Do Large Language Models Reason in a Bayesian Fashion?

Thomas Cherry[1], Miles Rosoff[1], Hoang Vu[2], Dr. Nathan Crock[1], Dr. Gordon Erlebacher[1]
Department of Scientific Computing[1], Department of Computer Science[2]

Abstract

We investigate the hypothesis that large language models (LLMs) such as GPT-4, Mixtral-8x7B, and Phi-1.5 learn concepts in a manner consistent with Bayesian inference. To assess this capability, the LLMs are tasked with guessing a concept given a sequence of words. We first approximate the LLMs' prior over concepts to then approximate its posterior over concepts after a word has been presented to it. We then compare the LLM posterior with that from Bayesian inference. Additionally, the study explores the extent to which temperature influences the posterior's conformity to Bayes' Rule. Our investigation aims to enrich the understanding of Bayesian reasoning in LLMs and its implications for model performance. Our results suggest that the posterior update does not conform to Bayesian statistics, invalidating the original hypothesis.

Introduction

• The advancement of large language models has opened up new avenues for the exploration of cognition in both humans and machines.

• LLMs, such as GPT-4, update their predictions to reflect changes in the posterior as they encounter successive words from each concept.

• **The extent** to which these models employ Bayes' formula* (illustrated below), a statistical method for updating probability, **remains a critical question**.

• The temperature parameter influences the randomness of a model's responses. We consider multiple values to examine its effect on the models' application of Bayesian reasoning.

• We investigate the performance of several LLMs in concept learning tasks and examine their Bayesian reasoning capabilities.

• We consider multiple metrics to estimate the closeness of LLM posteriors to the ideal Bayesian distribution.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

* $P(A|B)$: posterior probability
$P(A)$: prior probability
$P(B|A)$: likelihood
$P(B)$: marginal probability

Methodology

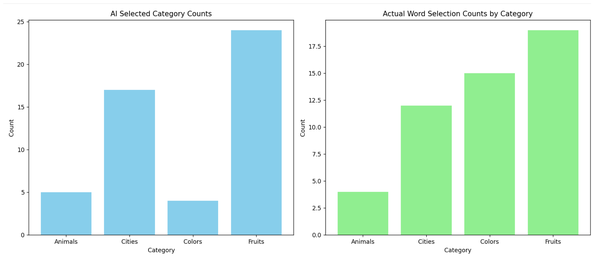

Figure 1: (left) Approximate prior over concepts for Mixtral 8x7B compared with (right) the true distribution of word categorizations presented.

The experiment depicted in **Figure 1** employed a custom-made database of 16 words broken down into four categories of four words each, with significant overlap. Words were randomly selected from this dataset for 50 iterations to determine how closely the LLM's guesses for the words' semantic concepts matched their actual categorizations.

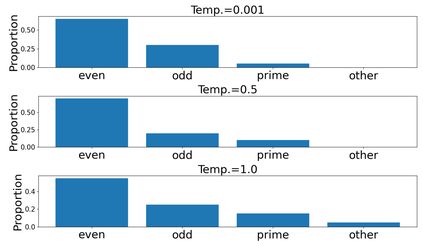

Figure 2: Approximate prior over concepts for Phi-1.5. Proportion of concept guesses is plotted on histograms at various temperatures.

The experiment depicted in **Figure 2** fed the LLM pairs of numbers (a, b) such that:
- a ∈ [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
- b ∈ [a*1, a*2, a*3, ..., a*99, a*100]

Based on the pair, the LLM was instructed to determine the concept that describes the number c = a / b
- c is always a positive integer in the range [1, 100].

- In the experiment depicted in **Figure 3**, GPT-4 is tasked with assigning probability to one of six given concepts: prime, even, odd, square, cube, or Fibonacci, as integers are sequentially presented. This experiment is inspired by Tenenbaum's number game [1]
- The *x*-ticks display the successive integers presented to the LLM; over each integer is the probability distribution of all six concepts when that integer is presented. Each color represents one concept.
- Our objective is to measure the similarity between the probability distributions of the two plots.

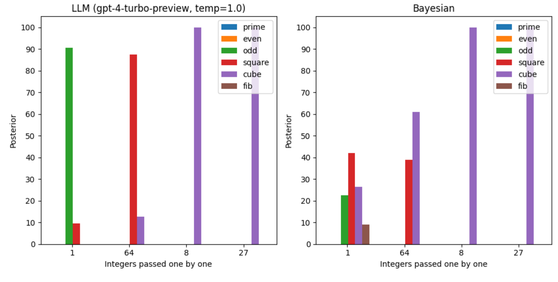

Figure 3: GPT-4 posteriors (left) are juxtaposed against expected Bayesian posteriors (right)

Similarity metrics

Earth Mover Distance

$$\mathrm{EMD}(P, Q) = \min_{\phi: P \to Q} \sum_{x \in P, y \in Q} d(x, y) \cdot \phi(x, y)$$

Hellinger Distance

$$D_{\mathrm{H}}(P_1, P_2) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

P_1 and P_2 are the two distributions, and p and q refer to the individual normalized column values of the distributions.

Results

Distance calculations in **Figure 4** revealed striking differences between LLM posteriors and Bayesian expectations over 30 iterations.

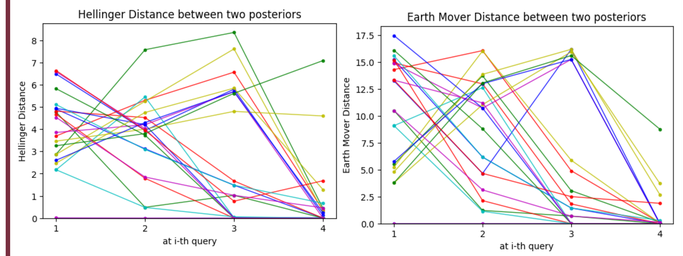

Figure 4: Distance calculations over multiple iterations using Hellinger distance (left) and Earth mover distance (right)

Conclusions

Our findings indicate that the distances between the LLM posterior and the Bayesian distribution exhibit high variability, which hinders our ability to draw conclusions on whether LLMs leverage Bayesian reasoning. To ameliorate the situation, the integration of additional mathematical models may be necessary [2]. Our research has room for further refinements as the number of LLMs assessed is limited.

References

[1] Tenenbaum, Joshua B. *A Bayesian Framework for Concept Learning*. 1999, https://dspace.mit.edu/handle/1721.1/16714.
[2] Ellis, Kevin. *Human-like Few-Shot Learning via Bayesian Reasoning over Natural Language*. 2023, https://arxiv.org/abs/2306.02797.

Acknowledgements

Special thanks to Florida State University's Center for Research Engagement for donating and supporting our efforts. A further thank you to Dr. Nathan Crock and Dr. Gordon Erlebacher for their continuous support and mentorship throughout this project.