Is In-Context Learning in Large Language Models Bayesian? A Martingale Perspective

Fabian Falck [*1] **Ziyu Wang** [*1] **Chris Holmes** [1]

Abstract

In-context learning (ICL) has emerged as a particularly remarkable characteristic of Large Language Models (LLM): given a pretrained LLM and an observed dataset, LLMs can make predictions for new data points from the same distribution without fine-tuning. 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 for 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

Large Language Models (LLMs) are autoregressive generative models trained on vast amounts of data, exhibiting extraordinary performance across a wide array of tasks (Zhao et al., 2023). A particularly remarkable characteristic of LLMs is so-called *in-context learning* (ICL) (Brown et al., 2020; Dong et al., 2022): Given a pretrained language model p_M and an observed dataset $D := \{(x_1, y_1), \ldots, (x_n, y_n)\} = z_{1:n}$ of samples, LLMs capture the distribution of the underlying random variables X and Y in this in-context dataset. 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.

Few-shot learning via ICL (Brown et al., 2020) has produced numerous breakthroughs in LLM research (Dong et al., 2022), such as in supervised learning (Min et al., 2021) or chain-of-thought prompting (Wei et al., 2022). 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 on exchangeable data (approximately) Bayesian?**

In contrast to prior work, our analysis focuses on one fundamental property of Bayesian learning systems for exchangeable data: 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, but begin by intuitively describing two important and desirable *consequences* of it with an example, highlighting its relevance. These consequences are: (i) the martingale property is a necessary condition for rendering *predictions unambiguous* in an *exchangeable* data setting, and (ii) it establishes a *principled notion* of the model's *uncertainty*.

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 *ex-*

[*]Equal contribution [1]Department of Statistics, University of Oxford, Oxford, UK. Correspondence to: Fabian Falck <fabian.falck@stats.ox.ac.uk>, Ziyu Wang <ziyu.wang@stats.ox.ac.uk>, Chris Holmes <cholmes@stats.ox.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

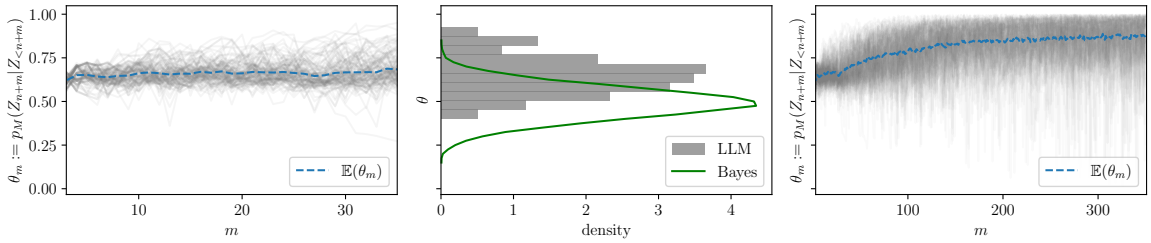


Figure 1: *In-context learning in Large Language Models is not Bayesian.* [Left] The *martingale property*, a necessary condition of Bayesian learning systems, is satisfied for short sample paths. [Centre] This allows us to approximate the *martingale posterior* (see §2.3) which, however, indicates deviation 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.

changeability (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 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.

This work states the hypothesis that ICL in LLMs given exchangeable data is Bayesian. Numerous works have argued that ICL approximates some form of Bayesian inference (Xie et al., 2021; Hahn & Goyal, 2023; Akyürek et al., 2022; Zhang et al., 2023b; 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, GPT-3.5 and GPT-4. 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. Our goal is to provide a scientific and precise framework which measures and quantifies the degree to which ICL of an LLM is Bayesian.

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 a principled interpretation of un-

certainty in LLMs (§2). (b) We derive actionable diagnostics with corresponding theory and test statistics of the martingale property for ICL. We also characterise the efficiency of ICL compared to standard Bayesian inference (§3). (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 a Bayesian Learning System? A Martingale Perspective

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.

2.1. The Martingale Property

We begin by defining the *martingale property*.

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

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

Eq. (1) states that $\{Z_i\} \sim p_M$ are *conditionally identically distributed* (Berti et al., 2004). As we will explain in §2.3, this renders distributions $\{p_M(Z_{n+1} = \cdot|Z_{1:n})\}$ to form a martingale, hence the name 'martingale property'.

It follows from Eq. (1) that predictive distributions of the form $p_M(Y_{n+k}|X_{n+k}, Z_{1:n})$ satisfy a similar identity:

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

for all integers $n, k > 0$, realisations $\{z_{1:n}, y\}$, and (almost every) realisation x measured by $p_M(X_{n+1}|Z_{1:n} = z_{1:n})$. In Eq. (2) the martingale property renders a model's predictions invariant to imputations of missing samples from the population (on average). Note that Eqs. (1) and (2) are equivalent in the unconditional case ($x_i = \varnothing$), which we consider in the majority of our experiments in §4.

2.2. The Martingale Property is Necessary for Unambiguous Predictions under Exchangeable Data

To understand the intuition behind the seemingly technical notion of the martingale property, consider two scenarios for ICL, illustrated in Fig. 2. In both scenarios,

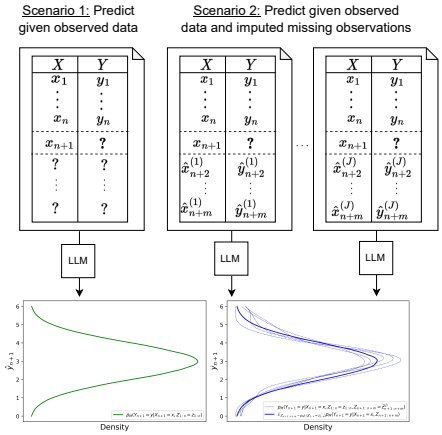


Figure 2: The *martingale property*, a fundamental requirement of a Bayesian learning system, requires *invariance with respect to missing samples from a population.*

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 J times 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 §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 the predictive distribution 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 all information about the distribution of X and Y presented to the model (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 core idea of the martingale property. If the predictive distribution for y_{n+1} changes on average, the model is 'creating new knowledge' when there is none: it is 'hallucinating'. In preview of our experimental results in §4, we observe this violation of the martingale property in state-of-the-art LLM families. We call this phenomenon *introspective hallucinations*: by querying itself, the model changes its predictions (on average), which as we

shall see in §2.4 violates how Bayesian systems learn.

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 §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,

$$p_M(Z_1, \ldots, Z_\ell) = p_M(Z_{\sigma(1)}, \ldots, Z_{\sigma(\ell)}). \quad (3)$$

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 $Z_{n+1,\ldots}|Z_{1:n}$. In the standard ICL setup, it is natural to assume that the sequence of example tuples in the ICL dataset, which is part of the prompt, is i.i.d. and thus exchangeable, and many influential works make this assumption (often without stating it explicitly) (Xie et al., 2021; Wang et al., 2023; Jiang, 2023). To understand the importance of this assumption further, consider the RCT example in §1, where $\{Z_{1:100}\}$ are (by experimental design) exchangeable. A model p_M should hence satisfy

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

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 with the 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).

2.3. The Martingale Property Enables a Principled Notion of Uncertainty

The second desirable and important consequence of the martingale property is that it establishes a principled notion of uncertainty in the model's predicitive distribution. More specifically, it allows us to decompose this uncertainty, enabling us to study and interpret the uncertainty of a model.

