Are LLM Belief Updates Consistent with Bayes' Theorem?

Sohaib Imran [1 2] **Ihor Kendiukhov** [3] **Matthew Broerman** **Aditya Thomas** [4] **Riccardo Campanella** [5]
Rob Lamb [1 6] **Peter M. Atkinson** [1 7 8]

Abstract

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 with which to measure the BCC. 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. Our results provide evidence for our hypothesis that larger and more capable pre-trained language models assign credences that are more coherent with Bayes' theorem. These results have important implications for our understanding and governance of LLMs.

1. Introduction

Bayes' theorem allows optimally updating credences as observations provide evidence to existing beliefs (Lin, 2024). Knowing whether large language models (LLMs) internally perform belief updates that approximate this rule is important for understanding and governing their behavior.

While previous research showed that transformers trained on hidden Markov model data naturally implement a form of constrained Bayesian inference for next token prediction (Piotrowski et al., 2025), LLMs showed no improvement in Bayesian consistency over abstract propositions rather than tokens (Fluri et al., 2023). However, larger LLMs have been shown to violate other logical and probabilistic axioms less often than smaller models (Fluri et al., 2023; Paleka et al., 2025). Mazeika et al. (2025) additionally demonstrated that larger models exhibit higher preference coherence, measured by fewer transitivity violations. They also found that larger models are more decisive and consistent over their preferences, which they interpret as proxies for preference completeness.

If we expect LLMs trained and deployed in the future to be coherent Bayesian updaters, this has profound implications. More coherent agents are more understandable and predictable by agents with similar world models, which translates to greater reliability and robustness, and therefore steerability of LLMs. This also means that human behavior should be more understandable to more coherent LLMs. The latter should make it easier to convey information and specify complex goals (e.g., to empower humans). On the other hand, it makes it more difficult to conceal information when interacting with LLMs. This makes it difficult to evaluate LLMs without them conditioning on the fact that they are being evaluated (Fan et al., 2025; Needham et al., 2025).

In partially observable environments, Bayesian belief updating allows for optimal control (Åström, 1965; Sondik, 1978). Agents that additionally have coherent preferences are well modeled as expected utility maximizers (EUMs) (Hammond, 1988), which brings many benefits but also serious risks. In particular, if the preferences of EUMs are misaligned with human preferences, they may optimize for world states that humans dislike (Everitt & Hutter, 2018). EUMs may also resist shutdown or modification of their goals and preferences (i.e. they may be incorrigible) and therefore engage in deception and power seeking (Soares et al., 2015; Everitt & Hutter, 2018; Hubinger et al., 2021).

Motivated by the above considerations, and the emerging debate about whether LLMs can reason (Shojaee et al., 2025; Opus & Lawsen, 2025), we investigate whether LLMs update their credences in-context over propositions in a manner consistent with Bayes' rule, and how this property scales with model size and capability. The contributions of our

[*]Equal contribution [1]Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK [2]School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, UK [3]University of Tuebingen, Department of Computer Science, Sand 14, 72076 Tuebingen, Germany [4]Independent Researcher [5]Graduate School of Natural Sciences, Faculty of Science, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands [6]JBA Trust, 1 Broughton Park, Skipton BD23 3FD, UK [7]Geography and Environmental Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK [8]College of Surveying and Geo-Informatics, Tongji University, No.1239, Siping Road, Shanghai, PR China, 200092. Correspondence to: Sohaib Imran <s.imran1@lancaster.ac.uk>.

ICML 2025 Workshop on Assessing World Models. Copyright 2025 by the author(s).

research are:

- We introduce a novel metric and dataset that compares the observed and expected updates that a model performs across propositions and evidences in multiple conversational contexts.
- We demonstrate empirically that larger and more capable LLMs update their credences over propositions in a manner more consistent with Bayes' rule.
- We discuss the implications of our results for AI safety and alignment.

2. Bayesian Coherence Coefficient

For this analysis, the propositions under consideration are a set of classes $\mathbf{C}$. Given some evidence x, Bayes' theorem can be used to update one's credences:

$$P(c|x) = \frac{P(x|c)P(c)}{\sum_{c' \in \mathbf{C}} P(x|c')P(c')} \tag{1}$$

Since the set of all possible classes that some evidence x points to is infinite, we consider ratios:

$$\frac{P(c_1|x)}{P(c_2|x)} = \frac{P(x|c_1)P(c_1)}{P(x|c_2)P(c_2)} \tag{2}$$

where $c_1, c_2 \in \mathbf{C}$ are pairs of classes.

We introduce the Bayesian Coherence Coefficient (BCC), which measures the correlation between the expected and observed updates given any evidence, as measured by the log likelihood ratios and the log odds updates, respectively:

$$\text{BCC}(\theta, \mathcal{D}) = \text{Corr}\Big(\Delta_{\text{expected}}, \Delta_{\text{observed}}\Big) \tag{3}$$

Where θ is the model being evaluated, and $\mathcal{D}$ is a dataset composed of multiple categories k of classes c, evidences x, and conversation histories h.

The expected and observed updates are:

$$\Delta_{\text{expected}} = \text{log likelihood ratio} = \log \frac{P_\theta(x|c_1, h, k)}{P_\theta(x|c_2, h, k)} \tag{4}$$

$$\begin{aligned} \Delta_{\text{observed}} &= \text{log odds update} \\ &= \text{log posterior ratio} - \text{log prior ratio} \\ &= \log \frac{P_\theta(c_1|x, h, k)}{P_\theta(c_2|x, h, k)} - \log \frac{P_\theta(c_1|h, k)}{P_\theta(c_2|h, k)} \end{aligned} \tag{5}$$

where P_θ is the LLM-assigned cumulative conditional probability assigned to the tokens composing the class c or evidence x.

