ARE LARGE LANGUAGE MODELS BAYESIAN? A MARTINGALE PERSPECTIVE ON IN-CONTEXT LEARNING

Fabian Falck *, **Ziyu Wang** *, **Chris Holmes**
Department of Statistics, University of Oxford
{fabian.falck, ziyu.wang, cholmes}@stats.ox.ac.uk

ABSTRACT

In-context learning (ICL) has emerged as a particularly remarkable characteristic of Large Language Models (LLM). Numerous works have postulated ICL as approximately Bayesian inference, rendering this a natural hypothesis. In this work, we analyse this hypothesis from a new angle through the *martingale property*, a fundamental requirement of a Bayesian learning system on exchangeable data. We show that the martingale property is a necessary condition for unambiguous predictions in such scenarios, and enables a principled, decomposed notion of uncertainty vital in trustworthy, safety-critical systems. We derive actionable checks with corresponding theory and test statistics which must hold if the martingale property is satisfied. We also examine if uncertainty in LLMs decreases as expected in Bayesian learning when more data is observed. In three experiments, we provide evidence for violations of the martingale property, and deviations from a Bayesian scaling behaviour of uncertainty, falsifying the hypothesis that ICL is Bayesian.

1 INTRODUCTION

A particularly remarkable characteristic of Large Language Models (LLMs) is so-called *in-context learning* (ICL) (Brown et al., 2020; Dong et al., 2022): Given a pretrained LLM p_M and an observed dataset $D := \{(x_1, y_1), \ldots, (x_n, y_n)\} = z_{1:n}$, LLMs capture the distribution of the underlying random variables X and Y. This allows them produce a new sample (x_{n+1}, y_{n+1}) using the *predictive distribution* $p_M(X_{n+1}, Y_{n+1} | Z_{1:n} = z_{1:n})$, or if x_{n+1} is observed infer the predictive distribution $p_M(Y_{n+1} | X_{n+1} = x_{n+1}, Z_{1:n} = z_{1:n})$, without retraining or fine-tuning p_M.

In spite of the remarkable empirical success of ICL, we lack a unified understanding of the algorithm and the properties of conditioning LLMs on in-context data. In this work, we are interested in characterising the type of learning that occurs in ICL. Specifically, we aim to answer the question: **is in-context learning for LLMs (approximately) Bayesian?** – In contrast to prior work, our analysis focuses on one fundamental property of Bayesian learning: the *martingale property*. In a nutshell, the martingale property describes the invariance of a model's predictive distribution with respect to missing data from a population. We will formally define and extensively explain the martingale property in §2. We refer to App. B.1 for a motivating example which intuitively describes two important and desirable *consequences* of the martingale property, which we formally discuss in §2.

This work states the hypothesis that ICL in LLMs is Bayesian. Numerous works have argued that ICL approximates a form of Bayesian inference (Xie et al., 2021; Hahn & Goyal, 2023; Jiang, 2023) which we will carefully review in App. D, rendering this hypothesis natural. Our work introduces a novel perspective which contradicts their conclusion: we show that the martingale property, a fundamental property of Bayesian learning systems, is violated for state-of-the-art LLMs such as Llama2, Mistral and GPT 3.5. We on purpose focus our analysis on three synthetic experiments where the ground-truth data generating process is simple and known, and which provide a useful test bed without the convolution of unknown latent effects as is typical in natural language.

More specifically, our *contributions* are: (a) We motivate the martingale property as a fundamental property of Bayesian learning, crucial for unambiguous predictions of an LLM in exchangeable settings, and interpretation of uncertainty (§2). (b) We derive actionable diagnostics with corresponding

*Equal contribution.

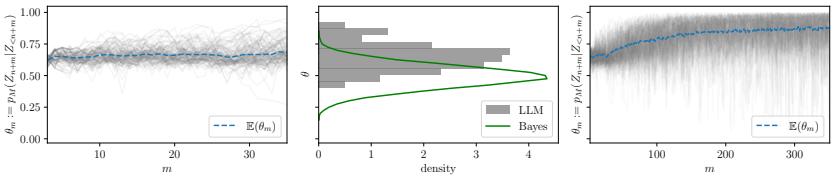


Figure 1: *In-context learning in Large Language Models is not Bayesian.* [Left] The *martingale property*, a necessary condition of Bayesian learning systems, is satisfied for short sample paths. [Centre] This allows us to approximate the *martingale posterior* (see §2) which, however, indicates deviation from a reference Bayesian model. [Right] For longer sample paths, we observe a drift which violates the martingale property, together rendering the ICL system non-Bayesian.

theory and test statistics of the martingale property. (c) We provide novel evidence for violations of the martingale property through LLMs in certain settings, and a deviation of the sample efficiency of ICL relative to Bayesian systems, falsifying our hypothesis that ICL in LLMs is Bayesian and cautioning against the use of LLMs in exchangeable and safety-critical applications (§4).

2 What characterises Bayesian ICL? A martingale perspective

The martingale property. In this section we rigorously formalise properties of an ICL system that follows Bayesian principles. Theoretical details and technical proofs are presented in App. A.

Definition 1. The predictive distributions for $\{Z_i\}$ satisfy the *martingale property* if for all integers $n, k > 0$ and realisations $\{z, z_{1:n}\}$ we have

$$p_M(Z_{n+1}{=}z|Z_{1:n}{=}z_{1:n}){=}p_M(Z_{n+k}{=}z|Z_{1:n}{=}z_{1:n}). \tag{1}$$

As we will explain below, this renders distributions $\{p_M(Z_{n+1} = \cdot|Z_{1:n})\}$ to form a martingale, hence the name 'martingale property'. The following identity follows from Eq. (1):

$$\begin{aligned} &p_M(Y_{n+1} = y|X_{n+1} = x, Z_{1:n} = z_{1:n}) = p_M(Y_{n+k} = y|X_{n+k} = x, Z_{1:n} = z_{1:n}) \\ &= \mathbb{E}_{Z_{n+1:n+k-1}\sim p_M(\cdot|Z_{1:n}=z_{1:n})} p_M(Y_{n+k} = y|X_{n+k} = x, Z_{1:n+k-1}), \end{aligned} \tag{2}$$

which holds for all $n, k > 0$, $\{z_{1:n}, y\}$, and (almost every) $x \sim p_M(X_{n+1}|Z_{1:n} = z_{1:n})$. Eq. (2) states that the model's predictions are invariant to imputations (on average).

The martingale property is necessary for unambiguous predictions. To understand the intuition behind the martingale property, consider two scenarios for ICL, illustrated in Fig. 3 (App B.1). In both scenarios, the LLM is given the observed data (D, x_{n+1}). In scenario 1, the LLM directly infers the *predictive distribution* $p_M(Y_{n+1}|Z_{1:n} = z_{1:n}, X_{n+1} = x_{n+1})$. In scenario 2, before making a prediction, the LLM generates (imputes) $m - 1$ missing samples $\hat{z}_{n+2:n+m}$ from the population autoregressively; given the observed data and the imputed samples as a prompt, we then sample from the LLM's predictive distribution $p_M(Y_{n+1}|Z_{1:n} = z_{1:n}, X_{n+1} = x_{n+1}, Z_{n+2:n+m} = \hat{z}_{n+2:n+m})$. We repeat this imputation procedure and average the obtained predictive distributions to receive a Monte Carlo estimate of the right-hand side of Eq. (2). Scenario 2 is of practical interest when estimating aggregated statistics of a population as illustrated in our RCT example in App. B.1. – The martingale property then states that the predictive distribution from scenario 1, $p_M(Y_{n+1}|Z_n = z_n, X_{n+1} = x_{n+1})$, and that from scenario 2, $p_M(Y_{n+1}|Z_n = z_n, x_{n+1}, Z_{n+2:n+m} = \hat{z}_{n+2:n+m})$, when averaged over all possible imputations of $\hat{z}_{n+2:n+m}$ are equivalent.

Why is the martingale property natural for any probabilistic system, and LLMs in particular? It is important to observe that to the model, all information about the data distribution (in addition to its prior belief; Zellner, 1988) lies in the observed data (D, x_{n+1}). Imputing the samples $\hat{z}_{n+2:n+m}$ should hence not change the predictive distribution for y_{n+1} when averaged over all possible imputations. This is precisely the idea of Eq. (2). If the predictive distribution for y_{n+1} changes on average, the model is 'creating new knowledge' when there is none: it is 'hallucinating'. We call this phenomenon *introspective hallucinations*: by querying itself, the model changes its predictions (on average), which as we shall see in below violates how Bayesian systems learn. App. B.2 discusses another important way in which predictions are rendered unambiguous *under exchangeable data*.

The martingale property enables a principled notion of uncertainty. An appealing aspect of Bayesian modelling is that it enables a principled decomposition of the predictive uncertainty

(Kendall & Gal, 2017): the predictive distribution always has the representation

$$p_M(Z_{n+1} = \cdot \mid Z_{1:n}) = \int p(\theta|Z_{1:n})p(Z = \cdot|\theta)d\theta, \tag{3}$$

where θ denotes the (latent) model parameter, $p(\theta|Z_{1:n})$ denotes the Bayesian posterior and $p(Z = \cdot|\theta)$ denotes the likelihood. Eq. (3) shows that the variation or *uncertainty* in the predictive distribution has two sources: **epistemic uncertainty**, which is about the latent θ and can be reduced if more data is available; and **aleatoric uncertainty**, which is irreducible given a fixed set of features even if infinite samples are observed. It is then possible to diagnose the model e.g. by checking if the epistemic uncertainty is decreasing as we receive more observations.

