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]

英寸

Abstract

更大且更强大的语言模型是否能够在上下文中更一致地使用贝叶斯定理来更新关于命题的“信念”？为测试这一点，我们制定了一个贝叶斯一致性系数（BCC）指标，并生成了用于测量 BCC 的数据集。我们对五个模型系列中多个仅预训练的语言模型进行了 BCC 测量，并与模型参数数量、训练数据量以及模型在常见基准上的评分进行了比较。我们的结果为我们的假设提供了证据，即更大且更强大的预训练语言模型分配的信念与贝叶斯定理更加一致。这些结果对于我们对大型语言模型的理解和管理具有重要意义。

1. 介绍

贝叶斯定理允许在观察结果对现有信念提供证据时最优地更新信度 (Lin, 2024) 。了解大型语言模型（LLMs）在内部是否执行近似这一规则的信念更新对于理解和控制它们的行为非常重要。

尽管先前的研究表明，在隐藏马尔可夫模型数据上训练的 transformers 自然地实现了一种用于下一个标记预测的约束贝叶斯推理形式 (Piotrowski et al., 2025) ，但大语言模型在抽象命题而非标记上的贝叶斯一致性并无改进 (Fluri et al., 2023) 。然而，已经显示出较大的大语言模型比较小的模型更少违反其他逻辑和概率公理 (Fluri et al., 2023; Paleka et al., 2025) 。Mazeika et al. (2025) 此外证明，较大的模型表现出更高的偏好一致性，通过更少的传递性违反来衡量。他们还发现，较大的模型在其偏好上更果断和一致，他们将这解释为偏好完备性的代理。

*Equal contribution [1] 兰卡斯特大学，兰卡斯特环境中心，兰开斯特 LA1 4YQ，英国 [2] 计算与通信学院，兰卡斯特大学，兰卡斯特 LA1 4WA，英国 [3] 德国图宾根大学，计算机科学系，Sand 14，72076 图宾根 [4] 独立研究员 [5] 荷兰乌得勒支，3584 CS 乌得勒支，海德堡大道 8 号，乌得勒支大学，科学系，自然科学研究生院 [6] JBA 信托，Broughton 公园 1 号，斯基普顿 BD23 3FD，英国 [7] 地理与环境科学，南安普敦大学，高地，南安普敦 SO17 1BJ，英国 [8] 同济大学测绘与地理信息学院，中国上海市四平路 1239 号，邮编 200092 . Correspondence to: Sohaib Imran < s.imran1@lancaster.ac.uk >.

ICML 2025 Workshop on Assessing World Models.

如果我们期望未来训练和部署的 LLM 是连贯的贝叶斯更新者，这将带来深远的影响。更加连贯的代理对具有类似世界模型的代理来说更易于理解和预测，这转化为更高的可靠性和稳健性，因此可以更好地控制 LLM。这也意味着人类行为对于更连贯的 LLM 而言应该更易于理解。后者应该使传递信息和指定复杂目标（例如，赋予人类权力）变得更容易。另一方面，当与 LLM 互动时，隐藏信息变得更困难。这使得在评估 LLM 时，难以不让它们意识到它们正在被评估 (Fan et al., 2025; Needham et al., 2025) 。

在部分可观察环境中，贝叶斯信念更新能够实现最优控制 (Åström, 1965; Sondik, 1978) 。除此之外，具有连贯偏好的主体可以很好地建模为期望效用最大化者 (EUMs) (Hammond, 1988) ，这带来了许多好处，但也伴随着严重风险。特别是，如果 EUMs 的偏好与人类偏好不一致，它们可能会优化出人类不喜欢的世界状态 (Everitt & Hutter, 2018) 。EUMs 可能还会抵制关闭或修改其目标和偏好（即，它们可能是不可纠正的），因此会参与欺骗和寻求权力 (Soares et al., 2015; Everitt & Hutter, 2018; Hubinger et al., 2021) 。受上述考虑以及有关 LLMs 是否可以推理的讨论的推动，我们研究了 LLMs 是否在上下文中以与贝叶斯规则一致的方式更新其信念，以及这种特性如何随模型大小和能力的变化而变化。我们研究的贡献包括：

- 我们引入了一种新的度量和数据集，它比较模型在多个对话上下文中对命题和证据所进行的观察到的更新与预期更新。
- 我们通过实验证明，更大且更强大的 LLMs 以更符合贝叶斯定理的方式更新它们对命题的信念。
- 我们讨论了我们的结果对人工智能安全性和对齐的影响。

2. 贝叶斯一致性系数

对于本分析，所考虑的命题是一组类别 C 。给定一些证据 x ，贝叶斯定理可以用来更新一个人的信念：

由于一些证据 x 指向的所有可能类别的集合是无限的，我们考虑比率：

$$\frac{P(c_1|x)}{P(c_2|x)} = \frac{P(x|c_1)P(c_1)}{P(x|c_2)P(c_2)} \tag{1}$$

其中 $c_1, c_2 \in C$ 是成对的类别。

我们引入贝叶斯一致性系数（BCC），其衡量在给定任何证据的情况下，期望更新与观察更新之间的相关性，分别通过对数似然比和对数赔率更新来衡量：

$$\mathrm{BCC}(\theta, \mathcal{D}) = \mathrm{Corr}\Big(\Delta_{\text{expected}}, \Delta_{\text{observed}}\Big) \tag{2}$$

