Guided Speculative Inference for Efficient Test-Time Alignment of LLMs

Jonathan Geuter [1 2] **Youssef Mroueh** [3] **David Alvarez-Melis** [1 2]

Abstract

We propose *Guided Speculative Inference* (GSI), a novel algorithm for efficient reward-guided decoding in large language models. GSI combines soft best-of-n test-time scaling with a reward model $r(x, y)$ and speculative samples from a small auxiliary model $\pi_S(y \mid x)$. We provably approximate the optimal tilted policy $\pi_{\beta,B}(y \mid x) \propto \pi_B(y \mid x) \exp(\beta\, r(x, y))$ of soft best-of-n under the primary model π_B. We derive a theoretical bound on the KL divergence between our induced distribution and the optimal policy. In experiments on reasoning benchmarks (MATH500, OlympiadBench, Minerva Math), our method achieves higher accuracy than standard soft best-of-n with π_S and reward-guided speculative decoding (Liao et al., 2025), and in certain settings even outperforms soft best-of-n with π_B. The code is available at: https://github.com/j-geuter/GSI.

1. Introduction

Large language models (LLMs) have demonstrated remarkable performance across diverse generation tasks, with scaling model and data size emerging as a reliable and efficient way to enhance their capabilities (Kaplan et al., 2020; Team, 2024; OpenAI, 2024). However, this scaling has resulted in significant computational and economic costs, prompting the need for more efficient alternatives. One such approach is *test-time scaling* (Snell et al., 2024; Muennighoff et al., 2025; Zhang et al., 2025), which focuses on scaling inference-time rather than training time compute. An orthogonal direction in LLM post-training is *model alignment* (Ouyang et al., 2022; Gao et al., 2022; Touvron et al., 2023), where models are optimized to maximize a given reward model $r(x, y)$ that quantifies the quality of a response y given a prompt x. Several techniques have been proposed for aligning LLMs with reward models (Yang & Klein, 2021; Mudgal et al., 2024; Huang et al., 2025). Recent work on reward-guided speculative decoding (RSD) (Liao et al., 2025) introduces a single-step speculative check to filter samples by a reward threshold, though it lacks theoretical guarantees on distributional fidelity. Alternatively, best-of-n sampling (Gao et al., 2022; Mroueh, 2024; Beirami et al., 2025) with temperature ("soft BoN") (Verdun et al., 2025) can interpolate between the base distribution π_B and reward maximization, helping to mitigate reward hacking (Skalse et al., 2025).

Contributions. In this paper, we introduce a novel algorithm, *Guided Speculative Inference* (GSI), which combines speculative decoding, soft best-of-n sampling, and rejection sampling. Importantly, by *tilting* (i.e., adjusting) the rewards according to the loglikelihoods under both π_B and π_S, GSI provably approximates the optimal reward-regularized policy under π_B (Section 3), namely

$$\pi_{\beta,B}(y \mid x) \;=\; \frac{\pi_B(y \mid x) \exp(\beta\, r(x, y))}{Z_{\beta,B}(x)}.$$

We evaluate GSI on several reasoning benchmarks and show that it outperforms both reward-guided speculative decoding (Liao et al., 2025) and soft best-of-n sampling (Section 4).

2. Background

Let $\mathcal{V}$ denote a (finite) vocabulary. Let $\mathcal{X} = \{x = (x_1, ..., x_n) : \; n \in \mathbb{N},\; x_i \in \mathcal{V}\}$ be the (countable) space of inputs (in practice, consisting of the prompt and already generated reasoning steps), and $\mathcal{Y} = \{y = (y_1, ..., y_n) : \; n \in \mathbb{N},\; y_i \in \mathcal{V}\}$ the (countable) space of reasoning steps (e.g. token sequences). Note that mathematically, these two spaces are identical, but we define both $\mathcal{X}$ and $\mathcal{Y}$ for notational convenience. By $\Delta(\mathcal{Y})$, we denote the set of probability measures over $\mathcal{Y}$. For $x \in \mathcal{X}$, let $\pi_B(y \mid x) \in \Delta(\mathcal{Y})$ and $\pi_S(y \mid x) \in \Delta(\mathcal{Y})$ be the *base* and *small* language model distributions over $y \in \mathcal{Y}$ given x. Note that we define the distributions over *reasoning steps* instead of single *tokens*. When we write $\pi_B(\cdot \mid x, y)$, it denotes the distribution of π_B over $\mathcal{Y}$ given a prompt x and a (partial) response y. We further assume we are given a process reward model (PRM) (Lightman et al., 2023) $r : \mathcal{X} \times \mathcal{Y} \;\to\; [0, R]$ for some $R < \infty$, which assigns a reward $r(x, y)$ to a reasoning step