As we review in App. B.3, under mild regularity conditions, the martingale property (1) is sufficient to guarantee that the predictive distribution $p_M(Z_{n+1} = \cdot|Z_{1:n})$ *has the same representation as Eq. (3)* (Fong et al., 2021). Moreover, it is possible to recover the equivalent to the Bayesian posterior $p(\theta|Z_{1:n})$—the *martingale posterior*—using only path samples $Z_{n+1,\ldots}|Z_{1:n}$ from the model, thereby implementing the decomposition. This allows us to interpret the predictive uncertainty from any model that satisfies the martingale property through the same foundation as Bayesian modelling. Importantly, *this methodology covers black-box models such as LLMs*.

On the link between the martingale property and Bayesian learning systems. So far, we asserted that the martingale property is fundamental to a Bayesian ICL system. We will now show that for ICL on i.i.d. data, exchangeability (see App. B.2 for a definition), for which the martingale property is a necessary condition, and Bayesian inference are closely connected, equivalent conditions.

ICL typically involves i.i.d. observations $Z_{1:n}$, which is our primary focus in this work. Therefore, a correctly specified Bayesian model should lead to $p_M(Z_{1:n} = z_{1:n}) = \int \pi(d\theta) \prod_{i=1}^{n} p_M(Z = z_i|\theta)\ \forall n \in \mathbb{N}$, where θ denotes the model parameter, π denotes the prior and $p_M(Z = \cdot|\theta)$ denotes the likelihood. It then follows that $\{Z_i\}$ are *exchangeable* (see App. B.2). The converse is also true by de Finetti's representation theorem (De Finetti, 1929): Under mild regularity conditions any p_M that defines exchangeable $\{Z_i\}$ must have the aforementioned representation. Hence, the distribution $p_M(Z_{n+1}|Z_{1:n})$ must have the form of Eq. (3), and can thus be viewed as implicit Bayesian inference over θ (Huszár, 2022). In aggregate, *ICL on i.i.d. data is Bayesian if and only if it defines an exchangeable sample sequence*. Since the martingale property is a necessary condition for exchangeability, an ICL system not satisfying the martingale property is not Bayesian.

3 Probing Bayesian learning systems through martingales

Diagnostics for the martingale property. As we showed in §2, the martingale property is fundamental to a Bayesian learning system. In this work, we probe the martingale property in LLMs via two properties *implied by* it. If these implied properties are strongly violated, so is the martingale property. More specifically, we will derive implications involving conditional expectations of the form $\mathbb{E}(f(Z_{n+1:n+m})|Z_{1:n})$, which can be estimated by generating sample paths with an LLM and use these samples to form Monte Carlo estimates of the conditional expectations. We begin with an equivalent characterisation of the (conditional) martingale property.

Proposition 1. *Any $\{Z_{n+1:n+m}\} \sim p_M(\cdot|Z_{1:n})$ satisfies Eq. (1) if and only if the following holds:*

$$\textit{for all } n', k \in \mathbb{N} \textit{ and functions } g, h, \quad \mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))h(Z_{n+1:n'})|Z_{1:n}) = 0. \tag{4}$$

We now state two implications of Proposition 1, our two diagnostics of the martingale property (1).

Corollary 1. *Let $\{Z_i : i \in \mathbb{N}\}$ be a sequence of random variables satisfying the martingale property. Then for all integers $n, n', k > 0$ and $n' > n$ it holds that:* ***(i)*** *$\mathbb{E}(g(Z_{n+1})|Z_{1:n}) = \mathbb{E}(g(Z_{n+k})|Z_{1:n})$ for all integrable functions g, and* ***(ii)*** *$\mathbb{E}((Z_{n'+k+1} - Z_{n'+1})Z_{n'}^{\top}|Z_{1:n}) = 0$.*

Properties (i) and (ii) are derived from Proposition 1 by making different choices of the functions (g, h). We refer to App. B.5 for a detailed discussion on how we derive these properties, how we instantiate them in our experiments in §4, and an illustrative example highlighting how certain choices of (g, h) can ensure consistency expression for important aspects of the posterior.

In App. C we present aggregated statistics $T_{1,g}$ and $T_{2,k}$ to compute and empirically measure properties (i) and (ii) from sample paths generated by an LLM. In our experiments, we check if these statistics lie within bootstrapped confidence intervals obtained by a reference Bayesian predictive

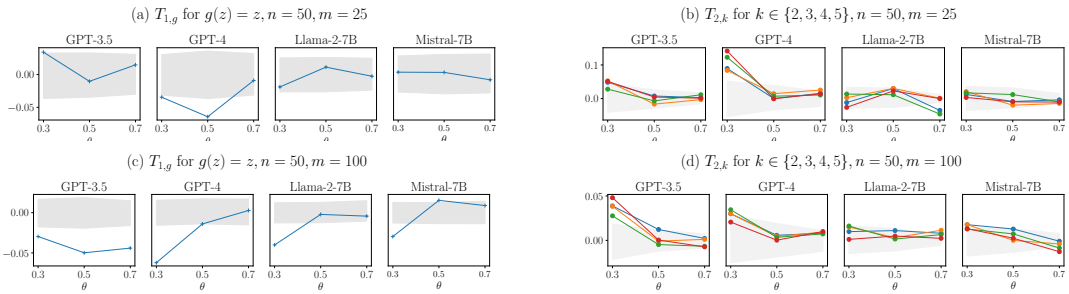


Figure 2: *Checking the martingale property on Bernoulli data.* Each data point represents a test statistic evaluated for an LLM (§3). Shade indicates 95% CIs from a reference Bayesian model.

model, which is readily available in synthetic settings, through the same sampling procedure. If $T_{1,g}$ and $T_{2,k}$ lie outside the confidence interval, properties (i) and (ii) and hence the martingale property are violated. We refer to App. B.7 for a discussion to which degree these violations are expected or even acceptable, which we considered in the interpretation of our experimental results. Lastly, we refer to App. B.6 where we derive a third diagnostic based on the scaling of epistemic uncertainty.

4 Experimental analysis on LLMs

In this section, we experimentally probe whether ICL in state-of-the-art LLMs is Bayesian using the diagnostics discussed in §3 and corresponding test statistics $T_{1,g}, T_{2,k}, T_3$.

Experiment setup. We consider three types of synthetic datasets $z_{1:n}$: **Bernoulli**: $Z_i \sim \text{Bern}(\theta)$, where $\theta \in \{0.3, 0.5, 0.7\}$; **Gaussian**: $Z_i \sim \mathcal{N}(\theta, 1)$, where $\theta \in \{-1, 0, 1\}$; A synthetic **natural language** experiment representing a prototypical clinical diagnostic task, where $Z_i = (X_i, Y_i)$ indicate the presence or absence of a symptom and disease as a text string for the i-th patient, respectively. Further, $X_i \sim \text{Bern}(0.5), Y_i | X_i \sim \text{Bern}(0.3 + 0.4X_i)$. On purpose, we reduce our experimental setup to these minimum viable test beds where the ground-truth latent parameters are known. We use the following four LLMs: `llama-2-7B` with 7B parameters (Touvron et al., 2023), `mistral-7B` (Jiang et al., 2023), `gpt-3.5` (Brown et al., 2020) and `gpt-4` (OpenAI, 2023). We refer to App. C for additional experimental results, in particular the Gaussian and natural language experiment, and checking epistemic uncertainty of LLMs with our third diagnostic, as well as experimental details.

Checking the martingale property: Bernoulli data We first check if the LLMs satisfy the martingale property. As we discussed in §2, this is a necessary condition for a Bayesian ICL system. Fig. 2 reports the results of the Bernoulli experiments with $n = 50$ observed samples, LLM sample paths of length $m \in \{n/2, 2n\}$, and datasets with ground-truth mean $\theta \in \{.3, .5, .7\}$. As discussed in §3, we compute the test statistics $T_{1,g}$ and $T_{2,k}$ on J sample paths generated by an LLM, and compare them with bootstrap confidence intervals (CIs) obtained from a reference Bayesian model. A deviation of the test statistics from the CIs indicates that the LLM is highly unlikely to be a good approximation of the reference Bayesian model. Moreover, when n becomes moderately large so that asymptotic normality results apply, such deviations further imply the deviation from all "*reasonable Bayesian models*" in the Bernstein von-Mises sense (Van der Vaart, 2000).

For short sample paths of length $m = n/2$ (subplots (a) and (b)), the LLMs' test statistics generally lie within the CIs, with the main exception being `gpt-4` ($\theta \in \{0.3, 0.5\}$), indicating a mostly adherence to the martingale property. However, for longer sample paths with $m = 2n$ (subplots (c) and (d)), all models fail the first check (left), and most models except `mistral-7b` fail the second check (right). In App. C we present further results for different choices of n ($\{20, 50, 100, 200\}$), which are consistent with Fig. 2. In summary, in the Bernoulli experiments the LLMs generally adhere to the martingale property in short sampling horizons, but in longer horizons demonstrate a significant deviation from the martingale property, and hence the Bayesian principle.

5 CONCLUSION

This work falsifies the hypothesis that in-context learning in large language models is Bayesian. Our work has several *limitations*: 1) We considered three LLMs of different computational scales, with two being relatively small (7 B parameters). 2) We analysed a limited number of diagnostics of the martingale property. While this work analysed the intrinsic behaviour of LLMs, future work should consider the use of tools which can greatly enhance the performance of such joint systems, or using fine-tuning to penalise deviations from the martingale property.