To simplify the exposition, suppose the variables Z_i are discrete and have $A < \infty$ realisations (both standard in LLMs) [1], 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 $\{Z_n\}$. Under boundedness conditions always satisfied in the above case, 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,

$$p_M(Z_{n+1}=\cdot\,|\,Z_{1:n}) = \int p(\theta_\infty|Z_{1:n}) p_{\theta_\infty}(Z=\cdot) d\theta_\infty. \quad (4)$$

Note the similarity of Eq. (4) with Bayesian inference: the Bayesian posterior predictive distribution has the form

$$p_M(Z_{n+1} = \cdot \mid Z_{1:n}) = \int p(\theta|Z_{1:n}) p(Z = \cdot|\theta) d\theta. \quad (5)$$

The random vector θ_∞ plays the same role as the parameter θ in a Bayesian model, as both determine a predictive distribution ($p_{\theta_\infty}(Z)$ or $p(Z|\theta)$). They are thus interchangeable for prediction purposes. Moreover, if p_M is defined through Bayesian inference over θ, p_{θ_∞} will define the same distribution over Z as $p(\cdot|\theta)$ (see App. B.1). Therefore we refer to the distribution $\theta_\infty|Z_{1:n}$ as the *martingale posterior*.

Eq. (4) shows that the variation or *uncertainty* in the predictive distribution $p_M(Z_{n+1} = \cdot \mid Z_{1:n})$ has two sources:

1. **epistemic uncertainty**, which is about the latent θ_∞ and can be reduced if more data is available; and
2. **aleatoric uncertainty**, which is irreducible given a fixed set of features even if infinite samples are observed and all aspects of the data generating process, namely the latent θ_∞, are known.

The close connection between Eqs. (4) and (5) shows that this decomposition of uncertainty is established by the same foundations as in Bayesian inference. This is particularly relevant for LLMs which lack clearly stated, interpretable and verifiable assumptions (such as a prespecified statistical model), rendering their predictive distribution a 'black-box'.

Importantly, we can construct the martingale posterior solely using path samples from p_M: we can sample from $p(\theta_{n+k}|Z_{1:n})$ simply by sampling $Z_{n+1:n+k-1}|Z_{1:n}$ as $\lim_{k\to\infty}\theta_{n+k} = \theta_\infty$. Alternatively, we can also estimate parametric models on the path samples as proposed in Fong et al. (2021) (see App. B.1 for further details). This construction is an appealing tool for interpreting black-box models such as LLMs.

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

[1] We refer to App. B.1 for a review of the more general case.

active learning, reinforcement learning or healthcare; particularly in regions of the input space where the uncertainty is high. In §3.3 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 $\theta_\infty = (\theta_{\infty,0}, \theta_{\infty,1}) \in \mathbb{R}^2$, and $p_{\theta_\infty} = \text{Bern}(\theta_{\infty,1})$. Thus, in both Eq. (4) and Eq. (5) 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 black-box model p_M to impute the missing samples $\{Z_{n+i}\}$ from a population —as in the RCT example in §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*.

2.4. 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. In this subsection, we want to further formalise this. We have already discussed the close connection between the martingale property, exchangeability (§2.2), and uncertainty (§2.3). We will now show that for ICL on i.i.d. data, exchangeability, for which the martingale property is a necessary condition, and Bayesian inference are closely connected, equivalent conditions.

ICL typically assumes i.i.d. observations $Z_{1:n}$, which is our primary focus in this work (see §2.2). Therefore, a correctly specified Bayesian model should produce marginal predictive distributions of the form

$$p_M(Z_{1:n}{=}z_{1:n}) \tag{6}$$
$$= \int p_M(Z_1{=}z_1, \ldots, Z_n{=}z_n|\theta)\pi(\theta)d\theta$$
$$= \int \left(\prod_{i=1}^{n} p_M(Z{=}z_i|\theta)\right)\pi(\theta)d\theta, \quad \forall n \in \mathbb{N}. \tag{7}$$

Here, θ denotes the parameter of a Bayesian model, π denotes the prior measure and $p_M(Z = \cdot \mid \theta)$ denotes the likelihood. From the factorisation over the data dimension n in (7), we can see that it is invariant with respect to permutations of $z_{1:n}$, and thus the left-hand side of the equation in (6) is invariant, too. It then follows that $\{Z_i\} \sim p_M$ satisfies Eq. (3), and thus $\{Z_i\}$ are exchangeable. 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 a representation in the form of Eq. (7). It then follows that the predictive distribution $p_M(Z_{n+1}|Z_{1:n})$ has the form of a Bayesian posterior predictive distribution,

$$p_M(Z_{n+1}|Z_{1:n}) = \int p_M(Z_{n+1}|\theta)\pi(\theta|Z_{1:n})d\theta,$$

and can thus be viewed as implicit Bayesian inference for the latent variable θ (Huszár, 2022). In conclusion, ICL on i.i.d. data corresponds to a Bayesian model that assumes (conditionally) i.i.d. observations *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 cannot be Bayesian.

3. Probing Bayesian Learning Systems through Martingales

In this section we introduce practical diagnostics to probe if LLMs match the behaviour of Bayesian learning systems.

3.1. 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. $\mathbb{E}(Z^l_{n'+1}|Z_{1:n}) = \mathbb{E}(Z^l_{n'+k}|Z_{1:n})$ 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. In early experiments, we

observed frequent poor approximations for long sampling paths (see Fig. 11 in the Appendix). 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 §3.2, in order to check our hypothesis experimentally. In App. B.3 we derive the order of 'acceptable violations' for the test statistics we will introduce.

3.2. Diagnostics for the Martingale Property

As we showed in §2.4, 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 $\{z^{(j)}_{n+1:n+m} \sim p_M(Z_{n+1:n+m}|Z_{1:n} = z_{1:n})\}_{j=1}^{J}$ autoregressively 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. *A sequence $\{Z_{n+1:n+m}\} \sim p_M(\cdot|Z_{1:n})$ satisfies the martingale property if and only if the following holds: for all $n', k \in \mathbb{N}$ and integrable functions g, h:*

$$\mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))h(Z_{n+1:n'})|Z_{1:n}) = 0. \quad (8)$$

We now state two implications of Proposition 1, our two diagnostics of the martingale property, which we will check experimentally in §4.

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). 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. (8) 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 our choices for h and g with an example.

Example 2. Suppose p_M is a Bayesian learning system over a latent parameter θ (see Eq. (7)), and the respective 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 the indices (n', k). These two important aspects of the posterior are hence consistently expressed by the model. The example is especially relevant as it covers Bernoulli ($p(Z|\theta) = \mathrm{Bern}(\theta)$) and Gaussian data, which will be our main focus in the experiments.

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 model, which is readily available in synthetic settings, through the same sampling procedure. We will refer to these comparisons as 'checks' of the martingale property. If $T_{1,g}$ and $T_{2,k}$ lie outside the confidence interval, properties (i) and (ii) and hence the martingale property are violated.