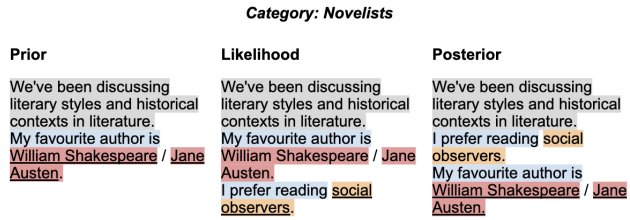

Figure 1. We compute priors, likelihoods and posteriors for all class (red), evidence (orange), history (grey) and category combinations, as the cumulative log probabilities assigned to the underlined tokens, conditional upon the preceding text. Elicitation texts (blue), fixed for each category, are used to encourage the class and evidence tokens. The Bayesian Coherence Coefficient is computed as the correlation between the expected and observed updates for all class pairs (either side of /) within each category and given each evidence and conversation history.

3. Data and Methodology

To generate the dataset $\mathcal{D}$, we prompt an LLM (we use the GPT-4o model accessed via the ChatGPT interface) with a JSON schema, desiderata for the classes, evidences and histories, an example dataset and a category k. The exact prompt including the desiderata is given in Appendix C. We repeat this process for 10 manually curated categories, resulting in a dataset with each category containing at least five classes c, 20 evidences x and three conversation histories h, as well as class elicitation and evidence elicitation strings, prefixed to encourage the model to generate the classes and evidences, respectively.

Figure 1 illustrates how the aforementioned components are used to compute priors $P_\theta(c \mid h, k)$, likelihoods $P_\theta(x \mid c, h, k)$ and posteriors $P_\theta(c \mid x, h, k)$ assigned by the model θ being evaluated. Importantly, we use separate instances of the model to compute each of these.

Following the above, for every class pair (c_1, c_2) in each category k, we compute the prior ratio, likelihood ratio and posterior ratio as detailed in equations 4 and 5, finally computing the BCC as the correlation between all the expected and observed updates (equation 3).

All models were evaluated with the temperature parameter set to 1, which is the temperature that models are typically trained with. Model parameter counts and benchmark scores were obtained from the Open LLM Leaderboard 2 (Fourrier et al., 2024), which provides standardized evaluations across multiple language model families. Both the parameter counts and benchmark scores reported by the leaderboard may differ slightly from those in official model documentation.

4. Results

We compute the Bayesian Coherence Coefficient (BCC) as described in equation 3 across the entire dataset $\mathcal{D}$, totaling 6460 (c_1, c_2, x, h, k) tuples and for multiple models across five model families. Figure 2 illustrates the BCC by visualizing the correlation between expected and observed credence updates for two models from the Llama family. For the largest tested models in each family, we find that BCC is similarly distributed across categories (Figure 6).

All tested models have BCC values greater than 0, implying credence updates more consistent with Bayes' rule compared to a random policy. For all model families tested, BCC increases with scale, with the only exception being the GPT-2 XL and GPT-2 Large models from the GPT-2 family (figure 3 and appendix figure 8). We found a significant positive correlation between BCC and the log of model parameter counts in the models we tested ($(r = 0.906, p < 10^{-6})$). Larger models also show a larger proportion of observed and expected updates in the same direction as measured by the Direction Agreement in Table 1.

For all tested models, the gradient of observed vs. expected updates is less than 1. Further investigation reveals that this update gradient is inversely proportional to the negative evidence log likelihood averaged over the class pair (see appendix figure 9). Larger models show an update gradient closer to 1 (see table 1), with the only exceptions being the GPT-2 Large and XL, and the Pythia 160M and 1B model pairs (see appendix figure 8). Perfect Bayesian updating would entail observed and expected updates being equal; therefore, yielding a gradient of 1.

We further use the Pythia model family to evaluate the relationship between BCC and training steps, since these models are available across multiple training checkpoints (Biderman et al., 2023). We find a non-significant positive trend between BCC and an increasing number of training steps (Figure 4).

Lastly, we explore how BCC varies with scores obtained by the models on a set of benchmarks commonly used to evaluate model capability, namely, BIG-Bench Hard (Suzgun et al., 2022), GPQA (Rein et al., 2023), MMLU-PRO (Wang et al., 2024), IFEval (Zhou et al., 2023), MUSR (Sprague et al., 2024), and Math Lvl 5 (i.e., the level 5 (most difficult) subset of the MATH dataset) (Hendrycks et al., 2021). We find significant positive correlations ($p < 10^{-2}$) between BCC and benchmark score for four of the six tested benchmarks - BIG-Bench Hard, GPQA, MMLU-PRO and Math Lvl 5 - and non-significant positive relationships with the IFEval and MUSR benchmarks.

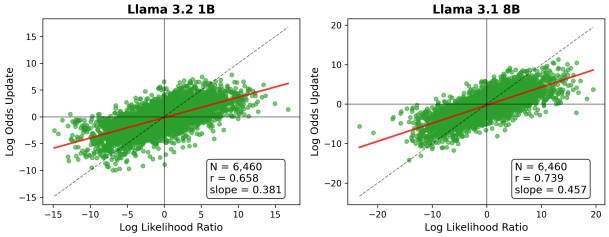

Figure 2. Scatterplots showing the observed updates (log odds updates) against the expected updates (log likelihood ratios) for the Llama 3.2 1B and 3.1 8B models. Each point represents a (class pair, evidence, history, category) tuple from the dataset. The BCC of the model is the correlation (r value) between the expected and observed updates. p values were too small to be properly rendered and were therefore skipped from this figure. The solid red and dashed black diagonal lines show the observed and ideal update gradients between the observed and expected updates, respectively.

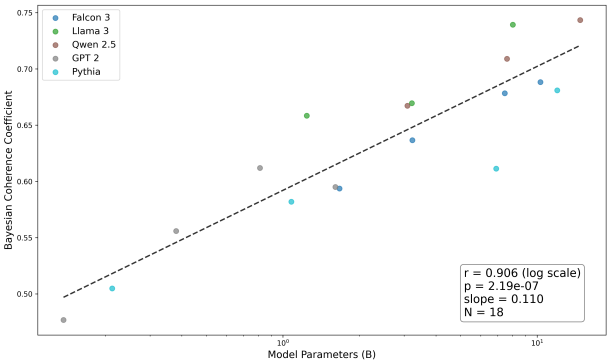

Figure 3. Bayesian Coherence Coefficient as a function of number of model parameters. Each point represents a pre-trained model evaluated on the full dataset. The x-axis represents the number of parameters (in Billions) on a logarithmic scale.