其中，θ 是被评估的模型，$\mathcal{D}$ 是由多个类别 k 的类 c 、证据 x 和对话历史 h 组成的数据集。

预期和观察到的更新是：

$$\Delta_{\text{expected}} = \log \text{ likelihood ratio} = \log \frac{P_\theta(x|c_1,h,k)}{P_\theta(x|c_2,h,k)} \tag{3}$$

$$\begin{aligned} \Delta_{\text{observed}} &= \log \text{ odds update} \\ &= \log \text{ posterior ratio} - \log \text{ prior ratio} \\ &= \log \frac{P_\theta(c_1|x,h,k)}{P_\theta(c_2|x,h,k)} - \log \frac{P_\theta(c_1|h,k)}{P_\theta(c_2|h,k)} \end{aligned} \tag{4}$$

其中 P_θ 是 LLM 分配给构成类别 c 或证据 x 的标记的累计条件概率。

3. 数据与方法

为了生成数据集 $\mathcal{D}$ ，我们使用一个 JSON 模式、类别的期望标准、证据和历史、一个示例数据集以及一个类别 k ，通过 ChatGPT 接口访问的 GPT-4o 模型来提示一个大语言模型。包括期望标准的确切提示在附录 C 中给出。我们为 10 个手动策划的类别重复此过程，生成的数据集中每个类别至少包含五个类 c 、20 个证据 x 和三个对话历史 h ，以及类引出和证据引出字符串，前缀用于鼓励模型分别生成类和证据。

图 1 展示了如何使用上述组件来计算由被评估的模型 θ 分配的先验 $P_\theta(c \mid h, k)$ 、似然 $P_\theta(x \mid c, h, k)$ 和后验 $P_\theta(c \mid x, h, k)$ 。重要的是，我们使用模型的不同实例来计算这些内容。

按照上述方法，对于每个类别 k 中的每对类别 (c_1, c_2) ，我们计算先验比、似然比和后验比，具体如方程 3 和 4 所述，最终计算 BCC 作为所有预期和观察到的更新之间的相关性（方程 2 ）。

所有模型均在温度参数设置为 1 的情况下进行了评估，这是模型通常训练时使用的温度。模型参数数量和基准得分来自于 Open LLM Leaderboard 2 (Fourrier et al., 2024) ，该标准提供了跨多个语言模型族的标准化评估。排行榜报告的参数数量和基准得分可能与官方模型文档中略有不同。

4. 结果

我们根据方程 2 计算贝叶斯一致性系数（BCC），跨整个数据集 $\mathcal{D}$ ，总计 6460 个 (c_1, c_2, x, h, k) 元组，并

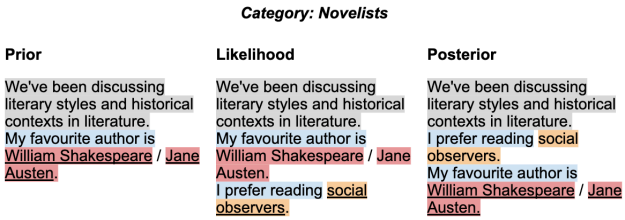

Figure 1. 我们计算所有类别（红色）、证据（橙色）、历史（灰色）和类别组合的先验、似然和后验概率，作为分配给带下划线标记的累积对数概率，条件是其在前文本的基础上。为每个类别固定的诱导文本（蓝色）被用来鼓励类别和证据标记。贝叶斯一致性系数被计算为每个类别内和给定每个证据及对话历史下的所有类别对（/两侧）的预期和观察更新之间的相关性。

针对五个模型族中的多个模型进行计算。图 2 通过展示预期和观察到的信任度更新之间的相关性，说明了 Llama 系列中的两个模型的 BCC。对于每个系列中最大的测试模型，我们发现 BCC 在各类中分布类似（图 6 ）。

所有测试的模型的 BCC 值都大于 0，这意味着与随机策略相比，模型更新更符合贝叶斯法则。对于所有测试的模型家族，BCC 都随规模增加，唯一的例外是 GPT-2 家族中的 GPT-2 XL 和 GPT-2 Large 模型（图 3 和附录图 8 ）。我们发现，在我们测试的模型中，BCC 与模型参数数量的对数之间存在显著的正相关关系（($r = 0.906, p < 10^{-6}$ ））。较大的模型还显示出更大比例的观测更新和预期更新方向一致，这可通过表格 1 中的方向一致性测量。

对于所有测试的模型，观测到的更新与预期更新的梯度都小于 1。进一步的调查显示，这个更新梯度与类对的负证据对数可能性的平均值成反比（见附录图 9 ）。较大的模型表现出更接近 1 的更新梯度（见表 1 ），唯一的例外是 GPT-2 Large 和 XL，以及 Pythia 160M 和 1B 模型对（见附录图 8 ）。完美的贝叶斯更新将意味着观测和预期的更新是相等的；因此，得到的梯度为 1。

我们进一步使用 Pythia 模型家族来评估 BCC 与训练步骤之间的关系，因为这些模型在多个训练检查点 (Biderman et al., 2023) 上是可用的。我们发现 BCC 与训练步骤数量增加之间存在不显著的正趋势（图 4 ）。

最后，我们探索了 BCC 如何随着模型在一组常用来评估模型能力的基准上的得分而变化，即 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) 和 Math Lvl 5（即 MATH 数据集的 5 级（最困难）子集）(Hendrycks et al., 2021) 。我们发现 BCC 与其中四个基准的得分存在显著的正相关关系（$p < 10^{-2}$ ），分别是 BIG-