[1]Harvard John A. Paulson School of Engineering and Applied Sciences [2]Kempner Institute [3]IBM Research. Correspondence to: Jonathan Geuter <jonathan.geuter@gmx.de>.

ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models, ICML 2025, Vancouver, Canada.

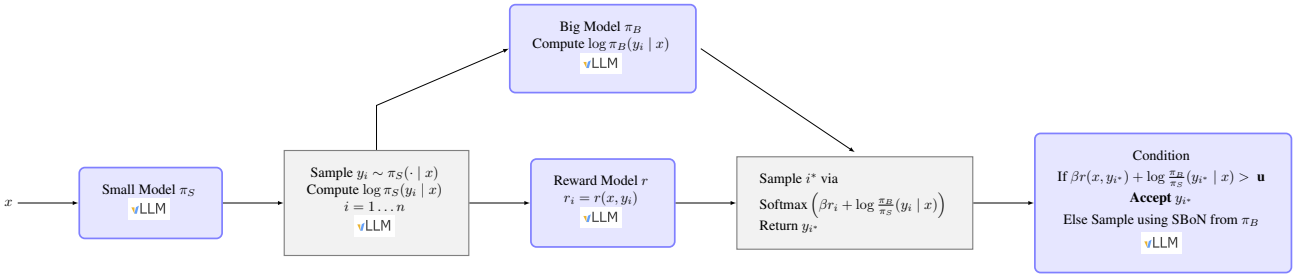

Figure 1. Guided Speculative Inference workflow. All computations involving the small, big, and reward models can be efficiently performed using `vLLM`.

$y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$. We assume that r approximates a *golden reward* (Gao et al., 2022) $r^* : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which can be thought of as the "true" reward function. Recall that the Kullback–Leibler divergence between two distributions $P, Q \in \Delta(\mathcal{Y})$ with $P \ll Q$ is defined as

$$\mathrm{KL}(P\|Q) \;=\; \mathbb{E}_{y\sim P}\big[\log \tfrac{P(y)}{Q(y)}\big],$$

and the chi-square divergence as

$$\chi^2(P||Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q = \int \frac{\mathrm{d}P^2}{\mathrm{d}Q} - 1.$$

KL Regularized Reward Alignment. A standard formulation for maximizing the reward $r(x, y)$ given $x \in \mathcal{X}$, while constraining how far the policy can move from the base policy $\pi_B(\cdot \mid x)$, is to add a KL regularizer, and find π_B^* maximizing

$$\max_{\pi\in\Delta(\mathcal{Y})} \mathbb{E}_{y\sim\pi}[\, r(x, y)\,] - \tfrac{1}{\beta}\,\mathrm{KL}\big(\pi(\cdot \mid x)\,\|\,\pi_B(\cdot \mid x)\big),$$

where $\beta > 0$ trades off maximizing the reward versus fidelity to π_B. It is well known (e.g. (Korbak et al., 2022)) that the optimal policy has the closed form

$$\pi_{\beta,B}(y \mid x) = \frac{\pi_B(y \mid x)\, \exp\big(\beta\, r(x, y)\big)}{Z_{\beta,B}(x)}, \qquad (1)$$

where $Z_{\beta,B}(x) = \mathbb{E}_{y'\sim\pi_B(\cdot|x)}\left[e^{\beta\, r(x,y')}\right]$.