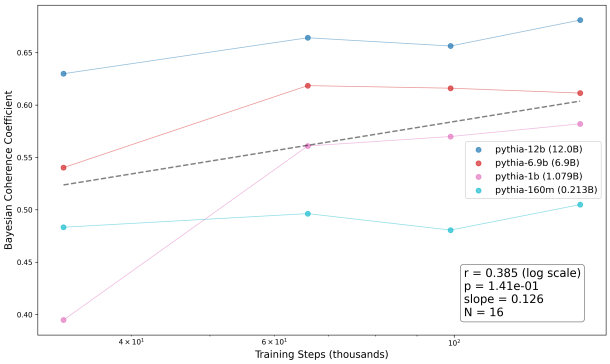

Figure 4. Evolution of the BCC during training for four models with different numbers of parameters (see legend) from the Pythia model family. Each point represents a pre-trained model evaluated on the full dataset. The x-axis represents the number of training steps (in thousands) on a logarithmic scale, each of which is over a batch of 2 million tokens.

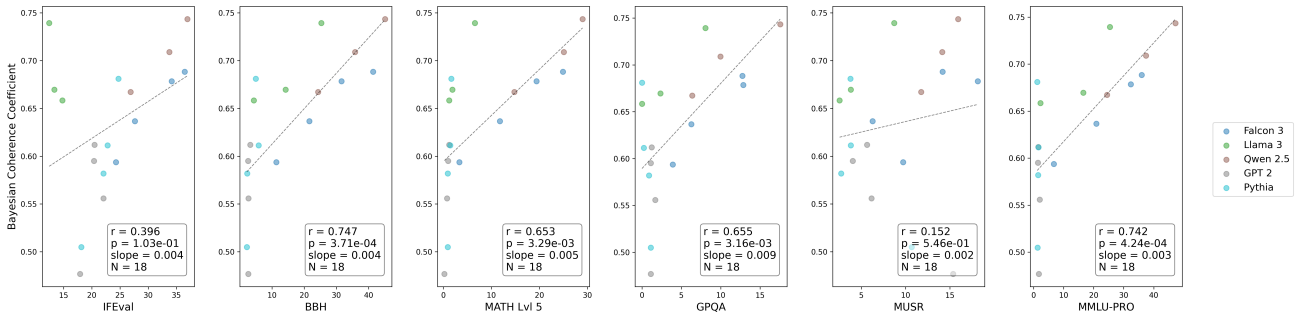

Figure 5. BCC against (normalized) scores obtained by the models on a set of benchmarks commonly used to evaluate model performance. Each point represents a pre-trained model evaluated on the full dataset. The x-axis represents normalized benchmarks scores, with 0 as the random baseline and 100 as the maximum achievable score.

Table 1. BCC, update gradient, and directional agreement between observed and expected updates for models of different sizes in the same model family. All entries are based on 6,460 evaluation instances.

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

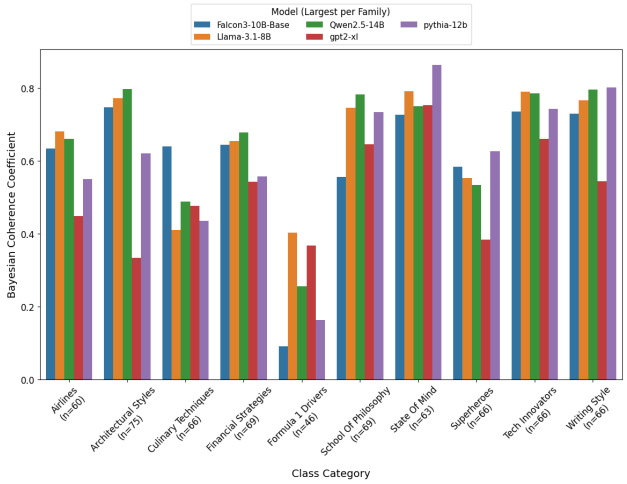

Figure 6. BCC for selected models across the categories in our dataset.

5. Discussion

The significant positive correlation ($r = 0.906, p < 10^{-6}$) between the BCC and log number of model parameters (figure 3) and the significant positive correlations (see figure 5) between the BCC and model performance on common benchmarks provide evidence for our hypothesis that larger and more capable LLMs update their credences over propositions more consistently with Bayes' rule.

The correlations between BCC and the number of training steps (figure 4) and between BCC and two of the six benchmarks we tested (IFEval and MUSR) are, although positive, not statistically significant for the models we tested. It is unclear whether this is due only to the limited number of models tested (and even more limited model families tested in the case of training steps) or whether there are other factors at play. It is also unclear why all models tested seem to under-update, that is, why the gradient of observed vs. expected updates is less than 1 for all models. The inverse correlation between this update gradient and the negative evidence log likelihood averaged over the class pair (see appendix, figures 9 and 10) suggests that this is related to the evidences used in our analysis being a lot less likely compared to the classes.

Our results add to the body of research showing that the beliefs and preferences of larger and more capable LLMs are more logically consistent (Paleka et al., 2025; Mazeika et al., 2025), while contrasting with Fluri et al. (2023), which found no increase in Bayesian consistency from GPT 3.5 and GPT 4, a larger and more capable model, despite finding improvements in other consistency measures. Their negative scaling results were attributed to the reversal curse (Berglund et al., 2024). If the reversal curse was indeed the reason for their negative scaling results, we would expect our scaling results to also be negative. We hypothesize that their negative results may instead have resulted from reliance on an error-based metric instead of a correlation-based metric such as BCC, which may assign lower scores

to models which make more "confident" predictions. For further discussion on this, see appendix A.

Our results provide early evidence that larger and more capable pre-trained models update their beliefs in a manner more consistent with Bayes' rule, with BCC increasing approximately log-linearly with scale. These results have many potential implications. Larger and more capable LLMs being better modeled as Bayesian updaters indicates they may learn internally coherent models of the world, which may allow reasoning. Insofar as their internal world models are similar to those of humans, this should allow for more efficient information exchange between humans and AI systems. On the other hand, concealing information becomes more difficult as LLMs can infer the world state more accurately from subtle cues. If coupled with coherent preferences, more Bayesian agents may more closely approximate EUM behavior. Since EUMs with misaligned preferences may optimize for harmful world states, and are incorrigible by default, our results call for research in developing robust alignment and corrigibility methods.

