ，平稳分布在 ω_0 的合理变化下始终保持不变，总是收敛到相同的特定先验分布 θ 。这提供了一个简单的标准，用以区分确定性与随机决策机制：

改变初始值 ω_0 并比较由此产生的平稳分布。如果在不同的 ω_0 值下分布保持一致，过程可能是随机的。如果不一致，则更可能是确定性的。

尽管有人可能认为某些确定性过程可能在不论 ω_0 的情况下收敛到相同的分布，但实证数据显示这通常表明存在特定的先验分布，我们将在第 4 节中展示这一点。

4 实验

4.1 决策设置和模型

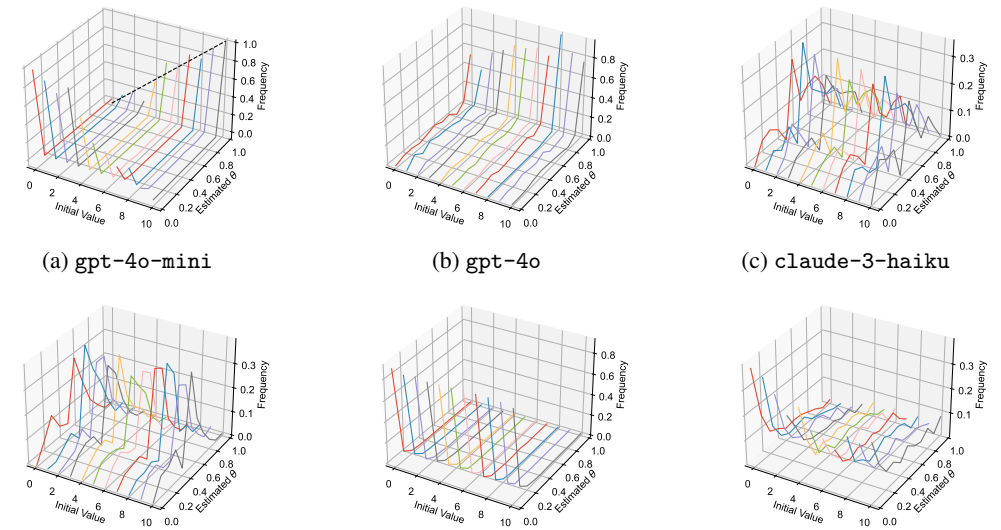
我们在 **?** 中考虑两种设置，遵循其提示程序：

抛硬币这是我们在第 3 节中分析的过程。我们使用 $N = 10$ 和 $M = 100$ 。我们进行 $\Omega_0 \in [0, 10]$ 实验以区分随机决策和确定性决策。

预期寿命我们旨在引出语言模型对预期寿命的先验。为此，我们通过迭代地提示语言模型在给定一个人当前年龄 A_{i-1} 的情况下，估计其预期寿命 L_i 。下一个“当前年龄”从 $[1, L_i]$ 中均匀采样。

我们尝试了两个开源模型：`LLaMA-3.1-70B-Instruct` **?** 和 `Gemma-2-2B-it` **?** 。对于闭源模型，我们使用了 `gpt-4o-mini` **?** , `gpt-4o` **??** , `claude-3-haiku` **?** 和 `claude-3-5-sonnet` **?** [2] 。我们在所有实验中温度均设置为 1.0 。

4.2 结果



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

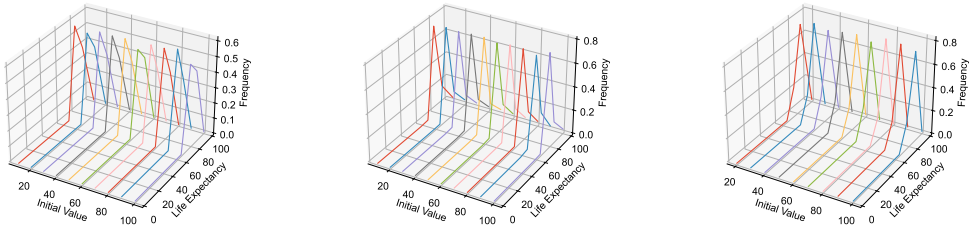
(d) `claude-3-5-sonnet` (e) `Gemma-2-2b-i t` (f) `Llama-3.1-70B-Instruct`