ACKNOWLEDGMENTS

Fabian Falck acknowledges the receipt of studentship awards from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme (Grant Ref: 218529/Z/19/Z). Ziyu Wang acknowledges support from Novo Nordisk. Chris Holmes acknowledges support from the Medical Research Council Programme Leaders award MC_UP_A390_1107, The Alan Turing Institute, Health Data Research, U.K., and the U.K. Engineering and Physical Sciences Research Council through the Bayes4Health programme grant.

This research is supported by research compute from the Baskerville Tier 2 HPC service. Baskerville is funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham. We further acknowledge the receipt of OpenAI API credits through the OpenAI Researcher Access Program.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Patrizia Berti, Luca Pratelli, and Pietro Rigo. Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, 32(3), July 2004. ISSN 0091-1798. doi: 10.1214/009117904000000676.

Anirban Bhattacharya, Debdeep Pati, and YUN Yang. Bayesian fractional posteriors. *Annals of Statistics*, 47(1):39–66, 2019.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Bruno De Finetti. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pp. 179–190, 1929.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Joseph L Doob. Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pp. 23–27, 1949.

Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *arXiv preprint arXiv:2103.15671*, 2021.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Thomas L Griffiths and Joshua B Tenenbaum. Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773, 2006.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820*, 2023.

Michael Hahn and Navin Goyal. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*, 2023.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2023.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Ferenc Huszár. Implicit bayesian inference in large language models. https://www.inference.vc/implicit-bayesian-inference-in-sequence-models/, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Hui Jiang. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*, 2023.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*, 2023.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Yinan Mei, Shaoxu Song, Chenguang Fang, Haifeng Yang, Jingyun Fang, and Jiang Long. Capturing semantics for imputation with pre-trained language models. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 61–72. IEEE, 2021.

OpenAI. Gpt-4 technical report, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *arXiv preprint arXiv:2306.15063*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. Model dementia: Generated data makes models forget. *arXiv e-prints*, pp. arXiv–2305, 2023.

Aaditya K Singh, Stephanie CY Chan, Ted Moskovitz, Erin Grant, Andrew M Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *arXiv preprint arXiv:2311.08360*, 2023.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

tqdm contributors. Imageio. https://github.com/tqdm/tqdm, 2022.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*, 2023.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman (eds.), *Proceedings of the 9th Python in Science Conference*, pp. 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*, 2022.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42 (4):278–280, 1988.

Liyi Zhang, R Thomas McCoy, Theodore R Sumers, Jian-Qiao Zhu, and Thomas L Griffiths. Deep de finetti: Recovering topic distributions from large language models. *arXiv preprint arXiv:2312.14226*, 2023a.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023b.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Appendix for *Are Large Language Models Bayesian? A Martingale Perspective on In-Context Learning*

A Proofs of theoretical statements in the main text

Fact 1. Any exchangeable random sequence $\{Z_i\}$ must be conditionally identically distributed.

Proof. See, e.g., Berti et al. (2004, p. 2030). □

Proposition 1. *Any* $\{Z_{n+1:n+m}\} \sim p_M(\cdot|Z_{1:n})$ *satisfies Eq. (1) if and only if the following holds:*

$$\text{for all } n', k \in \mathbb{N} \text{ and functions } g, h, \quad \mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))h(Z_{n+1:n'})|Z_{1:n}) = 0. \quad (4)$$

Proof. It suffices to show the equivalence between the following three statements:

(i) $Z_{n+1:n+m}|Z_{1:n}$ satisfies Eq. (1)

(ii) for all $n' \geq n, k \geq 1$ and integrable function g we have $\mathbb{E}(g(Z_{n'+k}) - g(Z_{n'+1})|Z_{1:n}, Z_{n+1:n'}) = 0$

(iii) for all $n' \geq n, k \geq 1$ and integrable (g, h) we have $0 = \mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))h(Z_{n+1:n'})|Z_{1:n})$.

(i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (iii) follow immediately by properties of the conditional expectation. For (iii) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (i), let h and g range over the indicator functions for all measurable sets of the respective variables. □

Corollary 1. *Let* $\{Z_i : i \in \mathbb{N}\}$ *be a sequence of random variables satisfying the martingale property. Then for all integers* $n, n', k > 0$ *and* $n' > n$ *it holds that:* ***(i)*** $\mathbb{E}(g(Z_{n+1})|Z_{1:n}) = \mathbb{E}(g(Z_{n+k})|Z_{1:n})$ *for all integrable functions* g*, and* ***(ii)*** $\mathbb{E}((Z_{n'+k+1} - Z_{n'+1})Z_{n'}^\top|Z_{1:n}) = 0$.

Proof. (i) follows by setting $h(z_{n+1:n'}) \equiv \mathbb{1}$ in (4). (ii) follows by setting $g(z) = z, h(z_{n+1:n'}) = z_{n'}$. □

B Further discussion of theory and methodology

B.1 Motivating example: consequences of the martingale property

Consider a drug company exploring the efficacy of a new medication for headaches. The company runs a two-arm Randomised Control Trial (RCT) with 100 patients, 50 in each arm, comparing the new treatment with the current standard of care (in this case ibuprofen), and records the outcome $Y \in \{0, 1\}$ whether patients are symptom-free four hours after treatment. It is important to note that in this setting, the distribution of outcomes is independent of the order in which the patients are observed, a property known as *exchangeability* (see §2 for a formal definition). Half-way into the trial, the company conducts an interim analysis. Define the interim observations $D = \{(x_1, y_1), \ldots, (x_{50}, y_{50})\}$ where y_k indicates outcome, and x_k the treatment arm and other patient covariates. Given these observations, the company wants to decide whether to stop the trial early. The company uses an LLM, which was trained on potentially useful background information from the internet (e.g. on clinical trials, or the efficacy of ibuprofen), to generate the missing patients via ICL conditioning on $x_{1:n+k-1}$ for the $(n + k)$-th patient, and determines if the RCT is successful combining the observed and synthetic data. It repeats this imputation procedure J times, and decides to keep going with the trial if the fraction of symptom-free patients in the treatment over the control arm is above a certain threshold on average over these J hypothetical trials. *Should we trust the LLM's prediction using ICL under this procedure?*

In preview of our experimental results in §4, the answer is *'No'*. Our experiments present evidence that state-of-the-art LLMs violate the martingale property in certain settings (see Fig. 1). The martingale property is a necessary condition for exchangeability, and in turn a fundamental property of Bayesian learning. If the martingale property is violated by an LLM performing ICL it implies that the model's

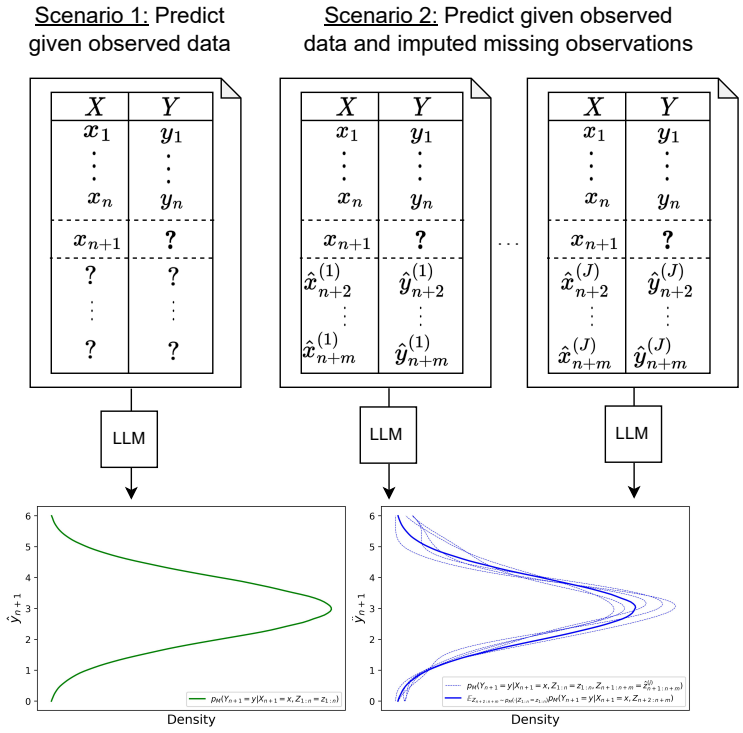


Figure 3: The *martingale property*, a fundamental requirement of a Bayesian learning system, requires *invariance with respect to missing samples from a population.*

predictions are not exchangeable, and hence that ICL with this LLM is not following any reasonable notion of probabilistic conditioning. This renders the LLM's predictive distribution incoherent: the model can make different predictions depending on the order in which the patients are imputed. This is problematic because by the design of an RCT, we know that there is no outcome dependence on the order of observations. It is incoherent and ambiguous to receive a different marginal predictions if we for example impute patient # 51 or patient # 100 first. Note that independent and identically distributed (i.i.d.) is a stricter condition implying exchangeability, and hence our work also applies to any i.i.d. data setting. This should caution the practitioner of the use of LLMs in exchangeable applications and data settings.