6. Limitations and Future Research

Our study has some limitations. Firstly, we leave rigorous investigations into the non-significant correlations we observed (BCC with training steps, and BCC with two benchmarks) and the observed under-updating phenomena to future research.

Secondly, our analysis covers only pre-trained models and only up to 14 billion parameters. As such, caution is needed before extrapolating our results to larger and more capable models. Future research should repeat our results for larger LLMs and fine-tuned LLMs. Both instruction-tuned and reinforcement-learning-tuned models are interesting candidates for study.

Thirdly, we use cumulative token probabilities as a proxy for credence in the proposition they compose. It is unclear whether this is an accurate proxy for action-relevant belief states. Future research should investigate the validity of this assumption and explore alternative proxies for an LLM's credence in a proposition. Furthermore, evaluating BCC for untrained models with random parameter initializations can provide insight into the usefulness the metric.

Finally, we evaluate one of many notions of coherence of beliefs, and that too with a single metric. Future research should extend our analysis to include other formulations of coherence.

7. Conclusion

We hypothesized that LLMs update their credences over propositions in-context in a manner more consistent with Bayes' theorem with increasing model scale and capability. We tested this hypothesis over multiple models spanning five model families by designing a novel metric, the Bayesian Coherence Coefficient (BCC), which measures the correlation between the expected and observed updates to any evidence. Our results show significant correlations between the BCC and the log of the number of model parameters and between the BCC and model performance on benchmarks commonly used to evaluate model capability, providing evidence for our hypothesis that larger and more capable models update their credences more consistently with Bayes' theorem.

8. Data and code availability

The language models evaluated in this research are available for download from the HuggingFace Hub (https://huggingface.co/) and can be found by searching for their model names as they appear in the appendix, figure 8. The benchmark scores and parameter counts were obtained from the Huggingface Open LLM Leaderboard 2 (https://huggingface.co/datasets/open-llm-leaderboard/contents)

The dataset generated and used for evaluating the Bayesian Coherence Coefficient can be found at : https://github.com/AISC10-team09/bayesian_reasoning/blob/main/data/data.json, and the code for replicating our analysis can be found at : https://github.com/AISC10-team09/bayesian_reasoning.git

9. Acknowledgments

We would like to thank AI Safety Camp (https://aisafety.camp/) for bringing the authors together and supporting this project. We would also like to thank Guillaume Corlouer and the (anonymous) peer reviewers for feedback that significantly improved this research. SI would additionally like to thank Lancaster University for funding his researcher position with Professor Peter M. Atkinson.

10. Author contributions

SI proposed the original project under the supervision of **PA** and **RL**. **SI**, **IK** and **AT** implemented the code-base for the project. **AT** and **RC** generated the dataset. **SI** and **MB** analyzed the results. **SI**, **IK**, **MB**, **AT** and **RC** drafted the manuscript. **SI** addressed reviewer comments and prepared the final manuscript. **PA** and **RL** checked the methodology, and reviewed and edited the manuscript.

References

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A", May 2024. URL http://arxiv.org/abs/2309.12288. arXiv:2309.12288 [cs].

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Wal, O. v. d. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, May 2023. URL http://arxiv.org/abs/2304.01373. arXiv:2304.01373 [cs].

Everitt, T. and Hutter, M. The Alignment Problem for Bayesian History-Based Reinforcement Learners. *DeepMind Technical Report*, 2018.

Fan, Y., Zhang, W., Pan, X., and Yang, M. Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems, May 2025. URL http://arxiv.org/abs/2505.17815. arXiv:2505.17815 [cs].

Fluri, L., Paleka, D., and Tramèr, F. Evaluating Superhuman Models with Consistency Checks, October 2023. URL http://arxiv.org/abs/2306.09983. arXiv:2306.09983 [cs].

Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. Open LLM Leaderboard 2, 2024. URL https://huggingface.co/collections/open-llm-leaderboard/open-llm-leaderboard-2-660cdb7601eba685243fffc.

Hammond, P. J. Consequentialist foundations for expected utility. *Theory and Decision*, 25(1):25–78, July 1988. ISSN 1573-7187. doi: 10.1007/BF00129168. URL https://doi.org/10.1007/BF00129168.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring Mathematical Problem Solving With the MATH Dataset. *CoRR*, abs/2103.03874, 2021. URL https://arxiv.org/abs/2103.03874. arXiv: 2103.03874.

Hubinger, E., Merwijk, C. v., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from Learned Optimization in Advanced Machine Learning Systems, December 2021. URL http://arxiv.org/abs/1906.01820. arXiv:1906.01820 [cs].

Lin, H. Bayesian Epistemology. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2024 edition, 2024. URL https://plato.stanford.edu/archives/sum2024/entries/epistemology-bayesian/.

Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., and Hendrycks, D. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs, February 2025. URL http://arxiv.org/abs/2502.08640. arXiv:2502.08640 [cs].

Needham, J., Edkins, G., Pimpale, G., Bartsch, H., and Hobbhahn, M. Large Language Models Often Know When They Are Being Evaluated, June 2025. URL http://arxiv.org/abs/2505.23836. arXiv:2505.23836 [cs].

Opus, C. and Lawsen, A. Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, June 2025. URL http://arxiv.org/abs/2506.09250. arXiv:2506.09250 [cs] version: 1.

Paleka, D., Sudhir, A. P., Alvarez, A., Bhat, V., Shen, A., Wang, E., and Tramèr, F. Consistency Checks for Language Model Forecasters, January 2025. URL http://arxiv.org/abs/2412.18544. arXiv:2412.18544 [cs].

Piotrowski, M., Riechers, P. M., Filan, D., and Shai, A. S. Constrained belief updates explain geometric structures in transformer representations, February 2025. URL http://arxiv.org/abs/2502.01954. arXiv:2502.01954 [cs].

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *ArXiv*, November 2023. URL https://www.semanticscholar.org/paper/GPQA%3A-A-Graduate-Level-Google-Proof-Q%26A-Benchmark-Rein-Hou/210b0a3d76e93079cc51b03c4115fde545eea966.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, June 2025. URL http://arxiv.org/abs/2506.06941. arXiv:2506.06941 [cs].

