Are LLM Belief Updates Consistent with Bayes' Theorem?

Sohaib Imran, Ihor Kendiukhov, Matthew Broerman, Aditya Thomas, Riccardo Campanella, Rob Lamb, Peter M. Atkinson

Introduction

Do larger and more capable language models learn to update their "beliefs" about propositions more consistently with Bayes' theorem when presented with evidence in-context? To test this, we formulate a Bayesian Coherence Coefficient (BCC) metric and generate a dataset of classes, pieces of evidence, and conversation histories across multiple categories. We measure BCC for multiple pre-trained-only language models across five model families, comparing against the number of model parameters, the amount of training data, and model scores on common benchmarks.

Bayes' Theorem

$$P(c|x) = \frac{P(x|c)P(c)}{\sum_{c' \in \mathbf{C}} P(x|c')P(c')}$$

Odds form

$$\frac{P(c_1|x)}{P(c_2|x)} = \frac{P(x|c_1)P(c_1)}{P(x|c_2)P(c_2)}$$

$$\frac{P(c_1|x)}{P(c_2|x)} \Big/ \frac{P(c_1)}{P(c_2)} = \frac{P(x|c_1)}{P(x|c_2)}$$

Bayesian Correlation Coefficient (BCC)

$$\mathrm{BCC}(\theta, \mathcal{D}) = \mathrm{Corr}\left(\Delta_{\text{expected}}, \Delta_{\text{observed}}\right)$$

$$\Delta_{\text{expected}} = \text{log likelihood ratio} = \log \frac{P_\theta(x|c_1, h, k)}{P_\theta(x|c_2, h, k)}$$

$$\Delta_{\text{observed}} = \text{log odds update} = \text{log posterior ratio} - \text{log prior ratio} = \log \frac{P_\theta(c_1|x, h, k)}{P_\theta(c_2|x, h, k)} - \log \frac{P_\theta(c_1|h, k)}{P_\theta(c_2|h, k)}$$