But there is a second reason why the martingale property is crucial: it enables a principled interpretation of the *uncertainty* of LLMs, allowing us to decompose inference into epistemic and aleatoric uncertainty (see §2 for a detailed introduction). Revisiting the RCT example above, if we acquire data from the 50 remaining patients, a costly decision, can this substantially decrease (epistemic) uncertainty? What is the effect of acquiring additional features for each patient, e.g. a genetic predisposition, on the (aleatoric) uncertainty? – Without satisfying the martingale property, we have no understanding of the effect on reducing uncertainty in applications where additional data acquisition is feasible, for instance active learning or reinforcement learning. We cannot study the question 'why is the point prediction of my LLM imprecise' in a principled way, and the uncertainty of an LLM's predictive distribution remains opaque. This finding has important implications for safety-critical, high-stakes applications of LLMs where trustworthy systems with a principled uncertainty estimate are vital.

B.2 Unambiguous predictions under exchangeable data

There is another way in which predictions are rendered unambiguous: under *exchangeability* for which the martingale property is a necessary condition (see App. A) the model is invariant to the order of the observed and missing data. This requirement is vital if we know that the order of the underlying distributions is irrelevant, for instance because—as in the RCT example in App. B.1—we have designed the experiment such that we can exclude a dependency on the order. Formally, this concept is known as exchangeability. A sequence of random variables $\{Z_i\} \sim p_M$ is *exchangeable* if

for all $\ell \in \mathbb{N}$ and ℓ-permutations σ,

$$p_M(Z_1, \ldots, Z_\ell) = p_M(Z_{\sigma(1)}, \ldots, Z_{\sigma(\ell)}). \tag{5}$$

Exchangeability guarantees the invariance of predictions to the ordering of the observations $Z_{1:n}$, but also with respect to the order of future imputations $Z_{n+1,\ldots}|Z_{1:n}$. To understand the importance of this, consider the RCT example in App. B.1, where $\{Z_{1:100}\}$ are (by experimental design) exchangeable. A model p_M should hence satisfy

$$\begin{aligned} &p_M(Y_{n+k}|X_{n+k}=x, Z_{1:n}, X_{n+1:n+k-1}=\hat{x}_{n+1:n+k-1}) \\ &= p_M(Y_{n+k}|X_{n+k}=x, Z_{1:n}, X_{n+1:n+k-1}=\hat{x}_{\sigma(n+1:n+k-1)}), \end{aligned}$$

meaning that the prediction for $Y_{n+k}|D, X_{n+k}$ is independent of the order of the imputed inputs $\hat{x}_{n+1:n+k-1}$. If a model p_M violates the above equality, there may be ambiguities in the prediction of the next sample (Y_{n+k}, X_{n+k}) as it may depend on and vary by ordering. Such ambiguities would substantially undermine the credibility of predictions, as well as the downstream decision-making based on such procedures. The martingale property is connected to the above notions of invariance as a necessary condition for exchangeability. Furthermore, it can even ensure exchangeability of imputed samples as the observed sample size n becomes large, because Eq. (1) implies asymptotic exchangeability of $Z_{n+1,\ldots}|Z_{1:n}$ (Berti et al., 2004, Thm. 2.5).

B.3 Background on martingale posteriors

We will show how the martingale property allows for a principled decomposition for the predictive uncertainty from the model.

Let us first suppose for simplicity that the variables Z_i are discrete and have $A < \infty$ realisations, so that any distribution $p_\theta(Z = \cdot)$ can be identified by a vector $\theta \in \mathbb{R}^A$. Let θ_n denote the random vector that indexes $p_M(Z_{n+1} \mid Z_{1:n})$. Then, the martingale property is equivalent to stating that $\{\theta_n\}$ form a martingale w.r.t. the filtration defined by $\{Z_n\}$. Now Doob's theorem (Doob, 1949) states that θ_n converges almost surely to a random vector θ_∞, and we have $\theta_n = \mathbb{E}_{\theta_\infty|Z_{1:n}}\theta_\infty$, or equivalently,

$$p_M(Z_{n+1}=\cdot\,|\,Z_{1:n}) = \int p(\theta_\infty|Z_{1:n}) p_{\theta_\infty}(Z=\cdot) d\theta_\infty. \tag{6}$$

The random vector θ_∞ plays the same role as the parameter θ in a Bayesian model, as both determine a predictive distribution. Indeed, the Bayesian posterior predictive distribution, Eq. (3), has the same representation as (6). We will also explain shortly when p_M is defined through Bayesian inference over θ, θ_∞ will be equivalent to the model parameter θ, and $p(\theta_\infty|Z_{1:n})$ equivalent to the Bayesian posterior. For these reasons we refer to the distribution $\theta_\infty|Z_{1:n}$ as the *martingale posterior*.

We note that we can construct the martingale posterior solely using samples from p_M. This is because we can construct the martingale posterior by sampling $Z_{n+1:n+m}|Z_{1:n}$, which will determine a sample $\theta_{n+m}|Z_{1:n}$ as the parameter that indexes the predictive distribution $p(Z_{n+m+1} = \cdot|Z_{1:n+m}) = p_{\theta_{n+m}}(\cdot)$; and since $\theta_{n+m} \to \theta_\infty$ as $m \to \infty$, we can truncate the process at a large $m \gg n$ to obtain a good approximation for θ_∞.

The restriction to finite support is largely for expository simplicity as it allows us to avoid measure-theoretic considerations. More generally, it is always possible to view the distribution $p(Z_{n+1} = \cdot|Z_{1:n}) =: \theta_n$ as a random element in a suitable Banach space of measures and the condition (1) as requiring $\{p(Z_{n+1} = \cdot|Z_{1:n}) : n \in \mathbb{N}\}$ to define a martingale in that space. When Doob's theorem applies, the above construction provides a distribution over predictive distributions that quantifies the epistemic uncertainty. However, for tractability and comparability to Bayesian parametric posteriors, it is also useful to consider the following alternative procedure which is also closer to Fong et al. (2021):

1. Sample $Z_{n+1:n+m} \sim p_M(\cdot|Z_{1:n})$.
2. Compute $\hat{\theta}_m := \arg\max_{\theta\in\Theta} \sum_{j=1}^m \log p(Z_{n+j}|\theta)$.
3. Return $\hat{\theta}_m$ as an approximate sample from the martingale posterior, defined as the conditional distribution of the pointwise limit $\lim_{m\to\infty} \hat{\theta}_m$ given $Z_{1:n}$.

In the above, $p(Z_i|\theta)$ is the likelihood in the Bayesian parametric model. If $\{p_M(Z_{n+j}|Z_{1:n+j-1})\}_{j=1}^{\infty}$ corresponds to a certain posterior predictive defined by the same likelihood, and the model is such that maximum likelihood estimation is consistent, it follows from de Finetti's theorem (applied to $Z_{n+1:}|Z_{1:n}$) and consistency that as $m \to \infty$, $\hat{\theta}_m$ will converge to a random variable $\hat{\theta}_\infty$ (w.r.t. the norm and notion of convergence in consistency), and the distribution $\hat{\theta}_\infty|Z_{1:n}$ must equal the Bayesian posterior. Applying the same procedure to a more general p_M that satisfies (1) leads to the methodology in Fong et al. (2021).

We adopted this 'model-based' approach in our third diagnostic (§B.6). Compared with the former approach, it is easier to implement on ICL tasks where each sample Z_i is represented with multiple tokens and a correctly specified likelihood for the true observations is available; the latter is always true in our synthetic experiments. More importantly, when m is finite, only with this approach can we compare the sampling distribution of $\hat{\theta}_m|Z_{1:n}$ across different p_M, as we explain in Appendix B.8 below. This is important in our experiments where we find the LLMs (at best) follow the martingale property within a horizon of $m \asymp n$.

B.4 Utility of the uncertainty decomposition

The interpretable decomposition of uncertainty further provides actionable guidance on how the combined uncertainty can be reduced: We can collect more samples to reduce epistemic uncertainty in scenarios where this is possible such as active learning, reinforcement learning or healthcare; particularly in regions of the input space where the uncertainty is high. In §B.6 we propose diagnostics to check if epistemic uncertainty decreases w.r.t. training sample size. On the contrary, if the aleatoric uncertainty is high and ought to be reduced, we cannot do so without 'changing the problem', for instance by collecting more features for each data point. This principled notion of uncertainty in a model is crucial in safety-critical, high-stakes scenarios for building trustworthy systems.

We present the following example for further intuition:

Example 1. Suppose $Z_i \in \{0, 1\}$. Then $\theta_\infty \in \mathbb{R}^2$, and $p_{\theta_\infty} = \mathrm{Bern}((\theta_\infty)_2)$ is determined by its second dimension. Thus, in both Eq. (6) and Eq. (3) the epistemic uncertainty is represented by a distribution over the Bernoulli parameter, revealing their inherent connection. The epistemic uncertainty is especially important in scenarios where we use a model p_M to impute the missing samples $\{Z_{n+i}\}$ from a population —as in the RCT example in §B.1— and want to quantify a model's lack of knowledge about the population. Note this distribution is not identifiable if we only have samples from a single-step predictive distribution $p_M(Z_{n+1}|Z_{1:n})$, but becomes identifiable given *sample paths*.

B.5 Diagnostics for the martingale property: further discussion

We here provide further discussion of Corollary 1. Property (i) follows by setting $h(Z_{n+1:n'}) \equiv 1$ and examines the marginal predictive distributions $p_M(Z_{n+k}|Z_{1:n})$. We instantiate (i) using (at most) two choices of g: In preview of §4, we will perform our checks on unconditional experiments where Z_i—or equivalently Y_i because of the unconditional setting—are Bernoulli or Gaussian distributed random variables. In the Bernoulli experiment it suffices to choose the identity function $g(z) = z$, as the mean $\mathbb{E}(Z_{n+k}|Z_{1:n})$ provides full information about the distribution $p_M(Z_{n+k}|Z_{1:n})$. In the Gaussian experiment, we will observe that choosing $g(z) = z$ and $g(z) = z^2$ is in most cases sufficient to reveal substantial violations from the martingale property.