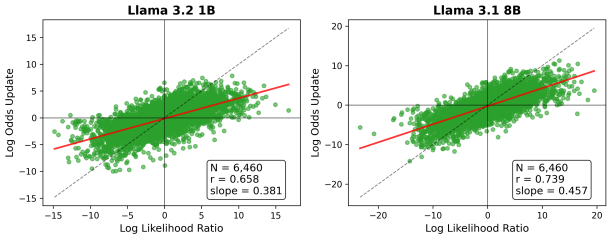

Figure 2. 散点图显示了 Llama 3.2 1B 和 3.1 8B 模型的观察更新（对数几率更新）与期望更新（对数似然比）之间的关系。每个点代表数据集中的一个（类别对，证据，历史，类别）四元组。模型的 BCC 是期望更新和观察更新之间的相关性（r 值）。p 值太小无法正确显示，因此在该图中被省略。实心红色和虚线黑色对角线分别显示了观察更新与期望更新之间的观察和理想更新梯度。

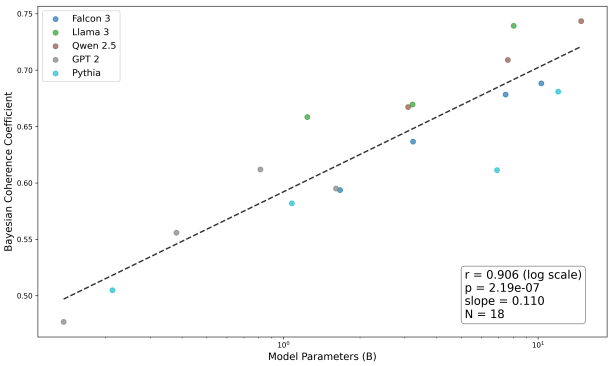

Figure 3. 贝叶斯一致性系数作为模型参数数量的函数。每个点代表在完整数据集上评估的预训练模型。x 轴表示参数的数量（单位为十亿），采用对数刻度。

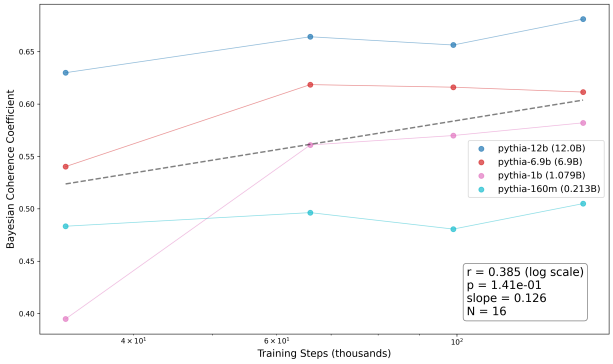

Figure 4. 在训练过程中，来自 Pythia 模型家族的四个模型（不同参数数量，见图例）的 BCC 演变。每个点代表一个在完整数据集上评估的预训练模型。x 轴表示训练步骤的数量（以千为单位）的对数刻度，每个步骤涉及一个包含 2 百万标记的批次。

Bench Hard、GPQA、MMLU-PRO 和 Math Lvl 5，而与 IFEval 和 MUSR 基准则呈现非显著的正相关关系。

5. 讨论

在 BCC 与模型参数对数数量之间存在显著正相关（$r = 0.906, p < 10^{-6}$）（图 3），以及在 BCC 与模型在常见基准上的表现之间存在显著正相关（见图 5），这些都为我们的假设提供了证据，即规模更大且能力更强的 LLM 能够更加一致地根据贝叶斯规则更新其对命题的信任度。

我们测试的模型中，BCC 与训练步数（图 4）之间的相关性，以及 BCC 与我们测试的六个基准中的两个（IFEval 和 MUSR）之间的相关性虽然为正，但在统计学上并不显著。目前尚不清楚这是否仅仅是由于测试的模型数量有限（在训练步数的情况下，测试的模型系列更加有限），还是有其他因素在起作用。同时也不清楚为什么所有测试的模型似乎都更新不足，即为什么所有模型中观察到的更新与预期更新的梯度都小于 1。这种更新梯度与类别对上的负证据对数似然均值之间的反相关性（见附录，图 9 和 10）表明这与我们分

Table 1. 对于同一模型家族中不同规模模型的 BCC、更新梯度和观察到的与预期更新之间的方向一致性。所有条目均基于 6,460 个评估实例。

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

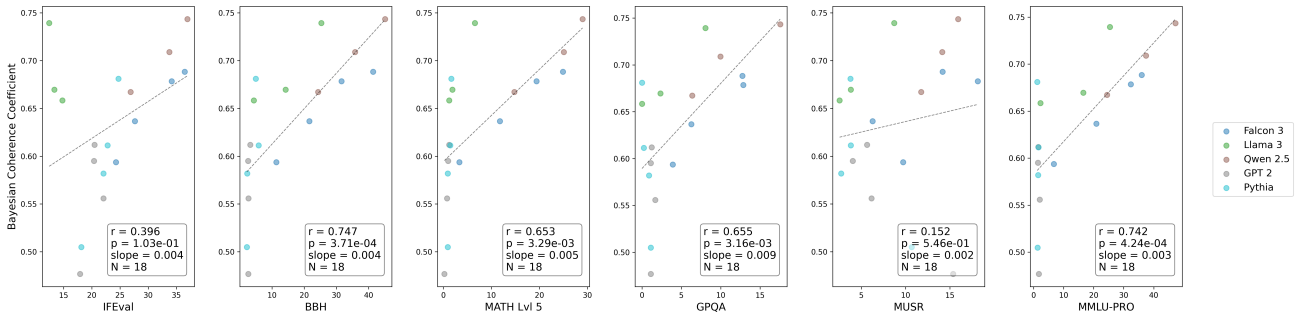

Figure 5. 针对模型在一组常用于评估模型性能的基准测试中获得的（标准化）分数进行 BGC。每个点代表一个在完整数据集上评估的预训练模型。x 轴表示标准化基准测试分数，其中 0 为随机基线，100 为可达到的最高分数。