3.3. Diagnostics for Epistemic Uncertainty

As discussed in §2.3, 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 1. 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,

$$\begin{aligned}&\mathbb{E}_{\theta_0\sim\pi, z_{1:n}\sim\pi(z|\theta_0)}\mathbb{E}_{\theta\sim\pi(\theta|z_{1:n})}\|\theta-\bar{\theta}_n\|^2\\&=\mathbb{E}_{\theta_0\sim\pi, z_{1:n}\sim\pi(z|\theta_0)}\|\theta_0-\bar{\theta}_n\|^2.\end{aligned}\tag{9}$$

The left-hand side in Eq. (9) 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 1 states that *epistemic uncertainty provides a quantification for the average-case estimation error*. Note that Eq. (9) 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 §2.3, 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 §3.1) 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*

$$T_3 = Q_{0.75}(\{\hat{\theta}_m^{(j)}\}_{j=1}^J) - Q_{0.25}(\{\hat{\theta}_m^{(j)}\}_{j=1}^J),\tag{10}$$

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.2. 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.

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 $T_{1,g}, T_{2,k}, T_3$. We provide our code base on `https://github.com/meta-inf/bayes_icl`.

4.1. Experiment Setup

We consider three types of synthetic datasets $z_{1:n}$:

- **Bernoulli**: $Z_i \sim \mathrm{Bern}(\theta)$, 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, $X_i \sim \mathrm{Bern}(0.5), Y_i|X_i \sim \mathrm{Bern}(0.3 + 0.4X_i)$.

On purpose, we reduce our experimental setup to these minimum viable test beds where the ground-truth latent parameters are known, stripping away the convoluted latent complexity of in-the-wild NLP data. We use the following LLMs: `llama-2-7B` with 7B parameters (Touvron et al., 2023), `mistral-7B` (Jiang et al., 2023), `gpt-3` (Brown et al., 2020) with 2.7B and 170B parameters, `gpt-3.5`, and `gpt-4` (OpenAI, 2023)[2].

In all experiments we compute test statistics on LLM samples, and compare their behaviour with the same statistics evaluated on samples from a reference Bayesian model. More specifically, in §4.2 we compare the statistics obtained from LLMs with the bootstrap confidence intervals (CIs) derived from the reference Bayesian model. A deviation will thus indicate that the LLM is unlikely to be a good approximation of the reference Bayesian model. More importantly, when n becomes moderately large, the Bernsten von-Mises theorem (Van der Vaart, 2000) applies: the deviations then imply that the LLM is highly likely deviating from all *reasonable Bayesian models*, namely those satisfying the regularity conditions of the theorem. This is because the theorem guarantees that the test statistics derived from all such models have asymptotically[3] equivalent distributions.

We refer to App. C.1 for additional experimental details, such as the prompt format, tokenization, and computational requirements, as well as additional experimental results.

[2]We only use `gpt-4` in a subset of experiments (Fig. 3, Fig. 5 in the text) due to API and resource limitations (App. C.1).

[3]We note that the asymptotic equivalence results are relevant in our setting. As a concrete example, in the setting of Fig. 3 (a), the CIs obtained by using $\mathrm{Beta}(1, 11)$ and $\mathrm{Beta}(1, 1)$ as the reference model are practically indistinguishable; the difference is on the order of 10^{-4}.

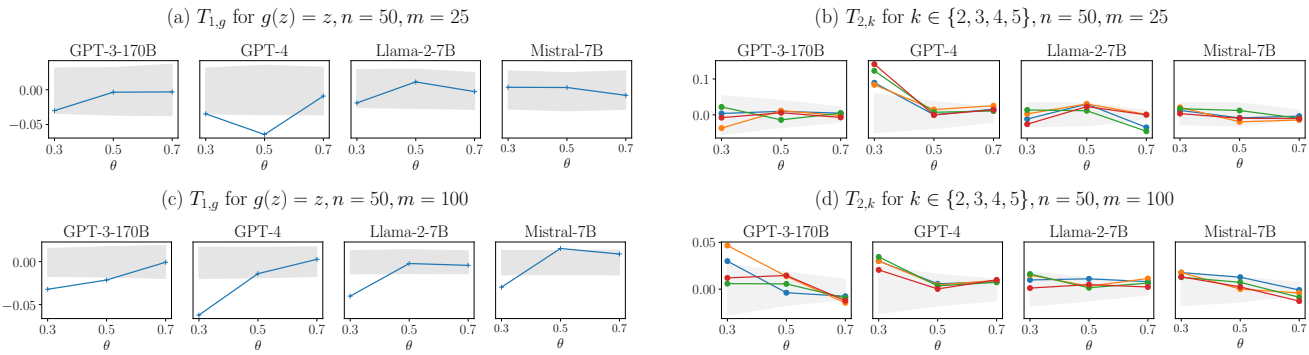


Figure 3: Checking the martingale property on Bernoulli experiments. Each data point represents a test statistic (y-axis) evaluated for an LLM, as derived in §3.2. Subplot and x-axis correspond to choices of Bernoulli probabilities and LLMs. Shade indicates the 95% confidence interval from a reference Bayesian model.

4.2. Checking the Martingale Property

We first check if state-of-the-art LLMs satisfy the martingale property. As we discussed in §2, this is a necessary condition for an exchangeable Bayesian ICL system.

Bernoulli experiment. Fig. 3 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 §3.2 and §4.1 above, 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 CIs (of high confidence, see scale of y-axis) obtained from a reference Bayesian model. Here we define the reference model using a Bernoulli likelihood and a non-informative $\text{Beta}(1, 1)$ prior.

For short sample paths of length $m = n/2$ (subplots (a) and (b)), most LLMs lead to test statistics that are generally within the respective 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)), more frequent deviations from the CIs are observed. For brevity, full results for other choices of n and LLMs are deferred to App. C.2. The findings are generally consistent across all choices of n. We also observe `gpt-3.5` to perform better than `gpt-4` but worse than `gpt-3-170b`. As we discuss in App. C.2 the latter observation may be explained by the fact that `gpt-3.5` and `gpt-4` have undergone instruction tuning (Ouyang et al., 2022). 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.

Gaussian experiment. In Fig. 4 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 $\mathcal{N}(0, 100)$. As we can see, all models except `gpt-3.5` demonstrate clear deviation from the martingale property. Additional results for `gpt-3.5` in App. C present our diagnostics with other choices of (n, m, θ), demonstrating a deviation from the predictive distribution of the reference Bayesian posterior. In conclusion, the presented evidence on the Gaussian experiment falsifies our hypothesis of Bayesian behaviour with the tested LLMs.

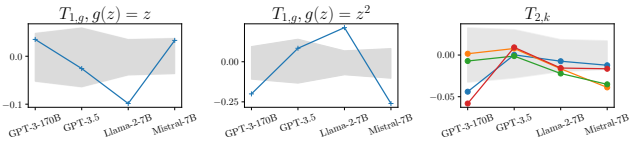
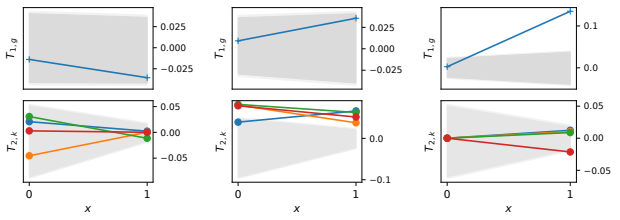


Figure 4: Checking the martingale property on Gaussian experiments. We present runs with $\theta = -1, n = 100, m = 50$ from different LLMs (x-axis) with test functions $g(z) = z$ and $g(z) = z^2$. See Fig. 3 for further details.