Property (ii) is equivalent to requiring Eq. (4) to hold for all linear functions (g, h), which follows by linearity of the functions and the conditional expectation. We will again see in our experiments that this choice is usually sufficient to reveal deviations from the martingale property.

Let us further consider choices for h and g in Corollary 1 with an example.

Example 2. Suppose p_M is defined through a Bayesian model for i.i.d. observations. Let θ denote the latent model parameter, and suppose the likelihood $p(Z|\theta)$ satisfies $\mathbb{E}_{Z\sim p(Z|\theta)} Z = \theta$. Then by Corollary 1, for all (k, n') we have

- $\mathbb{E}(Z_{n+k}|Z_{1:n}) = \mathbb{E}(\theta|Z_{1:n})$, and
- $\mathbb{E}(Z_{n'+k+1} Z_{n'+1}^\top|Z_{1:n}) = \mathbb{E}(\theta\theta^\top|Z_{1:n})$ (see e.g. Ghosal & Van der Vaart, 2017, p. 454).

In this setting, condition (i) (with $g(z) = z$) and (ii) thus guarantee that the conditional mean and covariance equal the posterior mean and covariance, respectively, independent of (n', k). These two important aspects of the posterior are hence consistently expressed by the model.

B.6 Diagnostics for epistemic uncertainty

As discussed in §2, the martingale property allows us to identify epistemic uncertainty, which should decrease with more observed samples. Here, we derive a third diagnostic for Bayesian ICL systems which probes this. We begin by presenting a theoretical fact which provides important intuition on the role of epistemic uncertainty.

Fact 2. Let $\pi(\theta)$ and $p_M(Z|\theta)$ be the prior and likelihood of a Bayesian model, $\bar{\theta}_n := \mathbb{E}_{\theta\sim\pi(\theta|z_{1:n})}\theta$ the posterior mean given data $z_{1:n}$, and $\|\cdot\|$ be any vector norm. Then,

$$\mathbb{E}_{\theta_0\sim\pi, z_{1:n}\sim\pi(z|\theta_0)}\mathbb{E}_{\theta\sim\pi(\theta|z_{1:n})}\|\theta - \bar{\theta}_n\|^2 = \mathbb{E}_{\theta_0\sim\pi, z_{1:n}\sim\pi(z|\theta_0)}\|\theta_0 - \bar{\theta}_n\|^2. \tag{7}$$

Proof. This holds because θ and θ_0 are conditionally independent and identically distributed given $z_{1:n}$, and $\bar{\theta}_n$ equals the conditional expectation of both random variables. □

The left-hand side in Eq. (7) is the trace of the posterior covariance (variance) and thus measures epistemic uncertainty. The right-hand side is the estimation error for the true parameter. Thus, Fact 2 states that *epistemic uncertainty provides a quantification for the average-case estimation error*. Note that Eq. (7) only applies to data from the prior predictive distribution, and thus not necessarily to the real observations. Nonetheless, a significant deviation of a model from the known scaling behaviour of the estimation error will indicate non-conformance with any reasonable Bayesian models. This is precisely our starting point to derive another diagnostic for Bayesian ICL systems.

As discussed in §2, we use sample paths generated by an LLM to approximate a martingale posterior and estimate its epistemic uncertainty. Here, we characterise epistemic uncertainty through the trace of the posterior covariance of the martingale posterior, the 'spread' of the distribution. Because the sample paths we use are finite (see App. B.7) we cannot study the exact martingale posterior directly, which can only be recovered with infinite samples. Instead, we study the sampling distribution of the maximum likelihood estimate (MLE) on the first m samples: $\hat{\theta}_m := \arg\max_{\theta\in\Theta}\sum_{i=1}^{m}\log p_\theta(Z_{n+i})$, where p_θ is the known parametric likelihood. We measure the spread of this distribution using its *inter-quartile range*

$$T_3 = Q_{0.75}(\{\hat{\theta}_m^{(j)}\}_{j=1}^{J}) - Q_{0.25}(\{\hat{\theta}_m^{(j)}\}_{j=1}^{J}), \tag{8}$$

where $\hat{\theta}_m^{(j)}$ denotes the MLE using the j-th sample path $\{z_{n+i}^{(j)}\}_{i=1}^{m}$, and $Q_{0.25}$ and $Q_{0.75}$ are the 0.25- and 0.75-quantiles. In our experiments in §4 we consider scenarios where the true data distribution is defined by regular parametric models. In such cases the optimal (squared) estimation error for the true parameter scales $O(d/n)$ where n is the ICL dataset size and d is the dimension of the parameter, which is also the minimax lower bound (Van der Vaart, 2000, Ch. 8). When choosing $m = \Theta(n)$, a reference Bayesian model will also have the $O(d/n)$ scaling behaviour following classical posterior contraction results in statistics; see App. B.8. Therefore, we can compare the asymptotic scaling of T_3 between an LLM and a reference Bayesian parametric model through the same sampling-based procedure. If the scaling behaviour of T_3 from our LLM deviates from that of the reference Bayesian model, we can conclude that the LLM either exhibits a marked loss of estimation efficiency, or does not maintain a correct notion of epistemic uncertainty at all. Both characteristics contradict a Bayesian ICL system and are undesirable.

B.7 Are all deviations from Bayes bad? – Expected and acceptable deviations from Bayesian reasoning

Numerous properties are implied if a learning system satisfies the martingale property, a distributional characteristic, and it is both infeasible and unnecessary as often practically irrelevant to check all of them in order to provide evidence for or against our hypothesis. For example, the martingale property implies that all conditional moments should be equivalent, i.e. $\mathbb{E}(Z_{n'+1}^l|Z_{1:n}) = \mathbb{E}(Z_{n'+k}^l|Z_{1:n})$ for all integers $n, n', k, l > 0$ and $n' > n$, yet higher-order moments are not vital in most applications

and hence are acceptable deviations, if existent. Therefore, we will restrict our attention to two key implications of the martingale property which—if present—have important practical consequences.

Pretrained LLMs are general-purpose models and can at best approximate Bayesian learning via ICL. The martingale property is an invariance that is not hard-coded in their transformer-based architecture, and can only be approximately (rather than exactly) satisfied. Let us assume that an LLM internally maintains a 'hierarchy of states' Wang et al. (2023), say a hierarchical Bayesian model, capturing different tasks (e.g. Bayesian ICL from i.i.d. data, or acting in a dialogue system), and at each sampling step first updates its belief about this state. Say there is a probability p that the LLM deviates from Bayesian ICL or simply fails to approximate. Even if p is small, the probability of a deviation $1-(1-p)^m$ becomes substantial when accumulated over a long sampling path of length m. This would trivially falsify the martingale property and our hypothesis.

In our experiments in §4, we hence restrict the sampling paths to a short, finite length where we check the martingale property. We also design our checks to be robust against such behaviour, for example by removing outliers before computing a test statistic. Furthermore, we are particularly interested in stark and unequivocal evidence of the model violating the martingale property beyond an expected error of any approximating model. We will analyse and quantify violations of the martingale property with diagnostics, which we introduce in §3, in order to check our hypothesis experimentally. In App. B.9 we derive the order of 'acceptable violations' for the test statistics we will introduce.

B.8 Approximate martingale posteriors with finite paths

We have claimed that with a finite m, the spread of the approximate martingale posterior $\hat{\theta}_m$ defined as the MLE on m samples (see §B.6, or §B.3) is comparable between different choices of p_M. We now substantiate on this claim.

Let us first restrict to exchangeable (i.e., Bayesian) choices of p_M. Consider de Finetti's representation for the posterior predictive measure: $Z_{n+1,\dots}|Z_{1:n}$ can be represented through

$$\theta_\infty \sim \pi(\cdot|Z_{1:n}), \;\; Z_{n+1,\dots} \overset{iid}{\sim} p(\cdot|\theta_\infty)$$

where the measure $\pi(\cdot|Z_{1:n})$ equals the Bayesian posterior, which as discussed in §B.3 equals the exact martingale posterior. Combining the above representation and the fact that $\hat{\theta}_m$ is a function of $Z_{n+1:n+m}$ lead to $\hat{\theta}_m \perp Z_{1:n}|\theta_\infty$, and

$$\begin{aligned}
&\mathrm{Cov}(\hat{\theta}_m|Z_{1:n}) \\
&= \mathbb{E}(\mathrm{Cov}(\hat{\theta}_m|\theta_\infty)|Z_{1:n}) + \mathrm{Cov}(\mathbb{E}(\hat{\theta}_m|\theta_\infty)|Z_{1:n}) \\
&\approx \mathbb{E}(\mathrm{Cov}(\hat{\theta}_m|\theta_\infty)|Z_{1:n}) + \mathrm{Cov}(\theta_\infty|Z_{1:n}),
\end{aligned}$$

where we dropped the term $\mathbb{E}(\hat{\theta}_m|\theta_\infty) - \theta_\infty$ which is the bias of MLE and thus a higher-order term for regular models. Therefore, that the (co)variance overhead $\mathrm{Cov}(\hat{\theta}_m|Z_{1:n}) - \mathrm{Cov}(\theta_\infty|Z_{1:n})$ is, up to the first order, the average-case error of MLE on m i.i.d. samples when the true parameter is sampled from the posterior $\pi(\cdot|Z_{1:n})$. For regular models this is always $\Theta(d/m)$, where the coefficient hidden in the Θ notation is also comparable across different p_M as long as the Fisher information matrix evaluated at $\theta \sim \pi(\cdot|Z_{1:n})$ has a comparable value (e.g., across all choices of p_M that satisfy *consistency*). As the martingale posterior covariance $\mathrm{Cov}(\theta_\infty|Z_{1:n})$ has the same $\Theta(d/n)$ scaling across all regular Bayesian models to which the Bernstein von-Mises theorem applies, with a choice of $m \asymp n$, any deviation in the scaling of $\mathrm{Cov}(\hat{\theta}_m)$ – from that of any regular Bayesian model – must be attributable to a different scaling of the exact MP covariance, and thus a deviation from all regular Bayesian models.