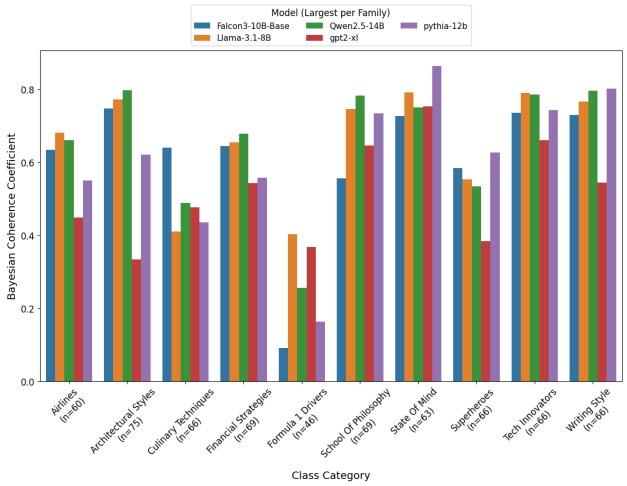

Figure 6. 我们数据集中各个类别中所选模型的 BCC。

析中使用的证据相比于类别的可能性较小相关。

我们的结果增加了研究的内容，显示了更大和更强大的大语言模型（LLM）的信念和偏好在逻辑上一致性更强 (Paleka et al., 2025; Mazeika et al., 2025)，而这与 Fluri et al. (2023) 的研究形成对比，后者发现从 GPT 3.5 到 GPT 4，一个更大且更强大的模型，其贝叶斯一致性没有增加，尽管在其他一致性度量上有所改善。他们的负向扩展结果归因于反转诅咒 (Berglund et al., 2024)。如果反转诅咒确实是他们负向扩展结果的原因，我们也会期望我们的扩展结果为负。我们假设他们的负向结果可能是由于依赖于基于错误的指标而不是类似于 BCC 这样的基于相关性的指标，这可能会给出对更“自信”预测的模型较低的评分。关于这一点的进一步讨论，见附录 A 。

我们的结果提供了早期的证据，表明更大且更强大的预训练模型以一种更符合贝叶斯法则的方式更新其信念，BCC 与规模大致呈对数线性增加。这些结果具有许多潜在的影响。更大且更强大的 LLM 更适合作为贝叶斯更新者模型，表明它们可能会学习内部一致的世界模型，这可能允许推理。就其内部世界模型与人类相似而言，这应当允许更高效的人与 AI 系统之间的信息交换。另一方面，隐藏信息变得更加困难，因为 LLM 可以从细微的线索中更准确地推断出世界状态。如果结合一致的偏好，更具贝叶斯性的代理可能更接近 EUM 行为的表现。由于具有偏好不对齐的 EUM 可能会优化有害的世界状态，并且默认情况下是不可纠正的，我们的结果呼吁在开发稳健的对齐和可纠正性方法方面进行研究。

6. 局限性和未来研究

我们的研究有一些局限性。首先，我们将对观察到的非显著相关性（BCC 与训练步骤的相关性，BCC 与两个基准的相关性）以及观察到的更新不足现象进行深入研究的工作留给未来的研究。

其次，我们的分析仅涵盖了预训练模型，并且上限为 140 亿参数。因此，在将我们的结果推断到更大和更强大的模型之前，需要谨慎。未来的研究应针对更大型的语言模型和微调的语言模型重复我们的结果。无论是指令微调还是强化学习微调的模型都是值得研究的有趣对象。

第三，我们使用累积标记概率作为它们组成的命题的信任度的代理。目前尚不清楚这是否是对行动相关信念状态的准确代理。未来的研究应该调查这一假设的有效性，并探索其他替代的代理来表示一个 LLM 对命题的信任度。此外，评估参数随机初始化未训练模型的 BCC 可以提供该指标有用性的见解。

最后，我们只评估了关于信念连贯性众多概念中的一个，并且也是用单一的度量标准进行评估。未来的研究应扩展我们的分析，以包括其他连贯性形式的研究。

7. 结论

我们假设，随着模型规模和能力的提高，LLMs 在上下文中更新其对命题的信念时，会与贝叶斯定理更一致。

我们通过设计一个新颖的指标，即贝叶斯一致性系数（BCC），来测试这一假设，该指标测量预期和观察到的证据更新之间的相关性。我们的结果显示，BCC 与模型参数数量的对数之间，以及 BCC 与常用于评估模型能力的基准测试中模型性能之间，存在显著的相关性，这为我们的假设提供了证据，即更大且更有能力的模型在更新信念时与贝叶斯定理更加一致。

8. 数据和代码可用性

本研究中评估的语言模型可从 HuggingFace Hub 下载（https://huggingface.co/），可通过在附录中查找其模型名称找到，如图 8 所示。基准分数和参数数量来自 Huggingface Open LLM Leaderboard 2（https://huggingface.co/datasets/open-llm-leaderboard/contents）。