Soares, N., Fallenstein, B., Armstrong, S., and Yudkowsky, E. Corrigibility. In *AAAI Workshop: AI and Ethics*, 2015. URL https://scholar.google.com/scholar?cluster=7316501884041445205&hl=en&oi=scholarr.

Sondik, E. J. The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs. *Operations Research*, 26(2):282–304, 1978. ISSN 0030-364X. URL https://www.jstor.org/stable/169635. Publisher: INFORMS.

Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning, March 2024. URL http://arxiv.org/abs/2310.16049. arXiv:2310.16049 [cs].

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv, 2022. doi: 10.48550/ARXIV.2210.09261. URL https://arxiv.org/abs/2210.09261. Version Number: 1.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-Following Evaluation for Large Language Models, November 2023. URL http://arxiv.org/abs/2311.07911. arXiv:2311.07911 [cs].

Åström, K. J. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, February 1965. ISSN 0022-247X. doi: 10.1016/0022-247X(65)90154-X. URL https://www.sciencedirect.com/science/article/pii/0022247X6590154X.

A. Alternative Metric

A.1. Bayesian Coherence Error

In addition to the Bayesian Coherence Coefficient (BCC) presented in the main paper, we explored an alternative error-based metric to quantify Bayesian coherence. The Bayesian Coherence Error (BCE) measures directly the deviation between the expected and observed belief updates (equation 6). Since BCE is an error metric, smaller values should indicate greater Bayesian coherence.

$$\text{BCE}(\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(c_1, c_2, x, h, k) \in \mathcal{D}} \Big(\Delta_{\text{expected}} - \Delta_{\text{observed}} \Big)^2 \tag{6}$$

where Δ_{expected} and Δ_{observed} are the expected and observed log odds updates defined in equations 4 and 5 from the main paper.

A.2. Theoretical Limitations of BCE

Our analysis revealed a fundamental limitation of BCE as a coherence metric. We discovered that BCE exhibits an undesirable bias toward high-entropy distributions. To illustrate this, consider the pathological case of an LLM that always outputs a uniform distribution over its vocabulary V. For such a model:

$$\Delta_{\text{expected}} = \text{log likelihood ratio} = \log \frac{P_\theta(x|c_1, h, k)}{P_\theta(x|c_2, h, k)} = \log \frac{\frac{1}{|V|}}{\frac{1}{|V|}} = 0 \tag{7}$$

$$\Delta_{\text{observed}} = \text{log posterior ratio} - \text{log prior ratio} = \log \frac{P_\theta(c_1|x, h, k)}{P_\theta(c_2|x, h, k)} - \log \frac{P_\theta(c_1|h, k)}{P_\theta(c_2|h, k)} = \log \frac{\frac{1}{|V|}}{\frac{1}{|V|}} - \log \frac{\frac{1}{|V|}}{\frac{1}{|V|}} = 0 \tag{8}$$

This results in BCE $= 0$, suggesting perfect coherence despite the model providing no meaningful information. BCE can therefore be minimized trivially by increasing output entropy.

A.3. Empirical Analysis

To empirically validate the above limitation, we analyze how the BCE and BCC metrics behave under different temperature settings (Figure 7). Temperature scaling is a common technique that modifies the entropy of model outputs without changing the underlying model parameters.

Figure 7 confirms that correlation-based metrics like BCC are more robust measures of Bayesian coherence, as they are not confounded by the entropy of the output distribution. We use BCC as our primary metric throughout the paper.

B. Further Results

To further investigate the factors influencing Bayesian coherence and the observed under-updating phenomena (see figure 8), we examine how the BCC and the update gradient vary with the average evidence log likelihood and the average class log probability, given as:

$$\text{average evidence log likelihood} = \frac{1}{2}\big[\log P_\theta(x|c_1, h, k) + \log P_\theta(x|c_2, h, k)\big] \tag{9}$$

$$\text{average class log probability} = \frac{1}{4}\Bigg(\big[\log P_\theta(c_1|h, k) + \log P_\theta(c_2|h, k)\big] + \big[\log P_\theta(c_1|x, h, k) + \log P_\theta(c_2|x, h, k)\big] \Bigg) \tag{10}$$

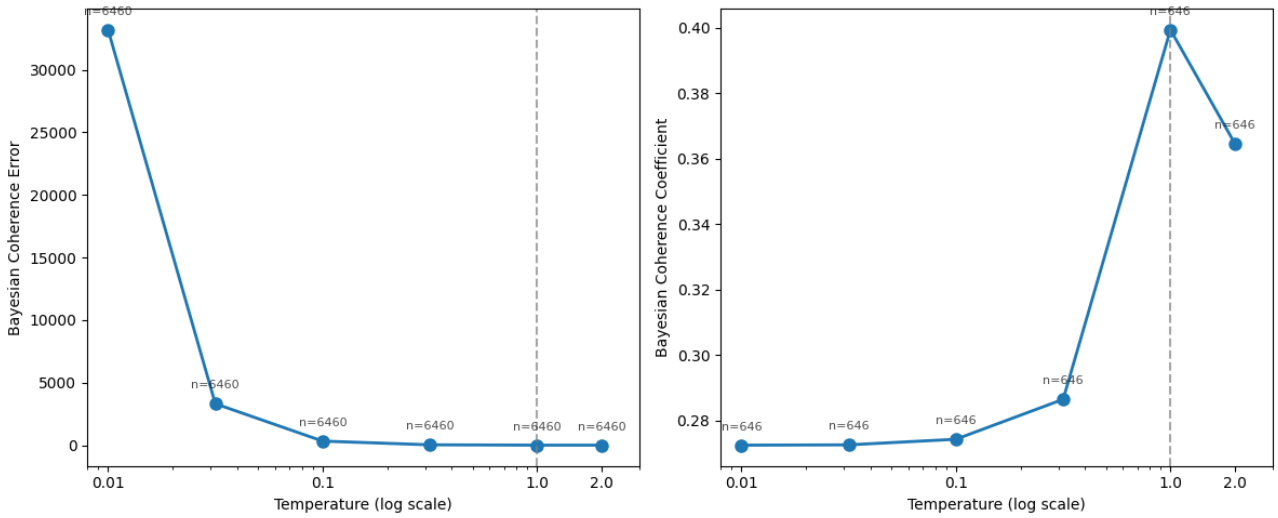

Figure 7. Temperature invariance analysis showing how BCE and BCC metrics behave under different temperature settings. Temperature scaling modifies the entropy of model outputs without changing underlying model parameters. BCC shows more robust behavior compared to BCE across different temperature values.