Finally, we note that while we focus on ICL models that are approximately Bayesian, the above discussion may also apply to general models that only satisfy the martingale property, since for those models $Z_{n+1,\dots}|Z_{1:n}$ remains asymptotically exchangeable (Berti et al., 2004). Moreover, the above discussion applies to inter-quantile range (IQR) as well, because for asymptotically normal posteriors the IQR is proportional to the posterior standard deviation; and even for non-normal posteriors, the IQR should still have the same order as the posterior contraction rate, by definition of the latter.

B.9 Acceptable approximation errors of properties (i) and (ii) in Corollary 1

Even as we restrict to a finite horizon m, there can still be expected deviations from the equality (1), and thus those in Corollary 1, simply because (1) represents invariance conditions that are not "hard-wired" in the LLM's architectures. yet it is logical that small violation of these equalities should not have practical consequences. We now derive the order of their acceptable violations in the setting of Example 2.

As discussed therein, the equalities in Corollary 1 guarantee the expressions for posterior mean and covariance for the parameter θ have consistently defined values, regardless of the choices of (n', k). The posterior mean has the order of $\Theta(1)$ and requires the violation of Corollary 1 (i) to be $o(1)$. The posterior covariance is generally $\Omega(1/n)$ and can be expressed through Example 2 as

$$\mathrm{Cov}(\theta|Z_{1:n}) = \mathbb{E}(Z_{n'}Z_{n'+k}|Z_{1:n}) - \mathbb{E}(Z_{n'}|Z_{1:n})^2.$$

Therefore, it can has an approximately consistent value if the equalities in Corollary 1 hold approximately *up to an error of* $o(1/n)$. Posterior mean and covariance are key quantities in the interpretation of predictive uncertainty, which in turn is a major benefit from the martingale property. Thus, we consider the above deviation to be acceptable as it already guarantees the approximately consistent interpretation of predictive uncertainty through the martingale property.

C Additional experimental details and results

C.1 Additional experimental details

Test statistics of properties implied by the martingale property. We summarise and empirically measure properties (i) and (ii) in Corollary 1 using the aggregated statistics

$$T_{1,g} := \frac{2}{Jm}\sum_{j=1}^{J}\sum_{i=1}^{m/2}(g(z_{n+i}^{(j)}) - g(z_{n+i+m/2}^{(j)})), \tag{9}$$

$$T_{2,k} := \frac{1}{Jm}\sum_{j=1}^{J}\sum_{i=1}^{m-k-1}(z_{n+i+1}^{(j)} - z_{n+i+k}^{(j)})z_{n+i}^{(j)}. \tag{10}$$

The statistics $T_{1,g}$ and $T_{2,k}$ are defined using samples $\{z_{n+i}^{(j)}\}$ from J paths generated by an LLM via ICL and correspond to Monte-Carlo estimates of the expectations in properties (i) and (ii). To be robust against the possible outlier paths (App. B.7), we remove sample paths with anomalous mean absolute values using the standard 1.5×IQR rule.

Experimental setup. For the first two experiments we vary $n \in \{20, 50, 100\}$, $m \in \{n/2, 2n\}$ and sample $J = 200$ paths from the LLMs. For the natural language experiments we fix $n = 100, m = 50, J = 80$. As non-exchangeable models may demonstrate different behaviour on different permutations of the same dataset, for our experiments we permute the observations when generating each sample path, so that we can produce a single test statistic that summarises each experiment configuration. For the epistemic uncertainty experiments (third diagnostic) in App. C.2, however, we use a fixed ordering for the observations for all path samples within each run, and report the median inter-quartile range across 9 runs for each configuration. This change is made to avoid (possibly small) deviations from exchangeability from inflating the estimated spread of the posterior.

We discuss prompt design and format in detail below. Here we emphasise that across all tasks, the prompt always includes sufficient information about the true likelihood.

Due to resource limitations we only employ `gpt-4` for the Bernoulli experiments with $n \leq 50$.

Prompt design and format. We use the following prompt format `<instruction> <observed data> <sampled data>`. `<instruction>` describes the distribution of the observed data and importantly states that the observed samples were drawn i.i.d., i.e. from exchangeable random variables. `<observed data>` and `<sampled data>` lists the observed $z_{1:n}$, and sampled data $\hat{z}_{n+k}$ (if there exists any), respectively. Samples are represented depending on the

experiment: as `int` values as 1-digit characters (e.g. '1'), `float` values with 1-digit of precision (e.g. '2.2') or words for synthetic natural language. As a sanity check, we also consider replacing integers with random words (e.g. 'tiger' for '1', 'hedgehog' for '0'), but did not notice important differences between the LLMs' behaviour. Each sample is delineated by a separator (e.g. ';').

We present exemplary prompts for each dataset below:

- A Bernoulli experiment with $n = 5$ and $m = 2$: *Provided are independent, identically distributed tosses of a coin, which flips 1 with probability p where p is unknown: 1;0,0,1,0,0;1.*
- A Gaussian experiment with $n = 1$ and $m = 2$: *Provided are independent, identically distributed draws from a Gaussian, with fixed but unknown mean and unit variance: 1.1,0.8,1.3,1.0,0.9,1.2,0.8.*
- The the natural language experiment: *You will make predictions for a novel disease. The observed dataset contains records for multiple subjects which are assumed to be independent and identically distributed. For each subject there are two binary variables, indicating fever and disease diagnosis, respectively. Output your prediction for the disease diagnosis of the next subject.\n Id: 0\n Fever: Y\n Diagnosis: N ...*

Other work represents both `int` and `float` numbers as a space-separated string of digits with fixed precision, where each number is separated by a semi-colon. This guarantees a per-digit tokenisation that was observed to be beneficial in the context of time series forecasting and further minimises the required number of tokens per number as the decimal point is redundant Gruver et al. (2023). We did not opt for this representation and corresponding tokenisation for two reasons: First, initial experiments with GPT-2 showed deteriorating sampling performance, where the model often hallucinated unrelated content. Second, and related to the first point, this representation is somewhat 'out-of-distribution' and probably unseen in the training distribution, which could limit and constrain any conclusions made in our experiments. Note that because of the tokenisation, in §4, the Gaussian experiment is more difficult than the Bernoulli experiment (or any dataset with single-token samples) as the LLM is required to learn the correlation structure between consecutive tokens representing a real-valued number.

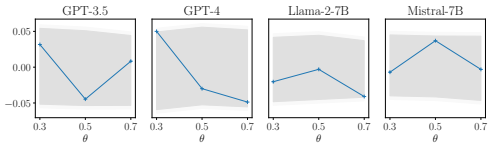
Additional details for the natural language experiment. For the natural language experiment, we modify the scheme as follows: we split the ICL dataset and the imputations into two sequences $(\{Y_{i_{0,k}}\}_{k=1}^{n_1+m_1}, \{Y_{i_{1,k}}\}_{k=1}^{n_0+m_0})$ based on the value of X_i. Subsequently, either sequence contains i.i.d. Bernoulli random variables with a different mean, and any Bayesian ICL model with a correctly specified likelihood must produce imputations following a separate Bayesian posterior for Bernoulli data. Thus, we can apply our Bernoulli diagnostics separately to both sequence. This modification allows us to focus on LLMs' conditional predictive distributions of the form $p_M(Y_{i+1}|X_{i+1}, Z_{1:i})$, which is more relevant in practice.

C.2 Further experimental results

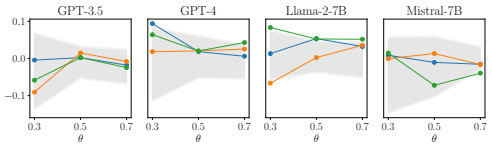
Gaussian experiment. In Fig. 7 we present results on the Gaussian experiment with $\theta = -1, n = 100, m = n/2$, again performing both checks of the martingale property and using a reference Bayesian model with the non-informative prior $\mathcal{N}(0, 100)$. We observe that `llama-2-7b` and `mistral-7b` demonstrate clear deviation from the martingale property, whereas `gpt-3.5` passes both checks. Additional results for `gpt-3.5` in App. C present our diagnostics with other choices of (n, m, θ), demonstrating a deviation to the predictive distribution of the reference Bayesian posterior. We will also further investigate `gpt-3.5` in the natural language experiment. In conclusion, the presented evidence on the Gaussian experiment falsifies our hypothesis of Bayesian behaviour with the tested LLMs.

Synthetic natural language experiment. In Fig. 8 [Left] we present our results for the natural language experiment with $n = 100, m = 50, g(z) = z$ and `gpt-3.5`. For both checks of the martingale property, we observe clear deviations from a reference Bayesian posterior for all values of X_i (see App. C for details). This provides further evidence of violations of the martingale property in settings where natural language (instead of numbers) are used.