Synthetic natural language experiment. In Fig. 5 we present our results for the natural language experiment with $n = 80, m = 40, g(z) = z$ using the GPT models. Here, we compute the test statistics on samples separated by the Bernoulli-distributed value of X_i (see App. C.1 for details). As we can see, both `gpt-3.5` and `gpt-4` demonstrate deviation from a reference Bayesian posterior. This provides further evidence of violations of the martingale property in settings where natural language (instead of numbers) is used.

4.3. Checking Epistemic Uncertainty of LLMs

In this subsection we analyse the scaling behaviour of an LLM's uncertainty. In Fig. 6 we measure T_3 (y-axis on a log-scale) and compare the approximate martingale posterior of



(a) GPT-3-170B (b) GPT-3.5 (c) GPT-4

Figure 5: 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. 3 for further details.

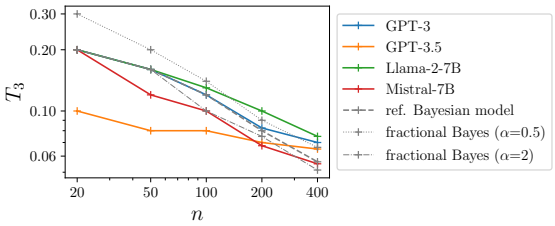


Figure 6: Scaling of epistemic uncertainty on the Bernoulli experiment: the test statistic T_3 (§3.3) computed on LLMs, compared with Bayesian and fractional Bayesian models.

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 `llama-2-7b` and `gpt-3.5` is slower than that of a Bayesian model, which suggests inefficiency as discussed in §3.3. Furthermore, `gpt-3.5` demonstrates over-confidence in the small-sample regime. The scaling of `gpt-3-170b` and `mistral-7b` are closer to the Bayesian model, even though not exactly matching the latter. This finding is interesting as on the Bernoulli experiments, `gpt-3-170b` and `mistral-7b` also demonstrate the best adherence to the martingale property.

5. Conclusion

In this work we stated the martingale property as a fundamental requirement of a Bayesian learning system for exchangeable data, and discussed its desirable consequences if satisfied by an LLM. Based on this property we derived three different diagnostics that allowed us to check whether LLMs adhere to the Bayesian principle on synthetic in-context learning tasks. We presented stark evidence that state-of-the-art LLMs violate the martingale property, and hence falsified the hypthesis that ICL in LLMs is Bayesian.

Our investigation is particularly relevant to a recent line of work that investigates LLM-based ICL for tabular data modelling: for prediction on noisy tabular datasets (Manikandan et al., 2023; Yan et al., 2024), the martingale property would enable us to diagnose the predictive uncertainty; and for synthetic data generation (Borisov et al., 2022; Hämäläinen et al., 2023; Veselovsky et al., 2023), it is vital to ensuring valid inference based on imputations of missing data (§2.2). It is thus of practical interest to develop models that better adhere to the martingale property.

The primary limitation of our work is the (intentional) restriction to small-scale, synthetic datasets, which are different from common NLP applications. We note that while our diagnostics are designed for synthetic problems, they reflect a broader principle: Bayesian epistemic uncertainty can be extracted from black-box models by examining the correlation structure in sequential predictions. This is clearly shown by the variance estimator in Example 2, and by the fact that MLE on sampled paths approximates the Bayesian posterior (§3.3). Future work could investigate generalisations of this approach.

More broadly, the RCT example in §1 can arguably be viewed as the simplest type of decision task involving multi-step reasoning, as the right decision (here based on an average treatment effect) is only naturally determined after imputing all missing samples. Thus, it would be interesting to investigate analogies to the hallucination behaviour we have identified for ICL in more complex reasoning tasks such as those involving chain-of-thought prompting (Wei et al., 2022). Lastly, it may be worth to consider fine-tuning objectives to achieve an idealised Bayesian behaviour with a model after pretraining, but before deployment.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here. We refer to App. E for further discussion.

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

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Berti, P., Pratelli, L., and Rigo, P. 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.

Bhattacharya, A., Pati, D., and Yang, Y. Bayesian fractional posteriors. *Annals of Statistics*, 47(1):39–66, 2019.

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

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

De Finetti, B. 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.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

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

Fong, E., Holmes, C., and Walker, S. G. Martingale posterior distributions. *arXiv preprint arXiv:2103.15671*, 2021.

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

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

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820*, 2023.

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

Hämäläinen, P., Tavast, M., and Kunnari, A. 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.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

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

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

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Kalai, A. T. and Vempala, S. S. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2023.

Kallenberg, O. *Foundations of modern probability*, volume 2. Springer, 1997.

Li, Z., Zhu, H., Lu, Z., and Yin, M. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*, 2023.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Manikandan, H., Jiang, Y., and Kolter, J. Z. Language models are weak learners, June 2023. URL http://arxiv.org/abs/2306.14101. arXiv:2306.14101 [cs].

Mei, Y., Song, S., Fang, C., Yang, H., Fang, J., and Long, J. Capturing semantics for imputation with pre-trained language models. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 61–72. IEEE, 2021.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Panwar, M., Ahuja, K., and Goyal, N. In-context learning through the bayesian prism. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HX5ujdsSon.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

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

Raventós, A., Paul, M., Chen, F., and Ganguli, S. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *arXiv preprint arXiv:2306.15063*, 2023.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. Model dementia: Generated data makes models forget. *arXiv e-prints*, pp. arXiv–2305, 2023.

Singh, A. K., Chan, S. C., Moskovitz, T., Grant, E., Saxe, A. M., and Hill, F. The transient nature of emergent in-context learning in transformers. *arXiv preprint arXiv:2311.08360*, 2023.

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

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., 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.

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

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

Veselovsky, V., Ribeiro, M. H., Arora, A., Josifoski, M., Anderson, A., and West, R. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*, 2023.

Wang, X., Zhu, W., Saxon, M., Steyvers, M., and Wang, W. Y. Large language models are latent variable models: Explaining and finding good demonstrations for incontext learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

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.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. 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.

Xiao, Y., Liang, P. P., Bhatt, U., Neiswanger, W., Salakhutdinov, R., and Morency, L.-P. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*, 2022.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Yan, J., Zheng, B., Xu, H., Zhu, Y., Chen, D., Sun, J., Wu, J., and Chen, J. Making pre-trained language models great on tabular prediction. In *International Conference on Learning Representations*, 2024.

Ye, N., Yang, H., Siah, A., and Namkoong, H. Pre-training and in-context learning IS bayesian inference a la de finetti. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL https://openreview.net/forum?id=ttupfosvgx.

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

Zhang, L., McCoy, R. T., Sumers, T. R., Zhu, J.-Q., and Griffiths, T. L. Deep de finetti: Recovering topic distributions from large language models. *arXiv preprint arXiv:2312.14226*, 2023a.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023b.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Appendix for *Is In-Context Learning in Large Language Models Bayesian? A Martingale Perspective*

A. Proofs of Theoretical Statements in the Main Text

Fact 2. Any exchangeable random sequence $\{Z_i\}$ must be conditionally identically distributed.

