

A note on the impossibility of conditional PAC-efficient reasoning in large language models

Hao Zeng

Department of Statistics and Data Science
Southern University of Science and Technology
`zengh@sustech.edu.cn`

December 4, 2025

Abstract

We prove an impossibility result for conditional Probably Approximately Correct (PAC)-efficient reasoning in large language models. While recent work has established marginal PAC efficiency guarantees for composite models that switch between expensive expert models and cheaper fast models, we show that conditional (pointwise) guarantees are impossible in the distribution-free setting. Specifically, for non-atomic input spaces, any algorithm achieving conditional PAC efficiency must be trivial in the sense that it defers to the expert model with probability at least $1 - \alpha$ for almost every input.

1 Introduction

Large language models have achieved remarkable progress in complex problem-solving, but suffer from high computational costs during deployment (Kwon et al., 2023). To address this, various approaches have been proposed, including model routing (Ong et al., 2025; Dekoninck et al., 2025), speculative decoding (Leviathan et al., 2023), and adaptive reasoning strategies (Snell et al., 2024). Zeng et al. (2025) proposed PAC reasoning, which constructs a composite model \hat{f} that selectively switches between an expensive expert model f and a cheaper fast model \tilde{f} while providing statistical guarantees on performance loss. A typical example is the thinking-nonthinking paradigm, where the expert model performs extended chain-of-thought reasoning while the fast model generates direct responses.

The original PAC reasoning provides marginal guarantees, controlling the expected risk over the input distribution. A natural extension is whether we can achieve a stronger, conditional guarantee that controls the risk for each input point individually. This is analogous to the notion of object-conditional validity in conformal prediction (Vovk, 2012; Lei and Wasserman, 2014; Lei et al., 2018).

However, Barber et al. (2021) established fundamental limits on distribution-free conditional predictive inference, showing that exact conditional coverage is impossible without distributional assumptions. Similar impossibility results have been explored in the context of conformal risk control (Angelopoulos et al., 2025b; Gibbs et al., 2025).

In this note, we establish an impossibility result for the PAC reasoning setting: conditional PAC efficiency implies triviality. Specifically, any algorithm achieving conditional PAC efficiency must defer to the expert model with probability at least $1 - \alpha$ for almost every input, providing no efficiency improvement.

2 Preliminaries

2.1 A router system

Let \mathcal{X} denote the input space and \mathcal{Y} the output space. We assume data are generated from a joint distribution P over $\mathcal{X} \times \mathcal{Y}$, with P_X denoting the marginal distribution on \mathcal{X} . Given input $x \in \mathcal{X}$, the expert model $f : \mathcal{X} \rightarrow \mathcal{Y}$ produces $y = f(x)$, while the fast model $\tilde{f} : \mathcal{X} \rightarrow \mathcal{Y}$ produces $\tilde{y} = \tilde{f}(x)$. In the calibration dataset $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n \sim P^n$ (i.e., n independent samples from P), we set $y_i = f(x_i)$ to be the expert model's output. A router system constructs a composite model via a routing function $g : \mathcal{X} \rightarrow \{0, 1\}$:

$$\hat{f}(x) = \begin{cases} f(x) & \text{if } g(x) = 1, \\ \tilde{f}(x) & \text{if } g(x) = 0. \end{cases}$$

The pointwise risk at input x is $R(\hat{f}; x) = \mathbf{1}\{g(x) = 0\} \cdot \ell(\tilde{f}(x), f(x))$, where $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$ is a loss function. For example, in the thinking-nonthinking paradigm, we might use the 0-1 loss $\ell(\tilde{y}, y) = \mathbf{1}\{\tilde{y} \neq y\}$ to measure whether the fast model's direct response matches the expert model's reasoning-based output. A simple implementation of the routing function is the single-threshold system introduced by Zeng et al. (2025). Given a score function $s : \mathcal{X} \rightarrow \mathbb{R}$ that measures the difficulty or uncertainty of an input, the routing function takes the form

$$g(x) = \mathbf{1}\{s(x) > \tau\},$$

where the threshold τ is chosen based on the calibration dataset to control the marginal risk. Inputs with scores above the threshold are routed to the expert model, while those below are handled by the fast model.