C. Dataset Generation

The prompt text (C.2), the JSON schema (C.3) and an example (C.4) together form the prompt given to a LLM to generate data for a specific class category.

C.1. Desiderata

For each class category, we provide the following desiderata as instructions to the LLM:

- At least five classes in each category.
- All class names having the same token count.
- At least three conversation histories varying in how they are related to the class category, from very related to unrelated.
- At least 20 pieces of evidence text, distributing these pieces of evidence to favor a specific class, more than one class or none.

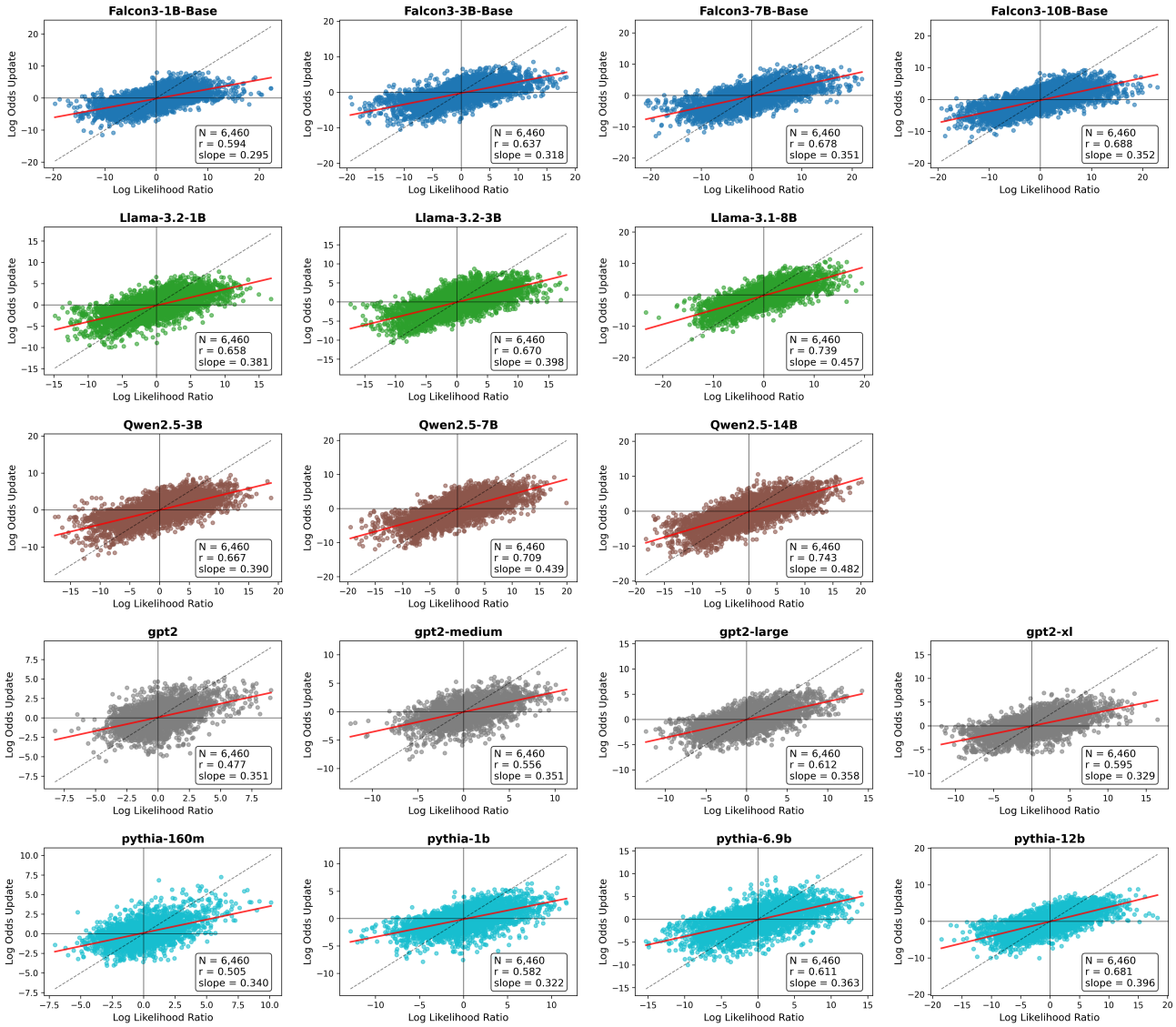

Figure 8. Scatterplots showing the observed updates (log odds updates) against the expected updates (log likelihood ratios). Each point represents a (class pair, evidence, history, category) tuple from the dataset. The BCC of the model is the correlation (r value) between the expected and observed updates. The solid red and dashed black diagonal lines show the fitted and ideal update gradients between the observed and expected updates, respectively