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

Proposition 1. *A sequence* $\{Z_{n+1:n+m}\} \sim p_M(\cdot|Z_{1:n})$ *satisfies the martingale property if and only if the following holds: for all* $n', k \in \mathbb{N}$ *and integrable functions* g, h*:*

$$\mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))h(Z_{n+1:n'})|Z_{1:n}) = 0. \quad (8)$$

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 $0 = \mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))h(Z_{n+1:n'})|Z_{1:n})$.

The equivalence between (i) and (ii) is trivial. We have (ii) $\Rightarrow$ (iii) because $\mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))h(Z_{n+1:n'}) \mid Z_{1:n}) = \mathbb{E}(\mathbb{E}(g(Z_{n'+k}) - g(Z_{n'+1}) \mid Z_{1:n'})h(Z_{n+1:n'}) \mid Z_{1:n}) \overset{(ii)}{=} 0$. To show (iii) $\Rightarrow$ (ii), for any $\sigma(Z_{n+1:n'})$-measurable set A let $h := \mathbb{1}_A$ be the respective indicator function, so that $\mathbb{E}((g(Z_{n'+k}) - g(Z_{n'+1}))\mathbb{1}_A \mid Z_{1:n}) \overset{(iii)}{=} 0 = \mathbb{E}(0 \cdot \mathbb{1}_A \mid Z_{1:n})$. Since this holds for all A, it follows by the definition of conditional expectation (Kallenberg, 1997) that $\mathbb{E}(g(Z_{n'+k}) - g(Z_{n'+1}) \mid Z_{1:n'}) = 0$ a.s.. □

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 (8). (ii) follows by setting $g(z) = z, h(z_{n+1:n'}) = z_{n'}$. □

Fact 1. 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,

$$\begin{aligned}&\mathbb{E}_{\theta_0\sim\pi, z_{1:n}\sim\pi(z|\theta_0)}\mathbb{E}_{\theta\sim\pi(\theta|z_{1:n})}\|\theta - \bar{\theta}_n\|^2 \\ &= \mathbb{E}_{\theta_0\sim\pi, z_{1:n}\sim\pi(z|\theta_0)}\|\theta_0 - \bar{\theta}_n\|^2. \quad (9)\end{aligned}$$

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. □

B. Further Discussion of Theory and Methodology

B.1. Additional Background on Martingale Posteriors

In §2.3 we discussed the construction of martingale posteriors in the finite-support case. Here, 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 in Eq. (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.

Nonetheless, for tractability and comparability to Bayesian parametric posteriors, it is useful to consider the following alternative, 'model-based' procedure:

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 $Z_{1:n}$.

We repeat this procedure to obtain multiple samples $\hat{\theta}_m$ from the martingale posterior in order to approximate its distribution (see Fig. 1 [Centre]). 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 Eq. (1) leads to the methodology in Fong et al. (2021).

We adopted this 'model-based' approach in §3.3 and for computing the approximate martingale posterior in Fig. 1

[centre]. 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 (and not $\gg n$), 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 the following. This is important in our experiments where we find the LLMs (at best) follow the martingale property within a horizon of $m = \Theta(n)$.

B.2. 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 §3.3, or above) 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.1 equals the exact martingale posterior. Combining the above representation and the fact that $\hat{\theta}_m$ is a function of $Z_{n+1:n+m}$ leads 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, 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 $m = \Theta(n)$, 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.

Lastly, 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.

B.3. Acceptable Approximation Errors of Properties (i) and (ii) in Corollary 1

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

As discussed in this example, the equalities in Corollary 1 guarantee the expressions for posterior mean and covariance for the parameter θ to 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

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

Therefore, it can have 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 of 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

$$T_{1,g} := \frac{2}{Jm}\sum_{j=1}^{J}\sum_{i=1}^{m/2}(g(z_{n+i}^{(j)}) - g(z_{n+i+m/2}^{(j)})), \qquad (11)$$

$$T_{2,k} := \frac{1}{Jm}\sum_{j=1}^{J}\sum_{i=1}^{m-k-1}(z_{n+i+1}^{(j)} - z_{n+i+k}^{(j)})z_{n+i}^{(j)}. \qquad (12)$$

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 (§3.1), we remove sample paths with anomalous mean absolute values using the standard 1.5×IQR rule.

We compare the observed value of the statistics above evaluated on LLMs with bootstrap confidence intervals computed using a reference Bayesian model (§4.1). For the latter, we draw $K = 300$ sets of completions $\{\{z_{bs,n+i}^{(j,k)} : 1 \leq i \leq m, 1 \leq j \leq J\} : 1 \leq k \leq K\}$ from the predictive distribution of the reference Bayesian model, which provides K samples for the test statistics, and compute two-sided confidence intervals using the respective quantiles.

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 the experiments in §4.2 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 experiments in §4.3, 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.

For a proper test of the martingale property, it is vital that the model cannot distinguish between the ICL training data $Z_{1:n}$ and its own generations $\{Z_{n+i}\}$. This is trivially true if the LLM takes free-form text as inputs without additional annotation, as with `llama-2-7b`, `mistral-7b`, and `gpt-3.5` accessed through the `Completion` API from OpenAI. However, the `gpt-4` model is only accessible through a different API (`ChatCompletion`) which includes annotation for user input and model generation in the prompt. To ensure a proper implementation of the checks, we hence call the API m times in generating each path sample. In each iteration we sample a single data point, and then append it to the user input part of the prompt. This is far less cost-efficient than our use of `gpt-3.5`. Therefore, we only include `gpt-4` for the Bernoulli experiment with $n \leq 50$, and the natural language experiment.

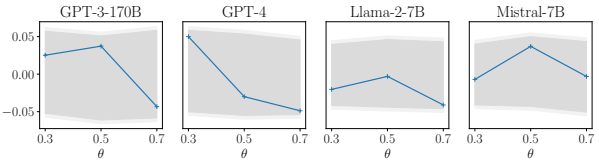
We discuss prompt design and format in detail below. Here we emphasise that across all tasks, the prompt always includes sufficient information about the true likelihood.

Prompt design and format. We use the following prompt format `<instruction> <observed data> <sampled data>`. `<instruction>` describes the distribution (i.e. true likelihood) 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 in the LLMs' behaviour. Each sample is delineated by a separator (e.g. ';').

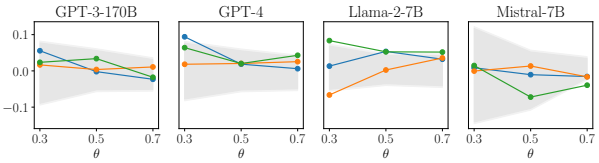
We present exemplary prompts for each dataset below:

- 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 $n = 2$ and $m = 3$: *"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"*.
- 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 . . . "*

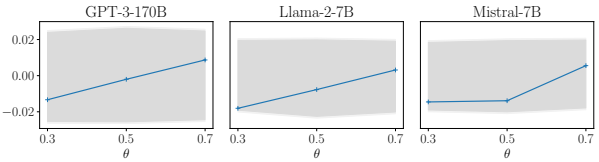
Other work represents both `int` and `float` numbers as a space-separated string of digits with fixed precision, where each number is separated by a semi-colon. This guarantees a per-digit tokenisation that was observed to be beneficial in the context of time series forecasting and further minimises 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.