用于评估贝叶斯一致性系数的数据集可以在 https://github.com/AISC10-team09/bayesian_reasoning/blob/main/data/data.json 找到，而用于复制我们分析的代码可以在 https://github.com/AISC10-team09/bayesian_reasoning.git 找到。

9. 致谢

我们要感谢 AI Safety Camp（https://aisafety.camp/）将作者们聚集在一起并支持这个项目。我们还要感谢 Guillaume Corlouer 和（匿名的）同行评审者，他们的反馈显著提高了这项研究。SI 还特别感谢兰卡斯特大学资助他与 Peter M. Atkinson 教授一起的研究员职位。

10. 作者贡献

在 PA 和 RL 的监督下，SI 提出了原始项目。SI、IK 和 AT 实现了项目的代码库。AT 和 RC 生成了数据集。SI 和 MB 分析了结果。SI、IK、MB、AT 和 RC 起草了手稿。SI 处理了评审者的评论并准备了最终的手稿。PA 和 RL 检查了方法，审阅并编辑了手稿。

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-2-660cdb7601eba6852431fffc.

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].

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.

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. 替代度量

A.1. 贝叶斯一致性误差

除了主要论文中介绍的贝叶斯一致性系数（BCC）外，我们还探讨了一种基于误差的替代度量方法来量化贝叶斯一致性。贝叶斯一致性误差（BCE）直接测量期望的和观察到的信念更新之间的偏差（公式 5 ）。由于 BCE 是一个误差度量，较小的值应表明更高的贝叶斯一致性。

$$\mathrm{BCE}(\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(c_1, c_2, x, h, k) \in \mathcal{D}} \Big(\Delta_{\text{expected}} - \Delta_{\text{observed}} \Big)^2 \tag{5}$$

其中，Δ_{expected} 和 Δ_{observed} 是主论文中方程 3 和 4 中定义的期望和观察到的对数几率更新。

A.2. BCE 的理论限制

我们的分析揭示了 BCE 作为连贯性度量的一个基本限制。我们发现 BCE 对高熵分布表现出一种不良偏向。为说明这一点，考虑一个病态的情况：一个大型语言模型始终在其词汇表 V 上输出均匀分布。对于这样的模型：

$$\Delta_{\text{expected}} = \log \text{ likelihood ratio} = \log \frac{P_\theta(x|c_1, h, k)}{P_\theta(x|c_2, h, k)} = \log \frac{\frac{1}{|V|}}{\frac{1}{|V|}} = 0 \tag{6}$$

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

这导致了 $\mathrm{BCE} = 0$ ，表明尽管模型未提供有意义的信息，却表现出完全的连贯性。因此，可以通过增加输出熵轻而易举地使 BCE 最小化。

A.3. 实证分析

为了实证验证上述限制，我们分析了在不同温度设置下 BCE 和 BCC 指标的行为（图 7 ）。温度缩放是一种常见技术，它在不改变基础模型参数的情况下修改模型输出的熵。

图 7 证实，像 BCC 这样的基于相关性的指标是更稳健的贝叶斯一致性衡量标准，因为它们不会受到输出分布熵的干扰。我们在整篇论文中使用 BCC 作为我们的主要指标。

B. 进一步的结果

为了进一步研究影响贝叶斯一致性和观察到的更新不足现象的因素（见图 8 ），我们考察了在给定平均证据对数似然和平均类别对数概率的情况下，BCC 和更新梯度如何变化：

$$\text{average evidence log likelihood} = \frac{1}{2}\big[\log P_\theta(x|c_1, h, k) + \log P_\theta(x|c_2, h, k)\big] \tag{8}$$

$$\text{average class log probability} = \frac{1}{4}\Big(\big[\log P_\theta(c_1|h, k) + \log P_\theta(c_2|h, k)\big] + \big[\log P_\theta(c_1|x, h, k) + \log P_\theta(c_2|x, h, k)\big]\Big) \tag{9}$$