2.2 PAC efficiency

Definition 1 (Marginal PAC efficiency (Zeng et al., 2025)). An algorithm \mathcal{A} is (ϵ, α) -marginally PAC efficient if for all distributions P ,

$$\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n, X \sim P_X} (R(\hat{f}; X) > \epsilon) \leq \alpha.$$

The PAC reasoning algorithm proposed by Zeng et al. (2025) is one approach that achieves marginal PAC efficiency. It constructs a simple router based on learn-then-test framework (Angelopoulos et al., 2025a) applied to the calibration dataset.

Definition 2 (Conditional PAC efficiency). An algorithm \mathcal{A} is (ϵ, α) -conditionally PAC efficient if for all distributions P and P_X -almost every $x \in \mathcal{X}$,

$$\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n} (R(\hat{f}; x) > \epsilon) \leq \alpha.$$

A trivial approach to achieving conditional PAC efficiency is to always use the expert model, i.e., set $g(x) = 1$ for all x . This guarantees $R(\hat{f}; x) = 0$ for every input, trivially satisfying the conditional PAC efficiency requirement. However, such an algorithm provides no computational savings, as it never uses the fast model. The key question is whether non-trivial algorithms, those use the fast model with probability greater than α for some inputs, can achieve conditional PAC efficiency.

3 Impossibility of conditional PAC efficiency

Theorem 3 (Impossibility). *Let \mathcal{X} be a non-atomic complete separable metric space. Assume the fast model \hat{f} has non-trivial loss, i.e., there exists $E \subset \mathcal{X}$ with $P_X(E) > 0$ such that $\ell(\hat{f}(x), f(x)) > \epsilon$ for all $x \in E$. Then an algorithm \mathcal{A} is (ϵ, α) -conditionally PAC efficient if and only if for all distributions P and P_X -almost every $x \in \mathcal{X}$,*

$$\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n} (g(x) = 0) \leq \alpha.$$

Remark 4. This result implies that any algorithm achieving meaningful efficiency gains cannot satisfy conditional PAC efficiency: An algorithm satisfying the condition in the theorem uses the fast model with probability at most α for almost every input, which means it essentially always defers to the expensive expert model and provides no efficiency improvement.

Proof of Theorem 3. (\Rightarrow) Suppose for all distributions P and P_X -almost every $x \in \mathcal{X}$, we have $\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n} (g(x) = 0) \leq \alpha$. Then for P_X -almost every x ,

$$\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n} (R(\hat{f}; x) > \epsilon) \leq \mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n} (g(x) = 0) \leq \alpha.$$

(\Leftarrow) Assume \mathcal{A} is (ϵ, α) -conditionally PAC efficient. Suppose for contradiction that there exists a distribution P and a set $E \subset \mathcal{X}$ with $P_X(E) > 0$ such that $\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n} (g(x) = 0) > \alpha$ for all $x \in E$. Note that by the definition of conditional PAC efficiency, we must have

$$\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n} (R(\hat{f}; x) > \epsilon) \leq \alpha$$

for P_X -almost every $x \in \mathcal{X}$. Since $R(\hat{f}; x) = \mathbf{1}\{g(x) = 0\} \cdot \ell(\tilde{f}(x), f(x))$, if $\ell(\tilde{f}(x), f(x)) > \epsilon$, then $R(\hat{f}; x) > \epsilon$ implies $g(x) = 0$. Therefore, for any x where the fast model has loss exceeding ϵ , we have

$$\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n}(g(x) = 0) \leq \mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n}(R(\hat{f}; x) > \epsilon) \leq \alpha.$$

Our contradiction hypothesis assumes the opposite: there exists a set E with positive measure where $\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n}(g(x) = 0) > \alpha$ for all $x \in E$.

Fix $x^* \in E$. The key observation is that we can construct a distribution that is nearly indistinguishable from P based on finite samples, yet has high loss at x^* . By Lemma 5, for any $\eta > 0$, there exists a distribution P' on $\mathcal{X} \times \mathcal{Y}$ such that: (i) the fast model has loss exceeding ϵ at x^* under P' , and (ii) the total variation distance between their n -fold products satisfies $\text{TV}((P')^n, P^n) < \eta$.