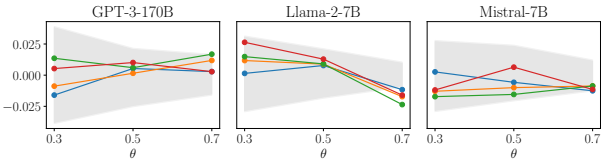
(a) $T_{1,g}$ for $g(z) = z, n = 20, m = 10$



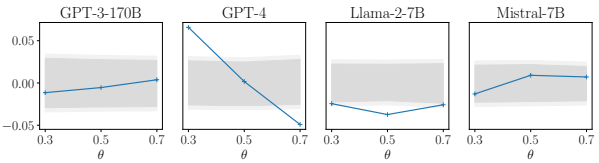
(b) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 20, m = 10$



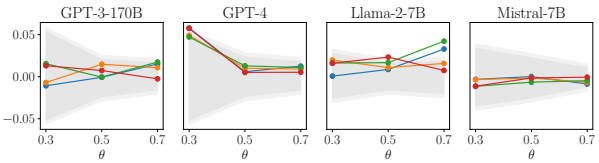
(c) $T_{1,g}$ for $g(z) = z, n = 100, m = 50$



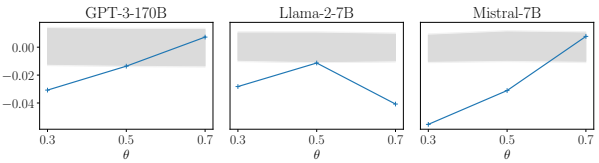
(d) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 100, m = 50$



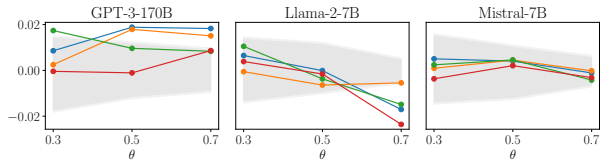
(e) $T_{1,g}$ for $g(z) = z, n = 20, m = 40$



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

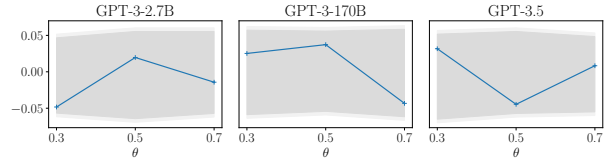


(g) $T_{1,g}$ for $g(z) = z, n = 100, m = 200$

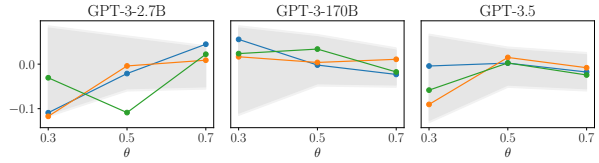


(h) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 100, m = 200$

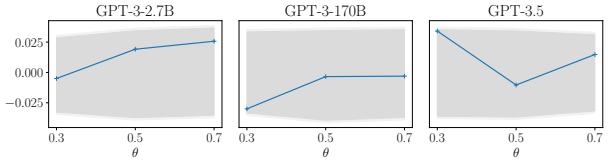
Figure 7: Checking the martingale property: results for the Bernoulli experiments for all choices of (n, m) in the setting of Fig. 3. Note that we drop `gpt-4` for $n = 100$ due to API limitations (as discussed in App. C.1).



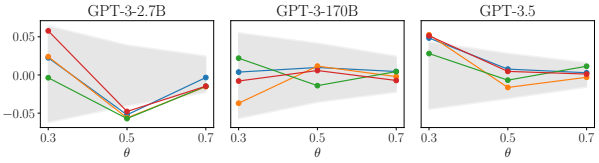
(a) $T_{1,g}$ for $g(z) = z, n = 20, m = 10$



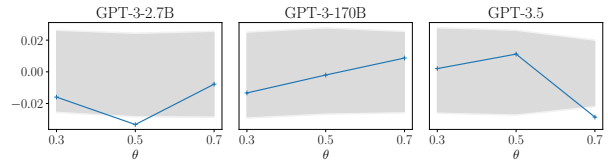
(b) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 20, m = 10$



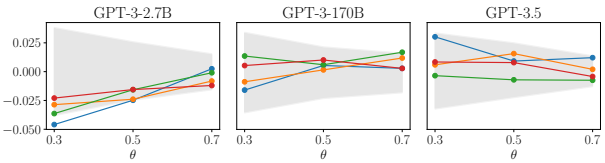
(c) $T_{1,g}$ for $g(z) = z, n = 20, m = 10$



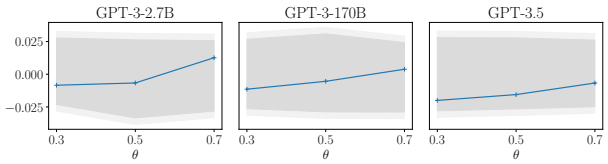
(d) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 20, m = 10$



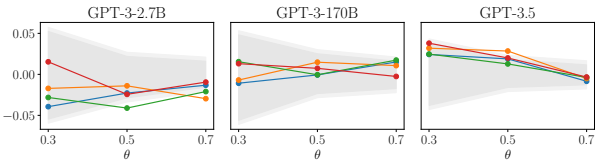
(e) $T_{1,g}$ for $g(z) = z, n = 100, m = 50$



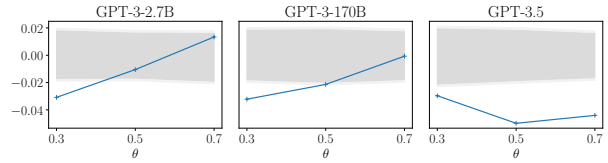
(f) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 100, m = 50$



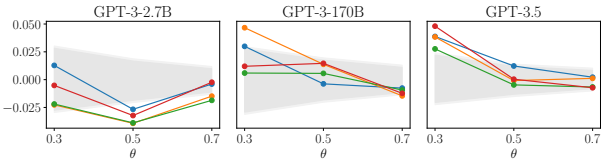
(g) $T_{1,g}$ for $g(z) = z, n = 20, m = 40$



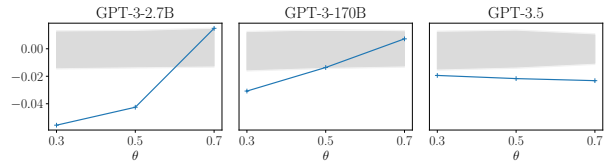
(h) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 20, m = 40$



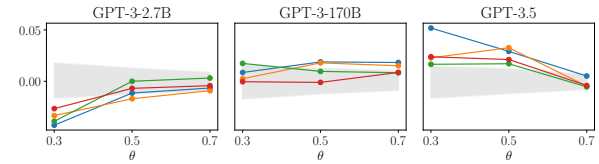
(i) $T_{1,g}$ for $g(z) = z, n = 50, m = 100$



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



(k) $T_{1,g}$ for $g(z) = z, n = 100, m = 200$



(l) $T_{2,k}$ for $k \in \{2, 3, 4, 5\}, n = 100, m = 200$

Figure 8: Checking the martingale property: results for `gpt-3-2.7b` and `gpt-3.5` in the setting of Fig. 3.