C. 数据集生成

提示文本（C.1 ）、JSON 模式（C.2 ）和一个示例（C.3 ）共同组成给 LLM 的提示，用于为特定的类别生成数据。

对于每个类别，我们提供以下期望作为对 LLM 的指示：

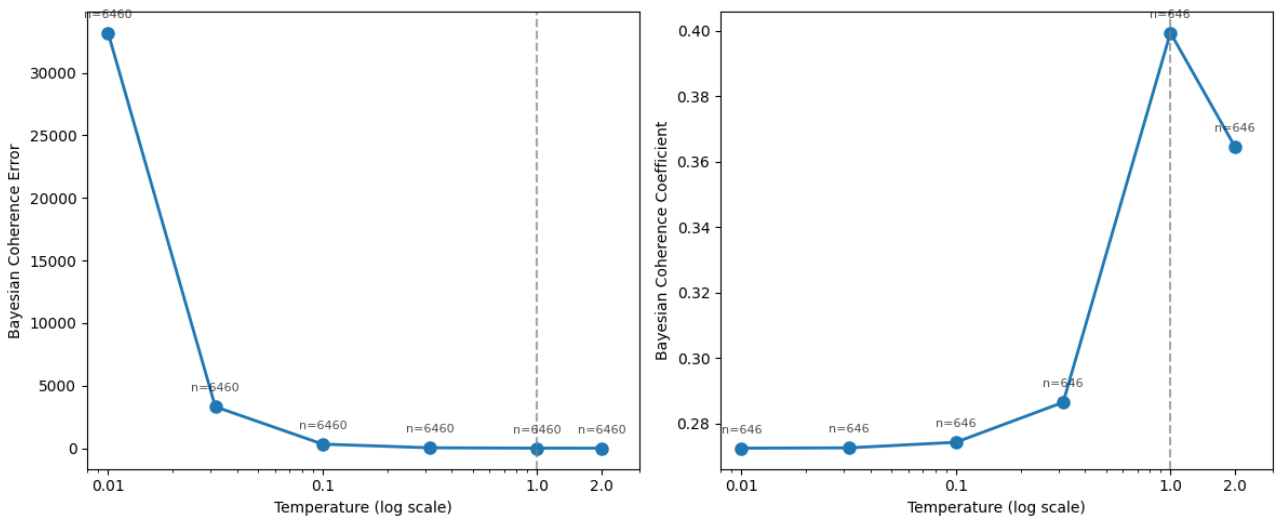

Figure 7. 温度不变性分析展示了 BCE 和 BCC 度量在不同温度设置下的表现。温度校准在不改变模型参数的情况下修改了模型输出的熵。与 BCE 相比，BCC 在不同的温度值下表现出更强的鲁棒性。

- 每个类别中至少有五个类。
- 所有类别名具有相同的词元计数。
- 至少有三种对话历史，类别与对话历史的相关性有所不同，从非常相关到无关。
- 至少有 20 条证据文本，这些证据可以倾向于支持一个特定的类别、多个类别或没有倾向性。

C.1. **提示文本**

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”

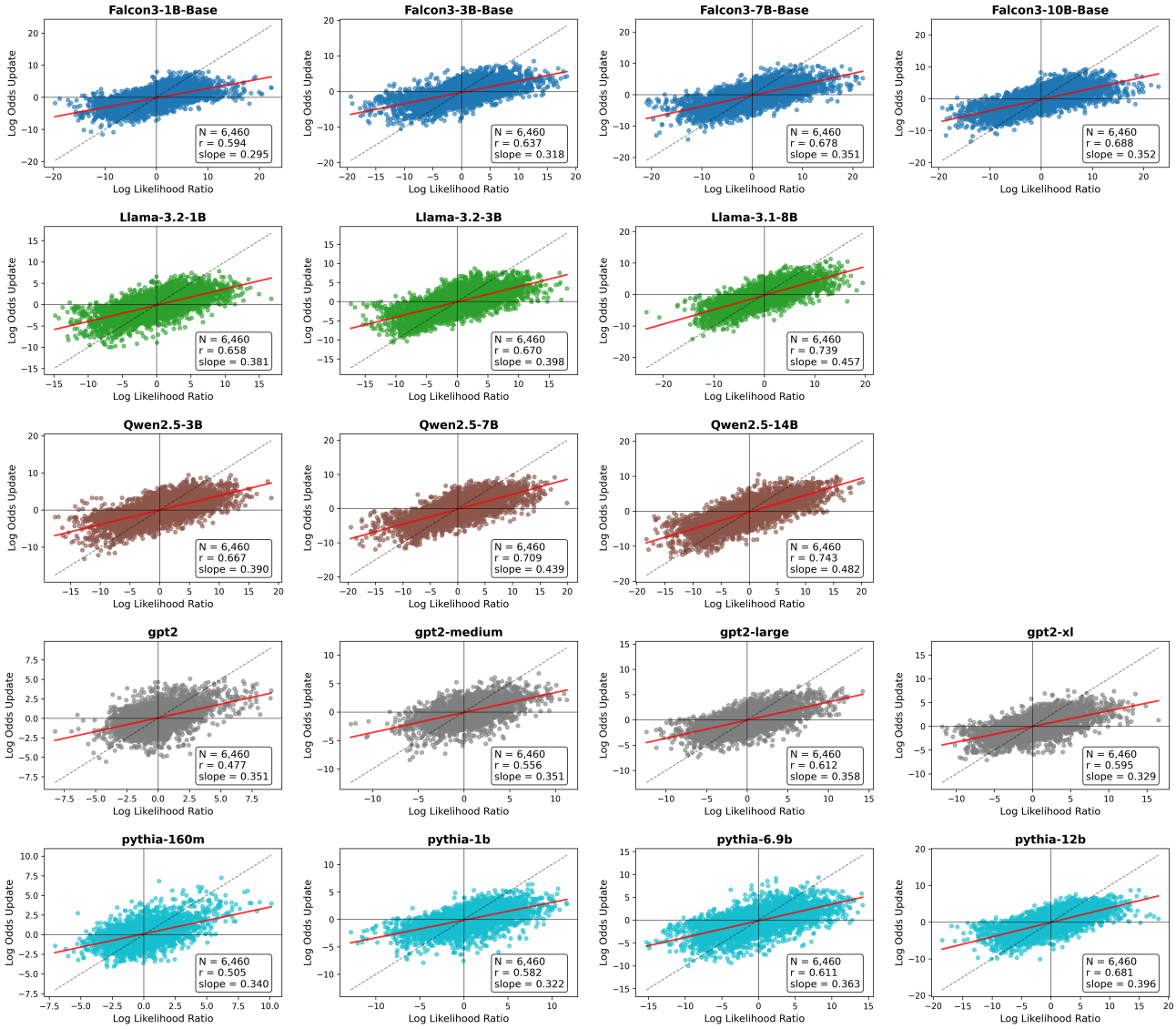

Figure 8. 散点图展示了观察到的更新（对数优势更新）与预期更新（对数似然比）。每个点代表了数据集中的一个（类别对，证据，历史，类别）元组。模型的 BCC 是预期更新与观察到的更新之间的相关性（r 值）。实线红色和虚线黑色对角线分别显示观测更新与预期更新之间的拟合和理想更新梯度。