Best-of-n Sampling. Best-of-n (BoN) (Beirami et al., 2025) is a common inference-time method for scaling LLMs. *Hard* best-of-n draws $y_1, \dots, y_n \sim \pi_B(\cdot \mid x)$, and selects

$$y^* \;=\; \arg\max_{i\in\{1,\dots,n\}} r(x, y_i).$$

Soft best-of-n (S-BoN) (Verdun et al., 2025) weighs each draw by $w_i \;\propto\; \exp\big(\beta\, r(x, y_i)\big)$, then sample a response y_i with probability $w_i / \sum_j w_j$. We denote the soft best-of-n distribution over y by $\pi^{r,n}_{\beta,B}(\cdot \mid x)$. Note that both soft and hard BoN can be applied both to one-shot generation (where the complete response it generated in one step) or reasoning tasks, where the y_i correspond to reasoning steps, and the BoN procedure is repeatedly applied. In this work, we focus on reasoning tasks. By moving from hard to soft best-of-n, the distribution $\pi^{r,n}_{\beta,B}(\cdot \mid x)$ enjoys a KL bound to the tilted distribution $\pi_{\beta,B}$ (Verdun et al., 2025):

$$\mathrm{KL}(\pi_{\beta,B}\|\pi^{r,n}_{\beta,B}) \le \log\left(1 + \frac{\mathrm{Var}_{y\sim\pi_B}[e^{\beta r(x,y)}]}{n(\mathbb{E}_{y\sim\pi_B}[e^{\beta r(x,y)}])^2}\right). \qquad (2)$$

Speculative Decoding. Speculative decoding (SD) (Leviathan et al., 2023) accelerates sampling from π_B by first drawing proposals from π_S and then accepting or rejecting them based on a criterion derived from the ratio π_B/π_S. On rejection, one falls back to direct sampling from π_B. SD provably samples from the distributions of π_B. The core idea is that k tokens can be sampled from π_S autoregressively, but verified by π_B in parallel, thus generating up to $k+1$ tokens from π_B with a single forward pass of π_B. Variants of SD include *block verification* (Sun et al., 2025) where sequences of draft tokens are verified jointly instead of token-by-token, and *SpecTr* (Sun et al., 2024) which allows for verification of multiple draft sequences in parallel by framing SD as an optimal transport problem. SD has also been combined with early-exiting (Liu et al., 2024), and (Bhendawade et al., 2024) propose using n-gram predictions of π_B as drafts, which alleviates the need for an auxiliary model.

A recent work proposes RSD (reward-guided speculative decoding) (Liao et al., 2025), where samples are generated from π_S, and a threshold on the reward of the samples from π_S determines whether one should accept the sample or resample from π_B. While this approach shares similarities with GSI, it only provides a guarantee on the expected reward: under the assumption that $\mathbb{E}_{\pi_B}[r(y \mid x)] \ge \mathbb{E}_{\pi_S}[r(y \mid x)]$, RSD satisfies $\mathbb{E}_{\pi_{\mathrm{RSD}}}[r(y \mid x)] \ge \mathbb{E}_{\pi_S}[r(y \mid x)]$, which in the worst case does not yield any improvement over the small model π_S, and also does not guarantee anything about the policy π_{RSD} itself. As we will see in Section 3, GSI provides guarantees on the induced policy directly.