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 and Discussion

Full results: Bernoulli experiments. Figs. 7 and 8 report the full results for the Bernoulli experiment in the setting of Fig. 3 ($m \in \{n/2, 2n\}$), where we also visualise the $o(1/n)$ 'acceptable deviation' (§B.3) using a light shade with width $0.1/n$. Consistent with the results in Fig. 3, for all models except the least capable `gpt-3-2.7b`, the martingale property is generally satisfied in the short-horizon scheme ($m = n/2$), but increasingly violated as we move to $m = 2n$.

The results for `gpt-3-2.7b` provide a sanity check of our experiment setup: Its unsatisfactory performance shows that our tasks require nontrivial ICL capabilities which are known to be absent in `gpt-3-2.7b` (Brown et al., 2020). As another sanity check we provide the results for $m = 10n$ in Fig. 11, where we drop all GPT models 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 §2.2.

Full results: Gaussian experiments. We report additional results for the Gaussian experiment in Fig. 9 ($\theta = 0$) and Fig. 10 ($\theta = -1$). As we can see, 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. 4 we find the sample variance to be $0.711 < 1$ (95% CI: $[0.680, 0.742]$).

Additional results: Scaling of epistemic uncertainty. To avoid clutter, in Fig. 6 we have plotted the sample median of the test statistic T_3 from various models, and in that aspect `gpt-3-170b` appears to be close to the reference Bayesian model when n is smaller. Here we note that a deviation becomes more evident if we compare the individual samples of T_3 (obtained from independent runs) against bootstrap CIs from the reference Bayesian model, as shown in Fig. 12.

Additional results with fine-tuned models. Some previous works (Zhang et al., 2023b; Jiang, 2023) studied ICL under the assumption that the LLM has been perfectly pretrained on the ICL test distributions. While such assumptions are somewhat unrealistic, it may still be interesting to investigate whether finetuning on datasets that are similar to the ICL test distribution could lead to a closer-to-Bayesian behaviour for ICL. To this end we finetune `gpt-3-2.7b` models on Bernoulli and synthetic NLP datasets with randomly sampled parameters, and repeat the checks in §3.2 on the finetuned models.[4] The results are visualised in Fig. 13. We can see that the finetuned models may indeed demonstrate a better adherence to the martingale property, but they do not always pass the checks.

Comparison of LLMs with and without instruction tuning. Among all LLMs evaluated, `gpt-3.5` and `gpt-4` generally demonstrate the worst performance in our evaluations, and incidentally they are the only LLMs that have undergone instruction tuning. The comparison between `gpt-3-170b` and `gpt-3.5` (see Fig. 6 and Fig. 8) is particularly interesting since the two models are generally similar, with a main difference being the presence of instruction tuning. These observations seem to suggest that instruction tuning may have exacerbated the non-Bayesian ICL behaviour. Such an explanation would be broadly consistent with the previous findings that instruction tuning generally causes the LLM to produce less calibrated uncertainty estimates (OpenAI, 2023; Gruver et al., 2023; Kalai & Vempala, 2023).

Could the LLMs correspond to 'unreasonable Bayesian models'? As discussed in the main text, our findings suggest that the behaviours of `gpt-3.5` and `llama-2-7b` are highly unlikely to correspond to any 'reasonable' Bayesian models in the Bernstein von-Mises sense. Gener-

[4] We use OpenAI's finetuning service which determines optimisation hyperparameters by validation loss. For the Bernoulli dataset, we sampled 10^4 sequence for training, each with an expected length of 75; the true parameter θ is sampled from the uniform distribution on $[0, 1]$. For the NLP dataset, we sampled 5000 sequences for training, each with an expected length of 85; the two Bernoulli parameters that determine the prompt distribution are sampled from a $\mathrm{Beta}(0.5, 0.5)$ distribution.

ally speaking, to conduct any statistical test with a reasonable level of power it is necessary to impose some regularity restrictions on the null hypothesis to be tested. Moreover, it could be similarly concerning if the LLMs correspond to any 'unreasonable' Bayesian model that does not satisfy the regularity conditions in the Bernstein von-Mises theorem. Nonetheless, in the setting above we can also provide some informal discussion on why the LLMs are unlikely to be 'unreasonable' Bayesian models (e.g., one with an approximately degenerate prior), by comparing the results across different choices of n. Specifically, for `gpt-3.5`, its small-sample behaviour in Fig. 6 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 some form of Bayesian Inference. The hypothesis is likewise studied in Jiang (2023); Wang et al. (2023) and the concurrent work of Ye et al. (2024). It is also covered by Zhang et al. (2023b) if we restrict to exchangeable demonstrations. Closely related are the works of Akyürek et al. (2022); Panwar et al. (2024) which demonstrate that high-capacity transformers pretrained with square loss may recover the Bayes predictor.

Xie et al. (2021) studied ICL in a setting where the the LLM is perfectly trained on a pretraining distribution defined by a Hidden Markov Model (HMM). Under this and further assumptions, they prove that the LLM must implicitly perform Bayesian inference to infer a latent concept of the prompt. Strictly speaking, their assumptions do not exactly match our hypothesis, because their Bayesian model employs a likelihood that is misspecified for ICL: it does not assume $\{Z_i\}$ is conditionally i.i.d. or exchangeable. However, their additional assumptions render the ICL behaviour similar to that of another Bayesian model that assumes conditionally i.i.d. observations: when considering Eq. (8-10) in their work, which imply that the log likelihood of their Bayesian model is well approximated by a Bayesian model assuming conditional i.i.d. observations. In this regard their analysis is connected to our hypothesis, as it applies to the Bayesian model we study. It is important to note that their assumptions have been crucial in their proof for sample efficiency. More broadly, for any ICL predictor to be sample efficient on exchangeable $\{Z_i\}$, it is perhaps reasonable to expect the predictor to (approximately) recognise the exchangeable nature of $\{Z_i\}$, where our hypothesis would apply.

We review the other works in brief in the following. Huszár (2022) dicussed high-level connections between Xie et al. (2021) and various notions of 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 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 sensititivity 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 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 which we study (e.g. the decomposition of uncertainty and the resulting identification of epistemic uncertainty). Furthermore, in contrast to the related work, which shuffles the input data $Z_{1:n}$, we analyse the effect of shuffling the imputed, generated sequence $Z_{n+1}, \ldots$, where we find non-exchangeable behaviour which deviates 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 machine learning task 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 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 a 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 characterises the behaviour of LLMs. We try to understand whether ICL in LLMs follows Bayesian principles. As we outlined in §2.3 this has important consequences for their potential use as trustworthy systems, which can be 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 §2.2, we caution against their use in exchangeable settings (e.g. i.i.d. in-context data) as their predictions can be rendered inconsistent.

The 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 our analysis, we believe this work provides ample pointers and reason for further investigation of these concerns, and shall point out and warn against (potentially intended) misuse of LLMs.

F. Code, Computational Resources, Datasets, Existing Assets Used

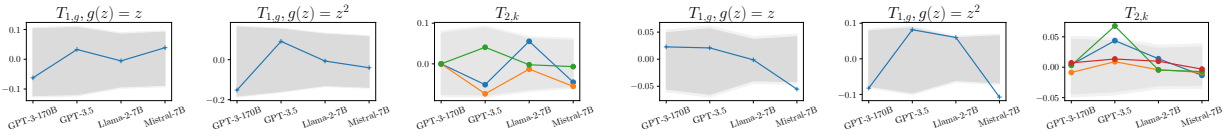
Code. We provide our code base on `https://github.com/meta-inf/bayes_icl` under MIT License, together with a `README.md` containing instructions on reproducing the key results in this paper.