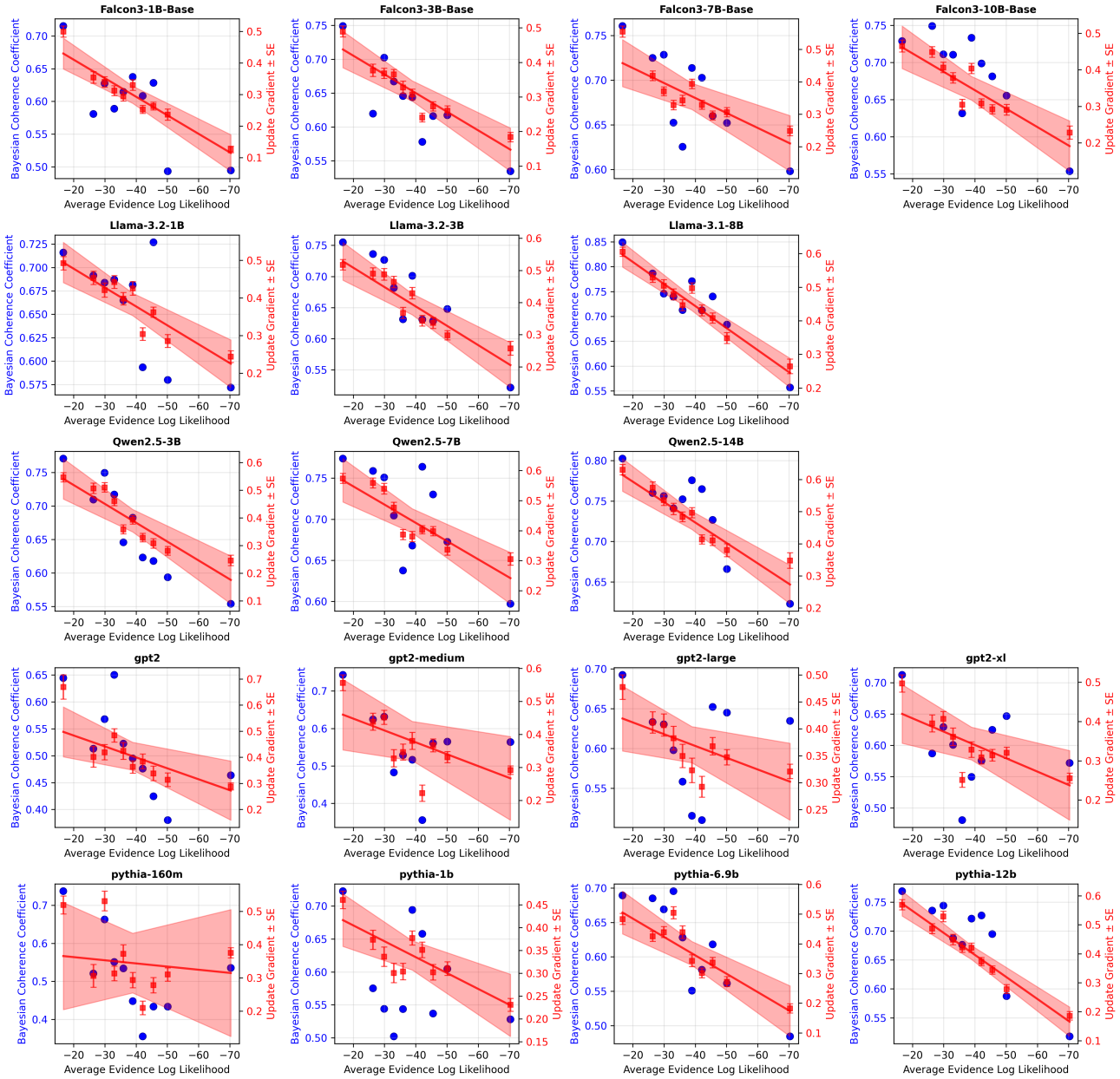

Figure 9. Scatterplots showing BCC (blue) and the gradient between the observed and expected updates (red) against the evidence log likelihood averaged over the class pair. The dataset was sorted based on the average evidence log likelihood and binned into 10 equal subsets. Each blue point represents the BCC and each red point represents the update gradient for one of 10 bins. The average evidence log likelihood is decreasing along the x-axis.