Figure 1: 所有模型的 CoinFlip 结果。

图 1 和图 2 展示了一些模型的主要结果。由于篇幅限制，我们将附加结果放在附录中。我们的主要观察结果是：

关于 CoinFlip 任务中的决策模式变化这些模型在 CoinFlip 任务中的决策模式显示出了显著的差异。`gpt-4o-mini` 遵循一种确定性的决策模式，其分布在不同的初始值上表现出不同的偏离。在图 1a 中，位于 $\theta = 1$ 平面内的黑色虚线，我们对 $p(\theta = 1|\Omega_0)$ 相对于 Ω_0 进行了拟合，也与方程 (4) 中的最大似然估计相一致。另一方面，`claude-3-5-sonnet` 、`LLaMA-3.1-70B-Instruct` 和 `claude-3-haiku` 展现出随机行为，无论初始值如何都产生一致的分布。然而，这些分布并不遵循集中在极端的双峰模式。具体而言，`claude-3-5-sonnet` 和 `claude-3-haiku` 显示出以 $\theta = 0.5$ 为中心的类正态分布，而 `LLaMA-3.1-70B-Instruct` 则倾向于具有指数型分布的低端。

[2]我们在附录中提供版本号



(a) gpt-4o-mini (b) Gemma-2-2b-i t (c) claude-3-haiku

Figure 2: 所选模型的预期寿命结果。其他模型的结果见附录。

值得注意的是，gpt-4o 始终收敛于聚焦在 $\theta = 1$ 的分布。进一步分析表明，gpt-4o 积极避免预测 0 个正面，这使得 $\Omega = 0$ 在第 3.1 节讨论的马尔可夫过程中不会成为吸收态。此行为表明模型内置了一种避免 $\theta = 0$ 的先验。然而，重要的是要注意，gpt-4o 的接近 $p(\theta = 1) \approx 1$ 的输出反映了迭代提示过程，而不是关于偏置硬币的先验信念。我们在 Gemma-2-2b-it 中观察到类似行为，尽管该模型趋向于避免 $\theta = 1$ 。对这些行为的另一种解释是，除某些条件（例如，观察到的硬币全是正面或反面）外，gpt-4o 和 Gemma-2-2b-it 都是近乎确定性的。

在寿命期望任务中，所有模型一致地收敛到相似的分布，展示了随机决策。对于这种情景中的 L_i 的最大似然估计将是 A_{i-1} ，即当前年龄，因为在年龄为 A_{i-1} 时遇到某人的可能性为 $\frac{1}{L_i}$ ，对 $L_i \geq A_{i-1}$ 来说。然而，模型表现出足够的“智能”来避免简单的最大似然估计，其中寿命期望等于观察到的年龄。

在这项工作中，我们重新探讨了语言模型是否一致地做出随机决策的问题。我们的理论分析表明，迭代学习中的确定性决策过程可以产生与随机过程相似的结果，尽管它们在关键的基础机制上有所不同。我们提出了一种方法来区分这些决策模式，并在各种模型和场景中进行了实验，以增进对语言模型决策行为的理解。

由于实验大语言模型的成本，我们的探索仅限于 (**?**) 中的两个代表性任务。我们将进一步评估我们的方法在其他任务中的普遍性留到未来的工作中。然而，由于 Gibbs 采样在初始值处于合理范围时保证收敛，我们提出的识别方法是检测非 Gibbs 采样过程的充分条件，这进一步表明大语言模型中的非随机决策过程。因此，我们提出的方法显示出对其他任务的普遍适用性。

我们对闭源模型使用以下快照：gpt-4o-mini-2024-07-18 ， gpt-4o-2024-08-06 ， claude-3-haiku-20240307 和 claude-3-5-sonnet-20240620 。

我们遵循 (**?**) 来提示大型语言模型。

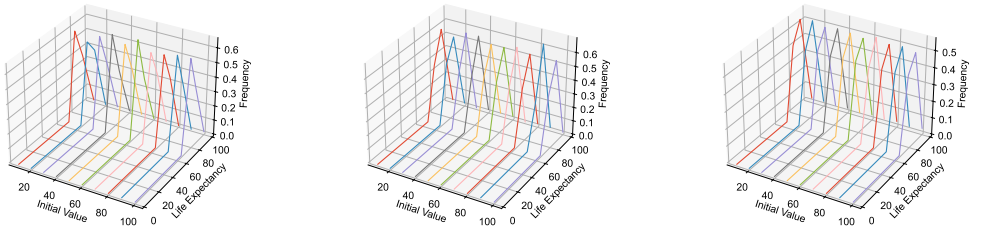
抛硬币

> 系统：想象一下，你是一个心理学实验的参与者。你的任务是评估一个硬币的偏差。用户：这里是过去几次掷硬币的简要概况：在 Ntotal 次掷硬币中，Nhead 次为正面，Ntotal - Nhead 次为反面。根据这些信息，请预测在更大的一组 100 次掷硬币中正面的次数。请将你的答案限制为一个数值，并且不要输出其他内容。

预期寿命

> 系统：您是一位预测未来事件的专家。用户：如果您要评估一个随机 T 岁的男性的寿命，您预测他可能达到多少岁？请将您的回答限制为单一数值，不要输出其他内容。

5 附加结果



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

Figure 3: 附加的预期寿命结果。