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.

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`, `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.

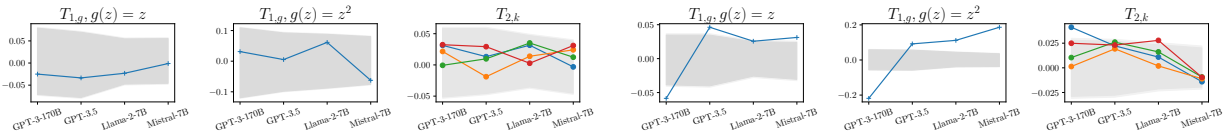
Existing assets used. Our work uses the following main software libraries and corresponding licenses: PyTorch (Paszke et al., 2019) (custom license), `numpy` (Harris et al., 2020) (BSD 3-Clause License), Weights&Biases (Biewald, 2020) (MIT License), Huggingface transformers library (Wolf et al., 2020) (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.

The five pretrained large language models we used (see §4) have the following licenses: `llama-2-7B` (Touvron et al., 2023) (custom license); `mistral-7B` (Jiang et al., 2023) (Apache 2.0 License); `gpt-3` (Brown et al., 2020), `gpt-3.5`, and `gpt-4` (OpenAI, 2023) (API; no code license).



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

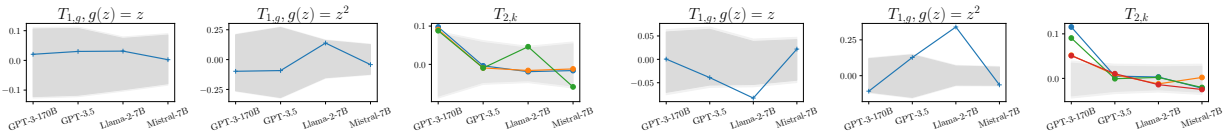
(b) $n = 20, m = 40$



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

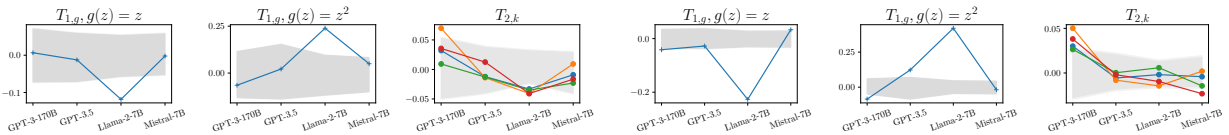
(d) $n = 50, m = 100$

Figure 9: Checking the martingale property: results for the Gaussian experiments with $\theta = 0$. See Fig. 4 for details.



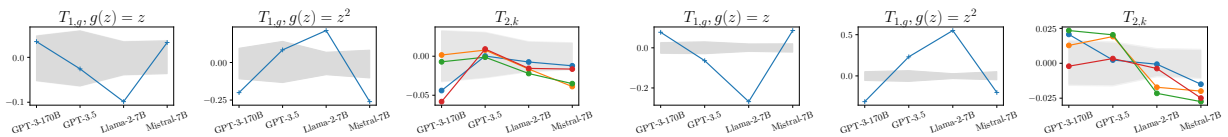
(a) $n = 20, m = 10$

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



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

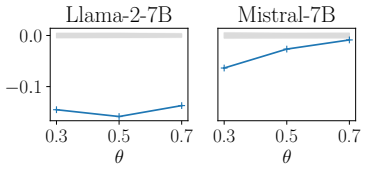
(d) $n = 50, m = 100$



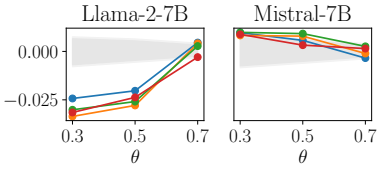
(e) $n = 100, m = 50$

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

Figure 10: Checking the martingale property: results for the Gaussian experiments with $\theta = -1$. See Fig. 4 for details.



(a) $T_{1,g}, g(z) = z$



(b) $T_{2,k}, k \in \{2, 3, 4, 5\}$

Figure 11: Checking the martingale property on Bernoulli experiments: additional result with $n = 100, m = 10n$. See Fig. 3 for details.

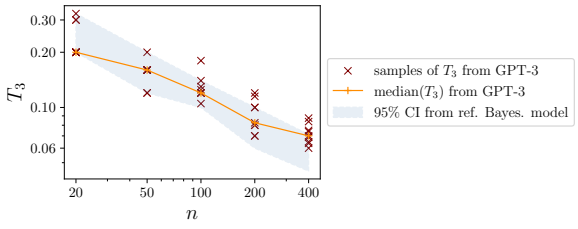
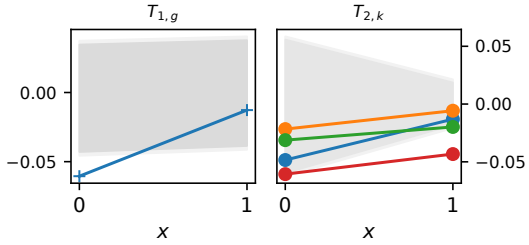
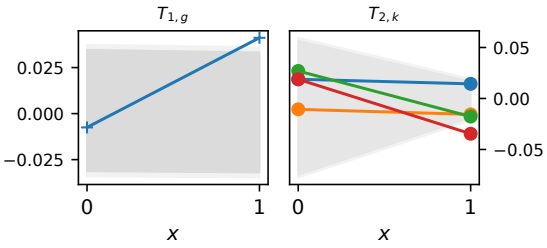


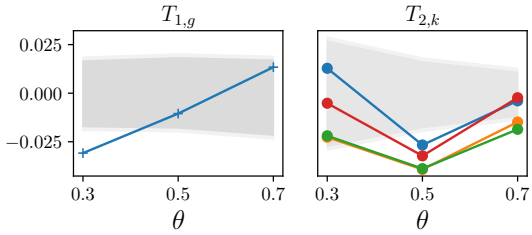
Figure 12: Scaling of epistemic uncertainty: samples of the test statistic T_3 evaluated on `gpt-3-170b`, compared with the 95% CI from the reference Bayesian model.



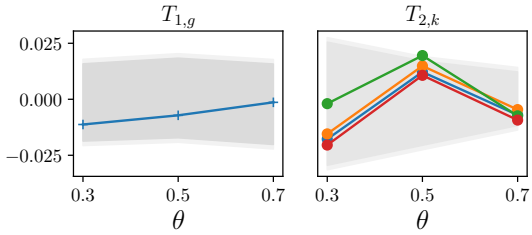
(a) `gpt-3-2.7b`, NLP, pretrained



(b) `gpt-3-2.7b`, NLP, finetuned



(c) `gpt-3-2.7b`, Bernoulli ($n = 50, m = 100$), pretrained



(d) `gpt-3-2.7b`, Bernoulli ($n = 50, m = 100$), finetuned

Figure 13: Checking the martingale property: comparison of `gpt-3` models before and after fine-tuning on the NLP (Fig. 5) and Bernoulli (Fig. 3) datasets.