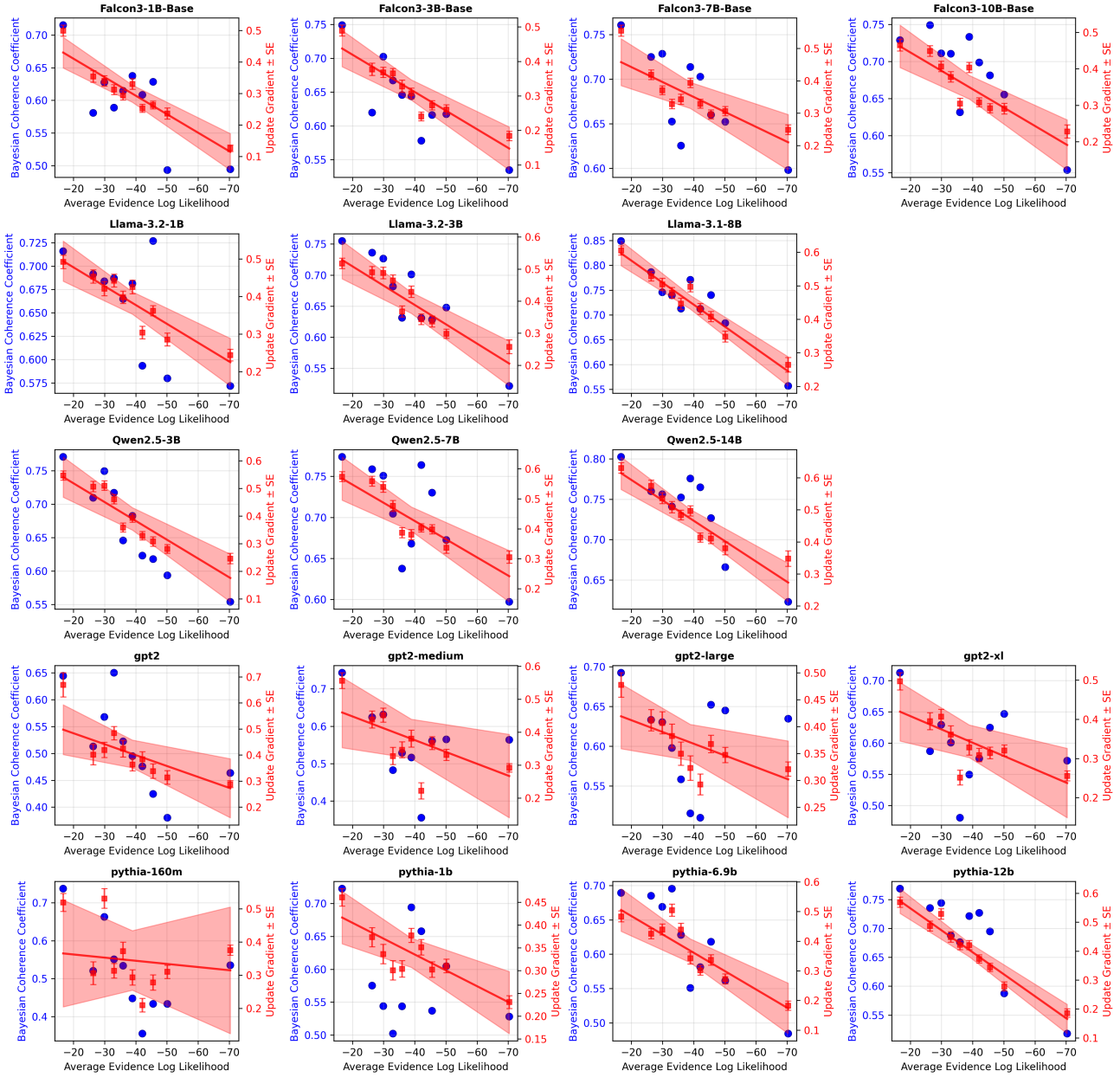

Figure 9. 散点图显示了 BCC（蓝色）和观察到的更新与期望更新之间的梯度（红色）与证据对数似然的关系，这些对数似然值是基于类别对的平均值。数据集根据平均证据对数似然值排序并分为 10 个相等的子集。每个蓝点代表一个 BCC，每个红点代表 10 个子集之一的更新梯度。平均证据对数似然值沿 x 轴递减。