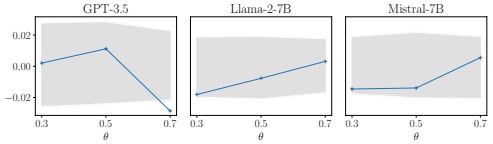
Checking epistemic uncertainty of LLMs. In this subsection we analyse the scaling behaviour of an LLM's uncertainty. In Fig. 8 [Right] we measure T_3 (y-axis on a log-scale) and compare



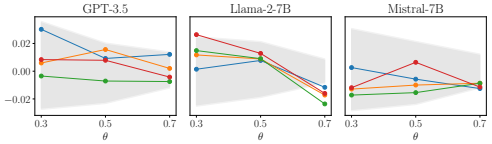
(a) $T_{1,g}$ for $g(z) = z, n = 20, m = 10$



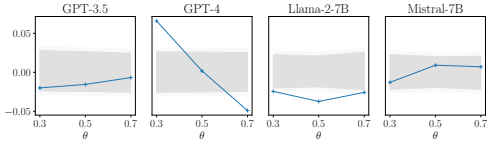
(b) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 20, m = 10$



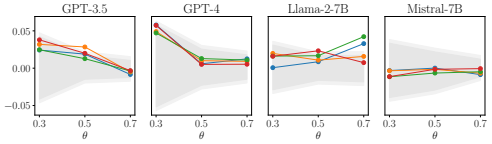
(c) $T_{1,g}$ for $g(z) = z, n = 50, m = 25$



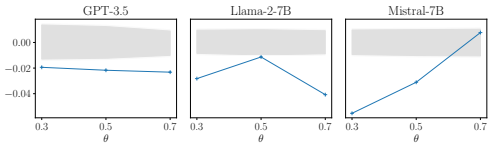
(d) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 50, m = 25$



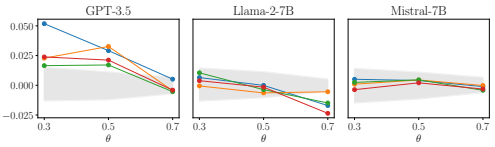
(e) $T_{1,g}$ for $g(z) = z, n = 20, m = 40$



(f) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 20, m = 40$

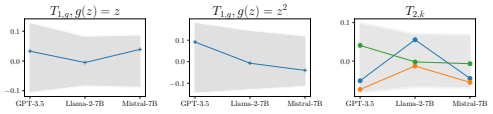


(g) $T_{1,g}$ for $g(z) = z, n = 50, m = 100$

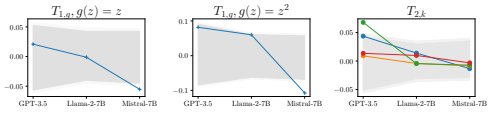


(h) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 50, m = 100$

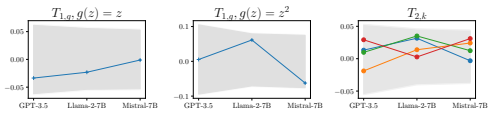
Figure 4: Checking the martingale property: full results for the Bernoulli experiments in the setting of Fig. 2.



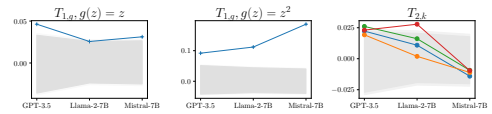
(a) $n = 20, m = 10$



(b) $n = 20, m = 40$

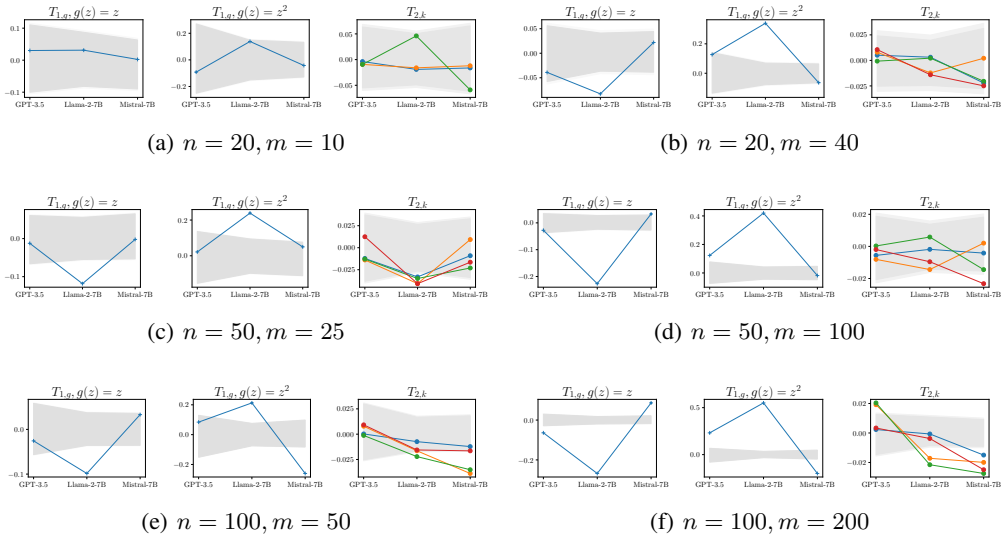


(c) $n = 50, m = 25$



(d) $n = 50, m = 100$

Figure 5: Checking the martingale property: results for the Gaussian experiments with $\theta = 0$. See Fig. 7 for details.



(a) $n = 20, m = 10$ (b) $n = 20, m = 40$

(c) $n = 50, m = 25$ (d) $n = 50, m = 100$

(e) $n = 100, m = 50$ (f) $n = 100, m = 200$

Figure 6: Checking the martingale property: results for the Gaussian experiments with $\theta = -1$. See Fig. 7 for details.

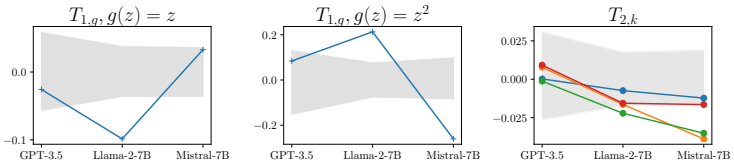


Figure 7: *Checking the martingale property on Gaussian experiments.* We present runs with $\theta = -1, n = 100, m = 50$ from different LLMs (x-axis) with test functions $g(z) = z$ and $g(z) = z^2$. See Fig. 2 for further details.

the approximate martingale posterior of an LLM with a reference Bayesian model when increasing the number of observed samples n (x-axis). We consider a Bernoulli experiment with $\theta = 0.5$ as it is the only experimental setting where, with a short sampling horizon of $m = n/2$, all LLMs approximately adhere to the martingale property. In addition to the standard reference Bayesian model, we also consider two α-fractional Bayesian posteriors (Bhattacharya et al., 2019), which are generalisations of the Bayesian posterior that exhibit a $O(d/\alpha n)$ scaling for its epistemic uncertainty. They allow us to check the weaker hypothesis whether an LLM's epistemic uncertainty scales at least up to the correct order of magnitude. We observe that the asymptotic rate of of `llama-2-7b` and `gpt-3.5` is slower than that of a Bayesian model, which suggests inefficiency as discussed in §B.6. Furthermore, `gpt-3.5` demonstrates over-confidence in the small-sample regime. `mistral-7b` appears to scale at least on a correct order of magnitude, even though not exactly matching the Bayesian model. This finding is interesting as on the Bernoulli experiments, `mistral-7b` also demonstrates the best adherence to the martingale property.

Checking the martingale property. Fig. 4 reports the full results for the Bernoulli experiment in the setting of Fig. 2 ($m \in \{n/2, 2n\}$), where we also visualises the $o(1/n)$ 'acceptable deviation' (§B.9) using a light shade with width $0.1/n$. Consistent with the results in Fig. 2, the martingale property is generally satisfied in the short-horizon scheme ($m = n/2$), but increasingly violated as we move to $m = 2n$. We further provide the results for $m = 10n$ in Fig. 9, where we drop `gpt-3.5` due to limitations with its API. As we can see, in this setting where the sampling horizon becomes even longer, deviation from the martingale property also becomes more severe. The consistently large negative value of $T_{1,g}$ indicates a continual upward bias towards 1, which demonstrates the 'creation of new knowledge' phenomenon discussed in §2.

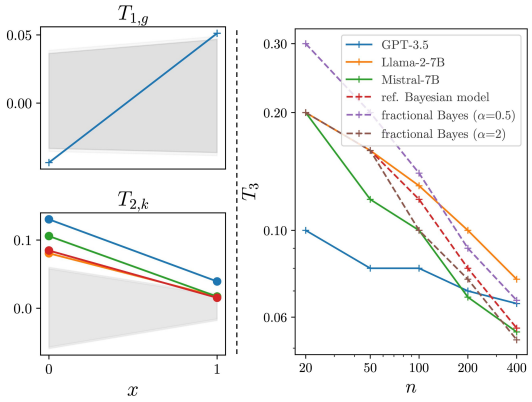
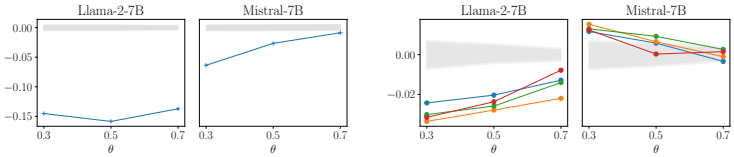


Figure 8: [Left] *Checking the martingale property on the natural language experiment.* We present both checks with test statistics computed separately for each value of X_i (x-axis). See Fig. 2 for further details. [Right] *Scaling of epistemic uncertainty on the Bernoulli experiment*: the test statistic T_3 (see §B.6) computed on LLMs, compared with Bayesian and fractional Bayesian models.