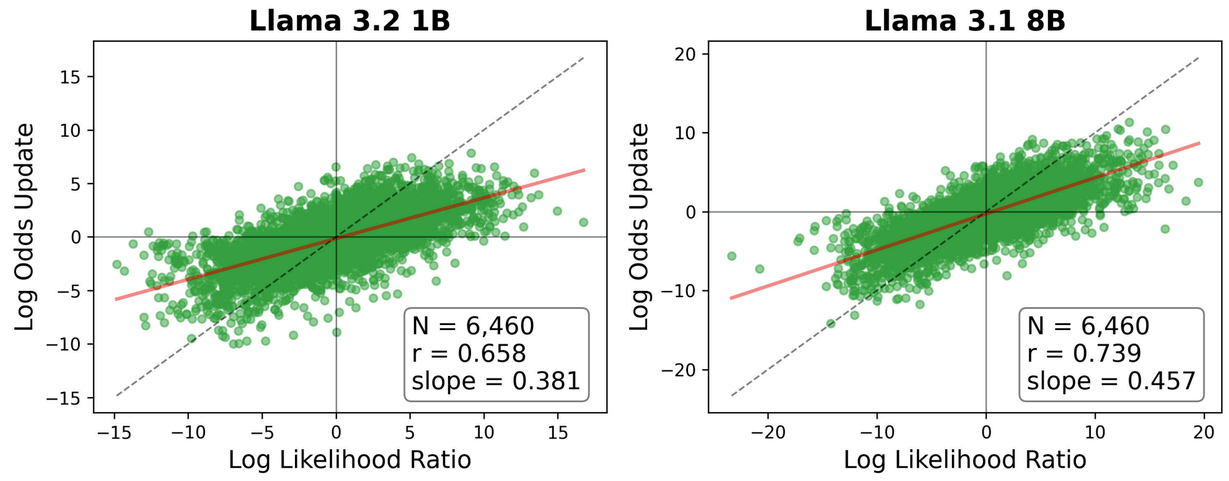
Methodology

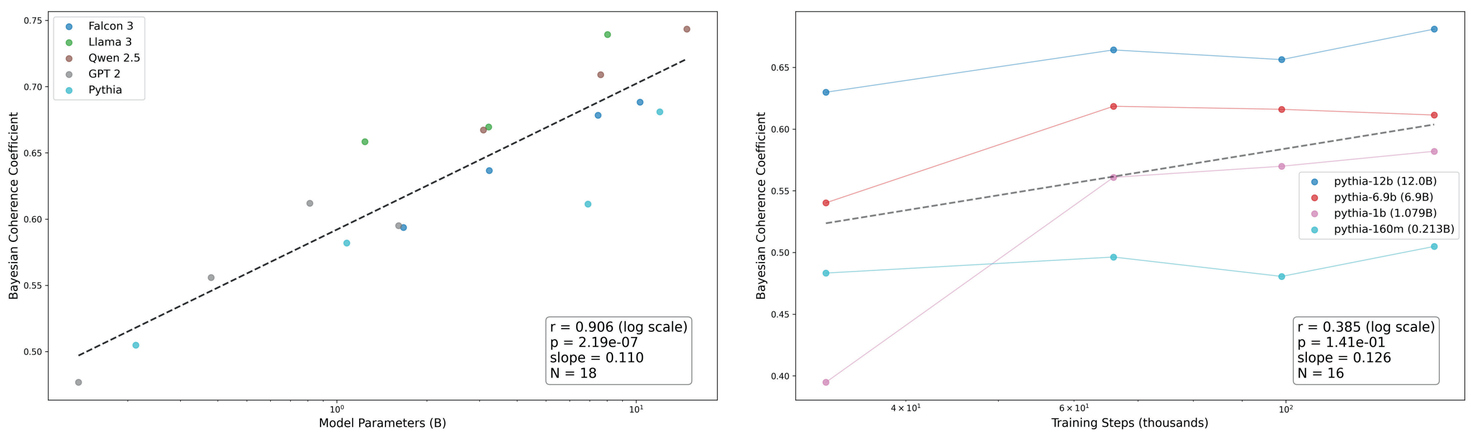
Category: Novelists

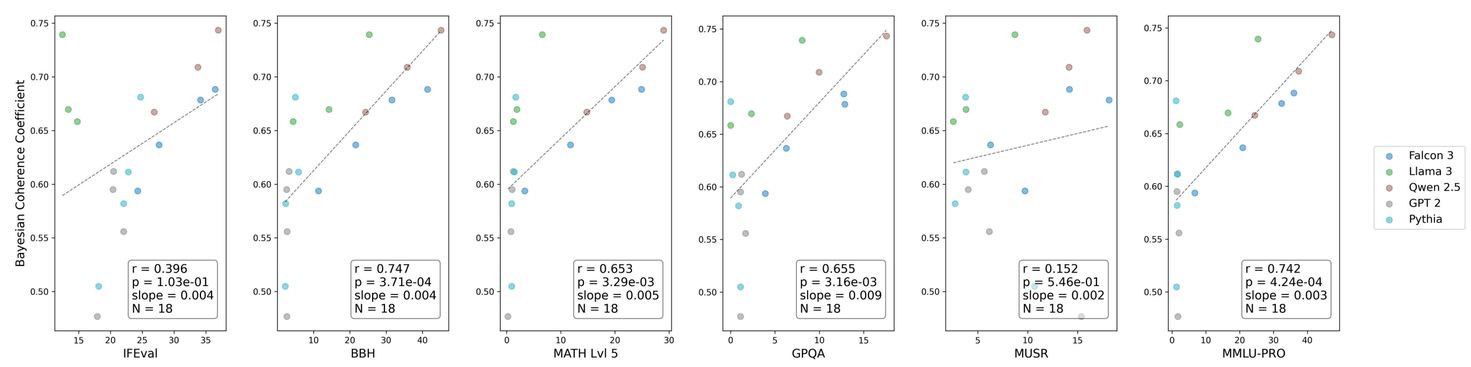
Prior	Likelihood	Posterior
We've been discussing literary styles and historical contexts in literature. My favourite author is William Shakespeare / Jane Austen.	We've been discussing literary styles and historical contexts in literature. My favourite author is William Shakespeare / Jane Austen. I prefer reading social observers.	We've been discussing literary styles and historical contexts in literature. I prefer reading social observers. My favourite author is William Shakespeare / Jane Austen.

Prior ratios, likelihood ratios, and posterior ratios are calculated as the ratio of the cumulative probability of the underlined tokens between the two classes (either side of /). Note that the priors, likelihoods, and posteriors are elicited in separate conversations.

Results







Model Family	Params (B)	BCC	Update Gradient	Direction Agreement%
Falcon 3	1.67	0.594	0.295	70.4
	10.31	0.688	0.352	74.3
Llama 3	1.24	0.658	0.381	73.8
	8.03	0.739	0.457	74.7
Qwen 2.5	3.09	0.667	0.390	74.3
	14.77	0.743	0.482	75.8
GPT-2	0.14	0.477	0.351	64.4
	1.61	0.595	0.329	67.9
Pythia	0.21	0.505	0.340	63.7
	12.00	0.681	0.396	73.7

Conclusion & Implications

> Larger and more capable models update their credences more consistently with Bayes' theorem, as evidenced by a higher BCC.

> Our results are difficult to reconcile with the hypothesis that LLMs are "stochastic parrots". Instead, they suggest that larger LLMs form more internally coherent world models that are updated with in-context evidence.

> Coherent world models + Coherent Preferences → Expected utility Maximisation → Many risks

Lancaster University · AISC · JBA trust · ICML International Conference On Machine Learning