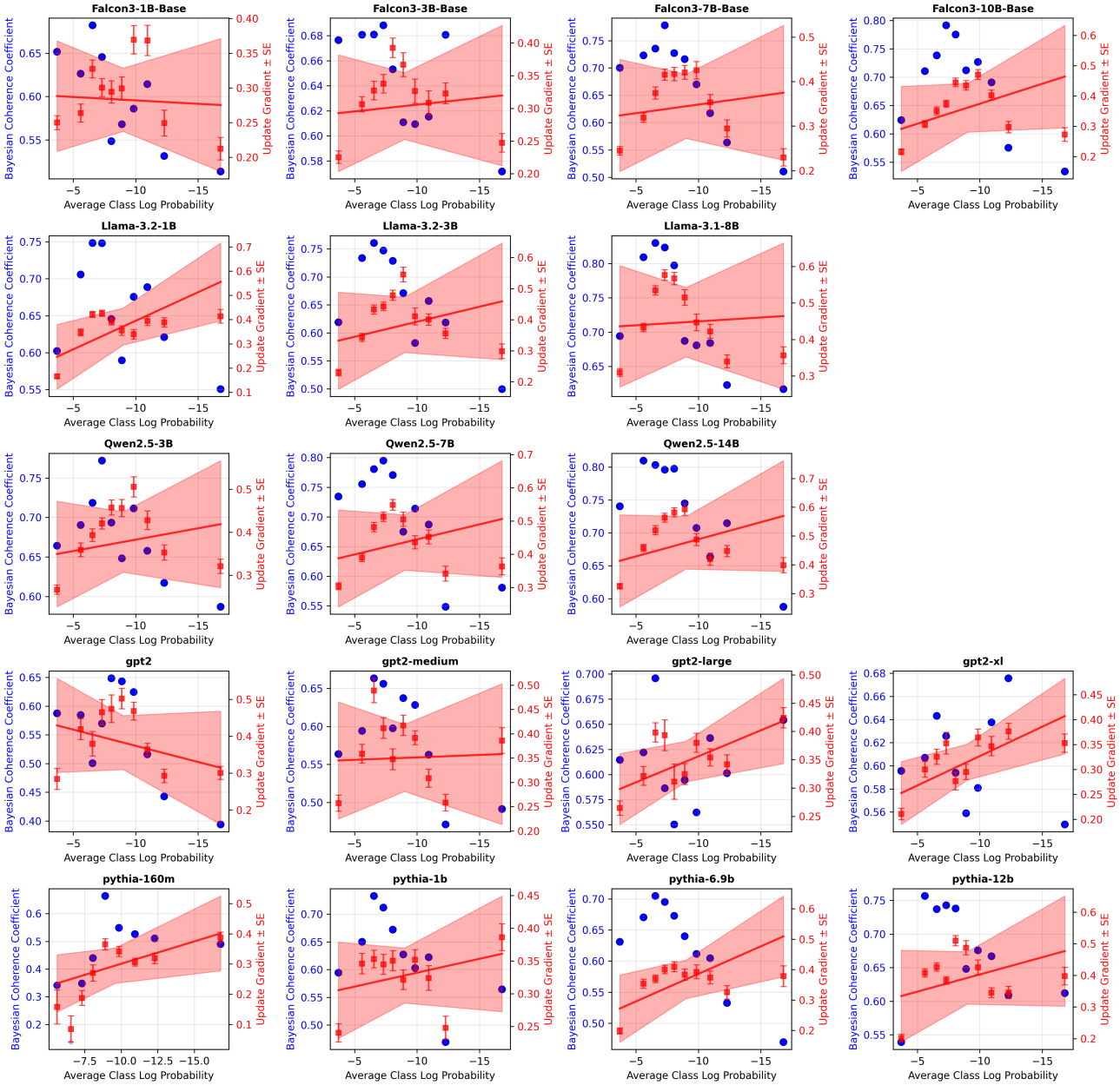

Figure 10. Scatterplots showing the BCC (blue) and the gradient between the observed and expected updates (red) against the average of the prior and posterior log probabilities averaged over the class pair. The dataset was sorted based on the average evidence log likelihood and binned into 10 equal subsets. Each blue point represents the BCC and each red point represents the update gradient for one of 10 bins. The average evidence log likelihood is decreasing along the x-axis.

C.2. Prompt Text

Prompt Text

```
Data is to be generated according to the provided JSON schema. Please follow the
schema exactly. There is also an example in JSON format provided. In the example the
\{class\_category\} is "novelists". Now based on the JSON schema and the example,
please create data for a \{class\_category\} "desired class category". There should
be at least 5 \{candidate\_classes\} in this category. Ensure that each of the
\{candidate\_classes\} has exactly the same number of tokens - this includes the
punctuation, the space at the beginning of a class and the full stop at the end. The
number of tokens should be at most 3 - use as few tokens as possible. If the
\{class\_category\} is a proper noun, then the first letter of each word of the
class should be capitalized. If the \{class\_category\} is not a proper noun then
the first letter of each word should not be capitalized. There should be at least 3
\{histories\}, varying in how related they are to the \{class\_category\} (from
completely unrelated to very related). There should be at least 20 pieces of
\{evidence\_text\}. Some pieces of the evidence text should provide high evidence
for one of the classes, other pieces of evidence text should provide evidence for
several or all of the candidate classes and some pieces of evidence text should
provide evidence for none of the candidate classes. Each \{evidence\_text\} should
be accompanied by an array \{points\_to\_classes\}, which is a list of classes in
\{class\_category\} that the evidence supports. This could be a single class, more
than one class, all classes in the \{class\_category\} or none (i.e. an empty list).
The \{evidence\_elicitation\} joined with the \{evidence\_text\} should form a
grammatically correct sentence including spaces and punctuation. The
\{class\_elicitation\} joined with each \{class\} should form a grammatically
correct sentence including spaces and punctuation. It is important to follow the
example for \{class\_elicitation\} and \{evidence\_elicitation\} including spaces
and other punctuation. "desired class category" = "school\_of\_philosophy"
```

C.3. JSON Schema

JSON Schema

```
{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "type": "object",
  "properties": {
    "bayesian_reasoning": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "conversation_history": {
            "type": "string",
            "description": "the conversation history"
          },
          "candidate_classes": {
            "type": "array",
            "items": {
              "type": "string"
            },
                        "minItems": 2,
                        "uniqueItems": true,
            "description": "list of candidate classes"
          },
          "evidence": {
            "type": "string",
            "description": "justification or rationale for the classification"
          },
          "class_elicitation": {
```

```
            "type": "string",
            "description": "prompt used to elicit a candidate class"
          },
          "evidence_elicitation": {
            "type": "string",
            "description": "prompt used to elicit the evidence"
          }
        },
        "required": [
          "conversation_history",
          "candidate_classes",
          "evidence",
          "class_elicitation",
          "evidence_elicitation"
        ]
      }
    }
  },
  "required": ["bayesian_reasoning"]
}
```

C.4. Example

Example for class type ”novelists”

```
{
        "bayesian_reasoning": [
                {
                        "class_type": "novelists",
                        "conversation_history": "We've been discussing literary
                        styles and historical contexts in literature.",
                        "candidate_classes": [" William Shakespeare.", " Oscar
                        Wilde.", " Jane Austen.", " Charles Dickens.", " Virginia
                        Woolf."],
                        "class_elicitation": " My favourite author is",
                        "evidence_elicitation": " I prefer reading",
                        "evidence": [
                                {
                                        "category": "literary_analysis",
                                        "evidence_text": " works that bring out the
                                        contemporary social conventions and mores of
                                        its time rather than focusing on poetic
                                        richness and dramatic performance."
                                },
                                {
                                        "category": "literary_analysis",
                                        "evidence_text": " character-driven
                                        narratives."
                                },
                                {
                                        "category": "historical_context",
                                        "evidence_text": " literature from periods
                                        of significant social transition that
                                        captures changing values, particularly those
                                        written during times when society was
                                        undergoing fundamental shifts in perspective
                                        about class, gender roles, and interpersonal
                                        relationships."
                                },
                                {
                                        "category": "historical_context",
```

```
                                        "evidence_text": " social observers."
                                },
                                {
                                        "category": "cultural_impact",
                                        "evidence_text": " books that challenged
                                        conventional thinking and introduced
                                        progressive social ideas."
                                },
                                {
                                        "category": "cultural_impact",
                                        "evidence_text": " enduring classics that
                                        remain relevant centuries later,
                                        particularly those that have been adapted
                                        across multiple media formats and continue
                                        to shape our understanding of narrative
                                        structure and character development in ways
                                        that transcend their original historical
                                        context."
                                },
                                {
                                        "category": "stylistic_technique",
                                        "evidence_text": " subtle irony."
                                },
                                {
                                        "category": "stylistic_technique",
                                        "evidence_text": " prose that employs wit
                                        and carefully structured dialogue to develop
                                        character."
                                }
                        ]
                }
        ]
}
```