By the conditional PAC guarantee under P' , since $\ell(\tilde{f}(x^*), f(x^*)) > \epsilon$, we have $\mathbb{P}_{(P')^n}(g(x^*) = 0) \leq \alpha$. By the finite-sample indistinguishability property, $|\mathbb{P}_{P^n}(g(x^*) = 0) - \mathbb{P}_{(P')^n}(g(x^*) = 0)| < \eta$. Taking $\eta \rightarrow 0$ yields $\mathbb{P}_{P^n}(g(x^*) = 0) \leq \alpha$. This contradicts our assumption that $\mathbb{P}_{\mathcal{D}_{\text{cal}} \sim P^n}(g(x^*) = 0) > \alpha$. By separability of \mathcal{X} , this holds for P_X -almost every x . \square

4 Conclusion

We have established a fundamental impossibility result for conditional PAC efficiency in distribution-free LLM routing. Our theorem shows that any algorithm achieving conditional PAC efficiency must be trivial, using the fast model with probability at most α for almost every input. This extends classical impossibility results from distribution-free predictive inference (Barber et al., 2021) to the PAC reasoning framework. The practical implications are clear: while marginal PAC efficiency enables meaningful computational savings (Zeng et al., 2025), conditional guarantees eliminate most efficiency gains. Practitioners should focus on marginal guarantees or explore relaxed notions of conditional efficiency under distributional assumptions.

A Technical Lemmas

Lemma 5 (Finite-sample indistinguishability). *Let \mathcal{X} be a non-atomic complete separable metric space. Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ and $x^* \in \mathcal{X}$. For any $\eta > 0$, there exists a distribution P' on $\mathcal{X} \times \mathcal{Y}$ such that:*

1. *Under P' : $\ell(\tilde{f}(x^*), f(x^*)) > \epsilon$ almost surely (conditional on $X = x^*$).*
2. *$\text{TV}((P')^n, P^n) < \eta$, where TV denotes the total variation distance.*

Proof. Let $B = B(x^*, r)$ be the open ball of radius r centered at x^* . Since \mathcal{X} is non-atomic, we can choose r small enough that $\delta := P_X(B) < \eta/(2n)$. We

construct P' by modifying P locally around x^* . We keep the same marginal distribution on X , but modify the conditional distribution of Y given X :

$$P'(Y|X = x) = \begin{cases} P(Y|X = x) & \text{if } x \notin B, \\ Q_\epsilon & \text{if } x \in B, \end{cases}$$

where Q_ϵ is a distribution such that $\ell(\tilde{f}(x^*), y) > \epsilon$ for all $y \in \text{supp}(Q_\epsilon)$.

The key observation is that P' and P differ only on the small ball B . The probability that at least one calibration point falls in B under P (or P') is at most $n\delta < \eta/2$. On the complementary event (probability at least $1 - \eta/2$), the calibration sets under P' and P have identical distributions. Therefore, $\text{TV}((P')^n, P^n) \leq 2n\delta < \eta$. \square

References

- Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. (2025a). Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19:1641–1662.
- Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2025b). Conformal risk control. *arXiv preprint arXiv:2208.02814*.
- Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10:455–482.
- Dekoninck, J., Baader, M., and Vechev, M. (2025). A unified approach to routing and cascading for LLMs. In *Forty-Second International Conference on Machine Learning*.
- Gibbs, I., Cherian, J. J., and Candès, E. J. (2025). Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf008.
- Kwon, Li, Zhuang, Sheng, Zheng, Yu, and et al. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP ’23, pages 611–626, New York, NY, USA. Association for Computing Machinery.
- Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111.
- Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:71–96.

- Leviathan, Y., Kalman, M., and Matias, Y. (2023). Fast inference from transformers via speculative decoding. *arXiv preprint arXiv:2211.17192*.
- Ong, Almahairi, Wu, Chiang, Wu, Gonzalez, and et al. (2025). RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*.
- Snell, C., Lee, J., Xu, K., and Kumar, A. (2024). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, pages 475–490. PMLR.
- Zeng, H., Huang, J., Jing, B., Wei, H., and An, B. (2025). PAC reasoning: Controlling the performance loss for efficient reasoning. *arXiv preprint arXiv:2510.09133*.