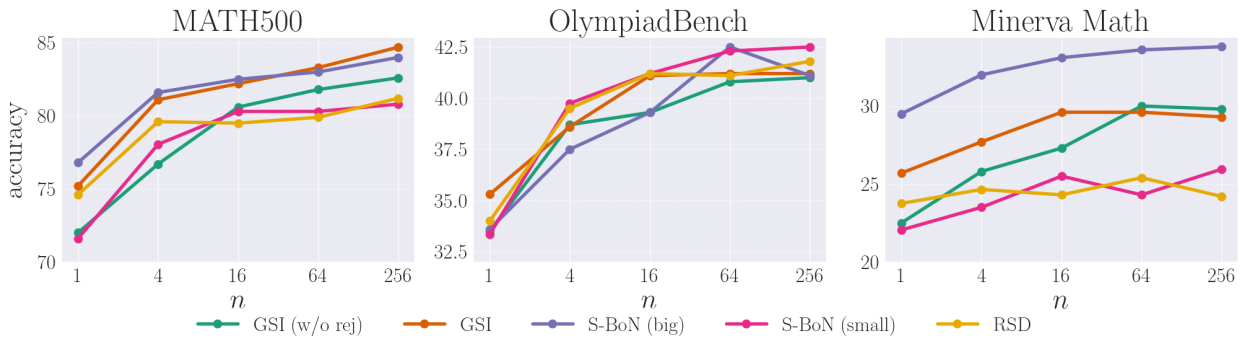

Figure 2. Accuracy over n. We plot GSI without rejection step ($\pi_{\beta,S}^{\tilde{r},n}$), GSI with rejection step (π_{GSI}), S-BoN with the small ($\pi_{\beta,S}$) and big ($\pi_{\beta,B}$) model, and RSD (Liao et al., 2025) (with $\beta = 20$ for all methods).

3. Guided Speculative Inference

Note that we can write the tilted distribution (1) as

$$\pi_{\beta,B}(y \mid x) = \frac{\pi_S(y \mid x) \exp\left(\beta r(x,y) + \log\left(\frac{\pi_B(y|x)}{\pi_S(y|x)}\right)\right)}{Z_{\beta,B}(x)},$$

i.e. we can rewrite it as a distribution over π_S (exponentially) tilted by the rewards

$$\tilde{r}(x,y) = r(x,y) + \frac{1}{\beta}\log\left(\frac{\pi_B(y \mid x)}{\pi_S(y \mid x)}\right)$$

Thus, one can sample from π_S and re-weight candidates to approximate $\pi_{\beta,B}$:

Reward-Likelihood Tilted S-BoN. For $x \in \mathcal{X}$, the reward-tilted S-BoN is defined as follows:

1. sample $y_1, ..., y_n \sim \pi_S(\cdot \mid x)$
2. compute $\tilde{r}_i = r(x, y_i) + \frac{1}{\beta}\log\left(\frac{\pi_B(y_i|x)}{\pi_S(y_i|x)}\right)$
3. sample $y_i \propto \exp(\beta \tilde{r}_i)$

We will denote the distribution generated by this sampling algorithm by $\pi_{\beta,S}^{\tilde{r},n}(\cdot \mid x)$. Of course, we can only hope that $\pi_{\beta,S}^{\tilde{r},n}(\cdot \mid x)$ is close to $\pi_{\beta,B}(\cdot \mid x)$ if the support of π_B is sufficiently covered by π_S.

Coverage Assumption. Throughout, we will assume that

$$C_\infty(x) := \sup_{y \in \mathcal{Y}: \pi_B(y|x) > 0} \frac{\pi_B(y \mid x)}{\pi_S(y \mid x)} < \infty. \tag{3}$$

Under this assumption, Reward-Likelihood Tilted S-BoN with π_S indeed approximates the tilted distribution $\pi_{\beta,B}$ in the sense of the following theorem.

Algorithm 1 Guided Speculative Inference

input base model π_B, small model π_S, PRM r, $\beta > 0$, threshold $u \in \mathbb{R}$, $n \in \mathbb{N}$, prompt $x \in \mathcal{X}$

$y \leftarrow ()$ `#empty response`
for $t = 0, 1, ...,$ until EOS **do**
  Sample $\{y_t^i\}_{i=1}^n \sim \pi_S(\cdot \mid x, y)$
  $\tilde{r}^i \leftarrow r(x, y_t^i) + \frac{1}{\beta}\left(\log \pi_B(y_t^i \mid x) - \log \pi_S(y_t^i \mid x)\right)$
  Sample index $c \sim \text{softmax}(\beta\tilde{r}^1, ..., \beta\tilde{r}^n)$
  if $\tilde{r}^c \geq u$ **then**
    $y \leftarrow (y, y_t^c)$ `#append y^c_t`
  else
    Sample $\{y_t^j\}_{j=1}^n \sim \pi_B(\cdot \mid x, y)$
    $r^j \leftarrow r(x, y_t^j)$
    Sample index $c \sim \text{softmax}(\beta r^1, ..., \beta r^n)$
    $y \leftarrow (y, y_t^c)$
  end if
end for

Theorem 1. *Let $x \in \mathcal{X}$. Assume that the coverage assumption (3) holds. Let $\epsilon > 0$ be arbitrary, and*

$$n \geq \frac{\left(\chi^2\left(\pi_B(\cdot \mid x) \,\|\, \pi_S(\cdot \mid x)\right) + 1\right) e^{2\beta\|r\|_\infty} - 1}{e^\epsilon - 1}.$$

Then,

$$\text{KL}\left(\pi_{\beta,B}(\cdot \mid x) \,\|\, \pi_{\beta,S}^{\tilde{r},n}(\cdot \mid x)\right) \leq \epsilon.$$

The proof can be found in Appendix B.