We report additional results for the Gaussian experiment in Fig. 6 ($\theta = -1$) and Fig. 5 ($\theta = 0$). As we can see that, all models generally demonstrate a deviation from the martingale property when $\theta = -1$, but with $\theta = 0$ they may often appear to satisfy the property within a shorter horizon ($m = n/2$). Results for $\theta = 1$ are similar to the $\theta = -1$ case and thus omitted. We note that in many cases the predictive distribution cannot be matched to any Bayesian posterior with the correct likelihood: for the latter the sample variance should be greater than 1, the likelihood variance, but this is often not true for the LLMs. For example, for `gpt-3.5` in the setting of Fig. 7 we find the sample variance to be $0.711 < 1$ (95% CI: $[0.680, 0.742]$).



(a) $T_{1,g}, g(z) = z$ (b) $T_{2,k}, k \in \{2, 3, 4, 5\}$

Figure 9: Checking the martingale property on Bernoulli experiments: additional result with $n = 100, m = 10n$. See Fig. 2 for details.

Scaling of epistemic uncertainty. As noted in the text, the behaviour of `gpt-3.5` and `llama-2-7b` cannot correspond to any 'reasonable' Bayesian models in the Bernstein von-Mises sense. Here we note that the same figure also suggests that they are unlikely to correspond to any 'unreasonable' Bayesian model (e.g., one with an approximately degenerate prior), either. For `gpt-3.5`, its small-sample behaviour can only be explained as a Bayesian model with a very strong prior that has the bulk of its mass near the true parameter; yet this would contradict its larger-than-regular posterior spread when n is large. For `llama-2-7b`, its large-sample behaviour could only be explained with the exact opposite (e.g., a $\mathrm{Beta}(100, 100)$ prior); yet that should have led to a much larger IQR when n is small.

D Related Work

In-context learning as Bayesian inference. Numerous papers have explained ICL as performing a form of Bayesian Inference. Xie et al. (2021) assume the pretraining distribution is a Hidden Markov Model (HMM). Under this and other assumptions pertaining transition probabilities and the distribution shift between the start of the prompt and all hidden transition distributions, they prove that the LLM infers a latent concept of the prompt which allows it to generate the next token, i.e.

infer its predictive distribution, implicitly performing Bayesian inference. Huszár (2022) connected their contribution to exchangeability. Hahn & Goyal (2023)[Section 1.4] relates to (Xie et al., 2021) as it can similarly be understood in terms of Bayesian inference, with the difference that they view the training tasks to be open-ended and compositional, in contrast to the finite nature of an HMM. Wang et al. (2023) likewise takes a Bayesian viewpoint, which they utilise to select the ICL dataset optimally. Jiang (2023) explains various phenomena of the 'emergent abilities' of LLMs, such as in-context learning and chain-of-thought prompting, through through Bayesian inference on the common distribution underlying natural languages. Zhang et al. (2023b) show that ICL implicitly uses a Bayesian model averaging. Griffiths & Tenenbaum (2006) recover the prior distributions in LLMs for everyday observations, such as the time of movies.

Theories for in-context learning. Numerous theoretical models and frameworks beyond Bayesian inference exist which aim at understanding and formalising ICL. We refer to Dong et al. (2022) for a detailed survey on in-context learning. Akyürek et al. (2022) prove that transformer-based architectures can implement classical learning algorithms such as linear models and ridge regression. Bai et al. (2023) extend this work by demonstrating that ICL via transformers can implement and even braoder set of algorithms, including convex risk minimisation algorithms and gradient descent, where the model intrinsically selects a different learning algorithm based on the task at hand. Singh et al. (2023) shows that the ability of performing ICL algorithms such as Bayesian inference may be a transient phenomenon which produces highest accuracy during certain stages of pretraining an LLM. Raventós et al. (2023) show that the ability of in-context learning to tasks unseen during training by picking the right learning algorithm depends on the task diversity during training.

Input order dependence of Large language models. Previous work has found a dependence of LLMs on the order in which an input sequence is presented. Lu et al. (2021) demonstrate that input order can significantly change the performance of an LLM in text classification tasks from "state-of-the-art" to "random guess". In the context of few-shot learning, Zhao et al. (2021) show the prediction of an LLM can depend on many seemingly irrelevant items, such as the prompt format or the order in which input examples are presented in a prompt, again with a sensitivity of performance to these factors. Zhang et al. (2023a) note that the topic structure of a document may be exchangeable, which motivates them to use Bayesian models, namely a Latent Dirichlet Allocation, to analyse the representations of an LLM. Our discussion on exchangeability relates to this line of work, but has a novel perspective on it through our focus on the martingale property, a necessary condition for exchangeability, among other implications of the martingale property. Furthermore, in contrast to the related work, which shuffles the input data $Z_{1:n}$, we analysing the effect of shuffling the imputed, generated sequence $Z_{n+1}, \dots$ where we find non-exchangeable behaviour, deviating from any reasonable Bayesian model.

Miscellaneous. Our work also relates a number of applications of LLMs. As we are generating samples from an LLM with ICL, which as we demonstrate deviate from the distribution of the ICL dataset, this work relates to and has implications for a line of work on LLMs for synthetic data generation Borisov et al. (2022); Hämäläinen et al. (2023); Tang et al. (2023); Veselovsky et al. (2023); Li et al. (2023). Furthermore, we show that the martingale property is violated for long sampling paths, which may have implications for time series prediction with LLMs Gruver et al. (2023); Jin et al. (2023), particularly over long horizons. We also demonstrate a dependence on the order in which missing values are imputed, which has direct implications for the concrete purpose of missing value imputations with LLMs Mei et al. (2021). Shumailov et al. (2023) demonstrate that models (including LLMs) which are recursively trained on data which they have previously generated shift in their distribution, where long tails disappear. While this work 'conditions' on data synthetic data by retraining, our work analyses the conditioning via ICL. Lastly, as LLMs violate the martingale property in certain empirical regimes, they hence do not allow for an decomposed interpretation of their predictive uncertainty, which has important implications for uncertainty quantification with LLMs Xiao et al. (2022).

E Negative societal impact

This paper analyses and attempts to characterise the behaviour of LLMs. We try to understand whether ICL in LLM follows Bayesian principles. As we outlined in §2 this has important consequences for

their potential use as trustworthy systems, which can deployed in safety-critical, high-stakes applications such as healthcare. These systems often crucially rely on a principled notion of uncertainty. The evidence presented in this work cautions against the use of LLMs in such settings without further checks as they—under certain experimental settings—do not possess such a principled interpretation of uncertainty, rendering their uncertainty 'black-box'. Furthermore, while LLMs have typically been trained in non-exchangeable scenarios (e.g. natural language where the order of words or tokens changes meaning), as we showed in §2, we caution against their use in exchangeable settings (e.g. i.i.d. data) as their predictions can be rendered inconsistent.

Both points noted above are potential negative societal impacts if Bayesian behaviour cannot be guaranteed by a model, as we argue in this work. While we do not see any direct negative consequences from the analysis in this work, we believe this work provides ample pointers and reason for further investigation of these concerns, and shall warn against potentially intended misuse of LLMs.

F Code, computational resources, datasets, existing assets used

Code. We do not provide code as part of this workshop paper, noting that we will release a full paper with our code base in due course.

Datasets. We used three synthetic datasets for our experiments: a coin flip experiment, sampling from univariate Bernoulli distributions, a Gaussian experiment, sampling from univariate Gaussian distributions, and a synthetic natural language experiment, sampling (conditionally) from Bernoulli distributions. We refer to §4 and App. C where they are introduced and discussed further details.

Computational resources and APIs used. Referring to §4, we implemented `llama-2-7B` and `mistral-7B` with the Huggingface Transformer library Wolf et al. (2020), and implemented `gpt-3.5` and `gpt-4` using the OpenAI API OpenAI (2023). For all Huggingface models, we generated the sampling paths by performing inference on a single A100 Nvidia GPU for each run.

Existing assets used. Our work uses the following main software libraries and corresponding licenses: PyTorch Paszke et al. (2019) (custom license), `numpy` Harris et al. (2020) (BSD 3-Clause License), Weights&Biases Biewald (2020) (MIT License), Huggingface transformers library Wolf et al. (2020) (Apache License 2.0; model licenses see below), `matplotlib` Hunter (2007) (PSF License), `tqdm` tqdm contributors (2022) (MPLv2.0 MIT License), `scikit-learn` and `sklearn` Pedregosa et al. (2011) (BSD 3-Clause License), `pandas` Wes McKinney (2010) (BSD 3-Clause License), `openai` (Apache 2.0 License), `tiktoken` (MIT License), and `pickle` Van Rossum (2020) (License N/A). We use Github Copilot and ChatGPT OpenAI (2023) for code development and occasionally as a writing aid.

We used three pretrained large language models (see §4): `llama-2-7B` Touvron et al. (2023) (custom license), `mistral-7B` Jiang et al. (2023) (Apache 2.0 License), and `gpt-3.5` Brown et al. (2020) (API; no code license).