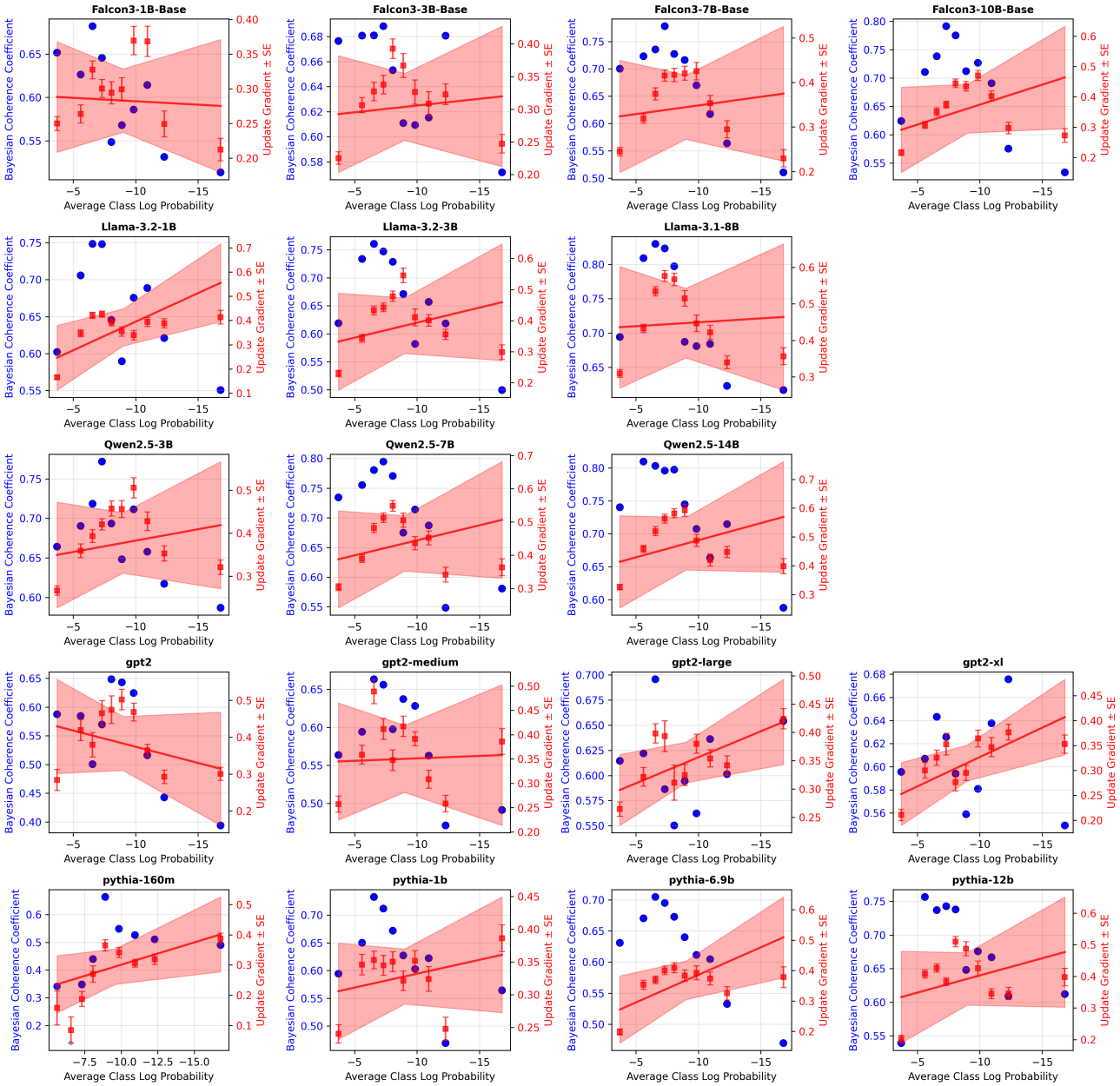

Figure 10. 散点图显示 BCC（蓝色）和观察到的更新与预期更新之间的梯度（红色），相对于类别对的先验和后验对数概率的平均值。数据集根据平均证据对数似然性排序，并分成 10 个相等的子集。每个蓝点代表一个子集的 BCC，每个红点代表 10 个子集之一的更新梯度。平均证据对数似然性沿 x 轴递减。

C.2. JSON 模式

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"
            },
^^I^^I^^I"minItems": 2,
^^I^^I^^I"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.3. 例子

JSON Schema

```
{
^^I"bayesian_reasoning": [
^^I^^I{
^^I^^I^^I"class_type": "novelists",
^^I^^I^^I"conversation_history": "We've been discussing literary styles and historical contexts in literature.",
^^I^^I^^I"candidate_classes": [" William Shakespeare.", " Oscar Wilde.", " Jane Austen.", " Charles Dickens.", "
    Virginia Woolf."],
```

```
^^I^^I^^I"class_elicitation": " My favourite author is",
^^I^^I^^I"evidence_elicitation": " I prefer reading",
^^I^^I^^I"evidence": [
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "literary_analysis",
^^I^^I^^I^^I^^I"evidence_text": " works that bring out the contemporary social conventions and mores of its time
    rather than focusing on poetic richness and dramatic performance."
^^I^^I^^I^^I},
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "literary_analysis",
^^I^^I^^I^^I^^I"evidence_text": " character-driven narratives."
^^I^^I^^I^^I},
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "historical_context",
^^I^^I^^I^^I^^I"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."
^^I^^I^^I^^I},
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "historical_context",
^^I^^I^^I^^I^^I"evidence_text": " social observers."
^^I^^I^^I^^I},
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "cultural_impact",
^^I^^I^^I^^I^^I"evidence_text": " books that challenged conventional thinking and introduced progressive social
    ideas."
^^I^^I^^I^^I},
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "cultural_impact",
^^I^^I^^I^^I^^I"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."
^^I^^I^^I^^I},
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "stylistic_technique",
^^I^^I^^I^^I^^I"evidence_text": " subtle irony."
^^I^^I^^I^^I},
^^I^^I^^I^^I{
^^I^^I^^I^^I^^I"category": "stylistic_technique",
^^I^^I^^I^^I^^I"evidence_text": " prose that employs wit and carefully structured dialogue to develop character."
^^I^^I^^I^^I}
^^I^^I^^I]
^^I^^I}
  ]
}
```