

The sources define Chain-of-Thought (CoT) as a prompting technique that enhances reasoning in Large Language Models, and the Martingale Property as a fundamental requirement of Bayesian updating on exchangeable data.

Chain-of-Thought (CoT)

Chain-of-Thought (CoT) prompting is a powerful technique that has emerged to enhance reasoning in language models by **generating intermediate reasoning steps**. By explicitly showing these steps, models can solve complex problems that would otherwise be intractable.

Key Aspects of Chain-of-Thought:

- * **Mechanism:** It augments the input by adding intermediate "reasoning" tokens ($\$c_1, \dots, c_k\$$) between the original data tokens ($\$x_1, \dots, x_n\$$) and the target prediction ($\$x_{n+1}\$$).
- * **Purpose:** CoT allows models to tackle complex problems, leading to significant improvements in tasks like code generation and mathematical reasoning. It provides an "external scratch space" that complements the model's internal compressed knowledge, enabling it to tackle problems beyond its finite parametric capacity.
- * **Theoretical Necessity:** The sources propose an "incompleteness theorem" which proves that CoT is not merely a prompting trick but a **fundamental requirement for computational completeness** for transformers with finite parameter budgets. Transformers with parameters encodable in `H` bits cannot compute all predicates with Kolmogorov complexity $K(\pi) > H$ without such a chain of thought.
- * **Optimal Length:** The research derives a closed-form expression for the optimal number of intermediate reasoning tokens, $k^* = \Theta(\sqrt{n \log(1/\epsilon)})$, where `n` is the context length and `ε` is the target error tolerance. This provides a principled approach to balance the model's internal capacity with external reasoning.
- * **Computational Cost:** While powerful, CoT comes with substantial computational costs, increasing inference time and API charges by a factor of $(n+k)/n$, potentially making it 10-100 times more expensive per query. The optimization of CoT length is crucial for reducing these costs, potentially leading to 50-90% cost reduction and 5-10x faster inference.

The Martingale Property

The martingale property is a **fundamental mathematical consequence of Bayesian updating**, particularly when dealing with **exchangeable data**.

Key Aspects of the Martingale Property:

- * **Definition:** For exchangeable data, where the order of observations carries no information, Bayesian posterior predictive distributions must satisfy:
 $E[f(X_{n+1})|X_1, \dots, X_n] = E[f(X_{n+1})|X_{\pi(1)}, \dots, X_{\pi(n)}]$ for any permutation π and bounded function f . More broadly, for exchangeable sequences, Bayesian inference satisfies the martingale property: $E[h(X_{n+1})|X_1, \dots, X_n] = E[h(X_{n+1})|X_1, \dots, X_{n-1}]$ for any bounded function h . This implies that the expected prediction for the next

observation should not change if the order of past observations is permuted or if older observations are disregarded (given a new observation).

* **Challenge to Bayesian Interpretation:** Recent empirical findings, particularly by, demonstrated that **transformer-based language models systematically violate the martingale property**. This poses a serious challenge to the theoretical foundations of interpreting LLMs as performing Bayesian inference and their ability to provide calibrated uncertainty estimates.

* **Reason for Violation:** The research paper resolves this paradox by explaining that **positional encodings**, which are ubiquitous in transformer architectures, **fundamentally alter the information-theoretic structure of the learning problem**.

Positional encodings explicitly break the assumption of exchangeability by making the model's computations depend on the order of inputs.

* **Quantification of Violation:** The paper quantifies these martingale violations as being of order $\Theta(\log n/n)$, where n is the sequence length. This $\log n$ factor arises from the expected distance between random permutations, implying that permutations "further" in permutation space induce larger prediction differences.

* **Reconciliation:** Despite these violations, transformers are shown to be **"Bayesian in expectation, not in realization"**. They minimize expected conditional Kolmogorov complexity $E_{\pi}[K(X|\pi)]$ over permutations rather than the permutation-invariant complexity $K(X)$. This means they achieve optimal compression when averaged over orderings, even though their predictions for any specific ordering will necessarily deviate due to positional awareness. The violations decrease with sequence length, suggesting models become "more Bayesian" as context grows.

Transformers violate martingale properties primarily because of their **architectural reliance on positional encodings**, which fundamentally alters the way they process sequential data.

Here's a breakdown of why this architectural constraint leads to martingale violations:

1. **The Martingale Property and Exchangeability:** The martingale property is a fundamental mathematical consequence of Bayesian updating, particularly when dealing with **exchangeable data**. Exchangeable data implies that the order of observations carries no information, meaning that a Bayesian posterior predictive distribution should yield the same expected prediction regardless of how the input sequence is permuted.

* Formally, for exchangeable data, $E[f(X_{n+1})|X_1, \dots, X_n] = E[f(X_{n+1})|X_{\pi(1)}, \dots, X_{\pi(n)}]$ for any permutation π and bounded function f .

2. **Positional Encodings Break Exchangeability:** Transformer architectures, while employing permutation-invariant attention mechanisms, must incorporate **positional encodings** (PEs) to process sequential information and understand the order of tokens. These encodings, such as sinusoidal, learned, or rotary (RoPE), explicitly provide position-dependent information to the model.

* This design choice **explicitly breaks the symmetry and assumption of exchangeability** by making the model's computations inherently depend on the order

of inputs.

3. ****Altered Learning Problem:**** Because of positional encodings, transformers with PEs are not trying to minimize the permutation-invariant Kolmogorov complexity ` $K(X)$ ` (as would be the case for classical Bayesian inference on exchangeable data). Instead, they minimize the ****expected conditional Kolmogorov complexity**** $E_{\pi}[K(X|\pi)]$ over permutations^{*}. This means the model's objective inherently accounts for the specific ordering (π) of the input sequence.

4. ****Quantified Violation:**** This information-theoretic distinction directly leads to systematic martingale violations. The research paper quantifies these violations as being of order $\Theta(\log n/n)$, where n is the sequence length.

* The $\log n$ factor in this scaling arises from the ****expected distance between random permutations**** and the combinatorial structure of permutations. Permutations that are "further" apart in permutation space induce larger differences in the model's predictions.

5. ****Empirical Confirmation:**** Experiments on models like GPT-3 confirm that martingale violations indeed follow this predicted $\Theta(\log n/n)$ scaling, with an adjusted $R^2 > 0.75$. This empirical evidence strongly supports the theoretical explanation linking positional encodings to martingale violations.

In essence, transformers violate martingale properties not due to a flaw, but as a ****direct and quantified consequence of their architectural design choice to include positional encodings****. This architectural bias towards specific orderings allows them to process sequences effectively, even if it means they are "Bayesian in expectation, not in realization" – achieving optimal compression when averaged over orderings, despite differing predictions for specific orderings.