This lies at the core of our proposed algorithm. In addition to sampling from the Reward-Likelihood Tilted S-BoN, we also add a rejection sampling-like threshold on the tilted reward, which triggers resampling from the base model π_B in case the tilted reward falls below it. This improves performance empirically. The complete GSI method can be seen in Algorithm 1. Note that in principle, it is possible to choose different n for the small and large model in the algorithm. We leave exploring this for future research. Note that

Table 1. Accuracies on reasoning benchmarks for $n = 16$ and $n = 64$. GSI performs better than RSD (Liao et al., 2025) and S-BoN with the small model. S-BoN with the big model is the target distribution.

Method	MATH500	OlympiadBench	Minerva Math	Mean Acc.
GSI, n=16 (ours)	**82.2**	41.4	**29.6**	**51.1**
RSD, n=16	79.5	**41.7**	24.3	48.5
S-BoN (small), n=16	80.3	41.2	25.5	49.0
S-BoN (big), n=16	82.5	39.3	36.0	52.6
GSI, n=64 (ours)	**83.3**	41.2	**29.6**	**51.3**
RSD, n=64	79.9	41.1	25.4	48.8
S-BoN (small), n=64	80.3	**42.3**	24.3	49.0
S-BoN (big), n=64	83.0	42.5	36.4	54.0

Table 2. Avg. inference time (in seconds) per reasoning step, avg. number of reasoning steps per sample, and avg. percentage of samples accepted (averaged across all datasets). GSI is significantly faster than S-BoN on the big model, but has slightly higher inference time per step than RSD (Liao et al., 2025) as the acceptance rate is lower.

Method	s / step	# steps	% accept
GSI, n=16 (ours)	0.87	13.9	91
RSD, n=16	0.74	11.6	97
S-BoN (small), n=16	0.79	11.0	–
S-BoN (big), n=16	1.14	11.8	–
GSI, n=64 (ours)	2.06	15.3	93
RSD, n=64	1.88	12.7	98
S-BoN (small), n=64	1.99	10.9	–
S-BoN (big), n=64	2.76	11.7	–

while GSI is, in theory, applicable to one-shot generation tasks, we consider y_t in Algorithm 1 to be a reasoning step (i.e., a subsequence of the full response), and r is a process reward model (PRM). The algorithm generates reasoning steps until an end-of-sequence (EOS) token is created.

We denote the distribution generated by Algorithm 1 as π_{GSI}. While Theorem 1 does not apply to π_{GSI} (only to $\pi_{\beta,S}^{\tilde{r},n}$), we can also guarantee that the expected difference in (golden) reward goes to 0 as n increases (see Theorem 2 in Appendix B):

$$\mathbb{E}_{\pi_{\beta,B}}[r^*] - \mathbb{E}_{\pi_{\text{GSI}}}[r^*] \xrightarrow{n\to\infty} 0.$$

4. Experiments

Models. We use Qwen2.5-Math-1.5B-Instruct as π_S, Qwen2.5-Math-7B-Instruct as π_B, and Qwen2.5-Math-PRM-7B as the PRM r throughout. The rewards lie in $[0, 1]$.

Implementation. We implement all models with vLLM (Kwon et al., 2023). The log-likelihoods for π_S are computed without any additional computational overhead within the forward pass of π_S. The log-likelihoods for π_B can be computed with minimal computational overhead, as they only require a single forward pass through π_B. We host each of the three models on its own NVIDIA H100 GPU.

Datasets. We evaluate on three mathematical reasoning benchmarks: MATH500 (Lightman et al., 2023), OlympiadBench (He et al., 2024) (the OE_TO_maths_en_COMP split which is text-only math problems in English), and Minerva Math (Lewkowycz et al., 2022). We decode stepwise with chain-of-thought; rewards are computed on each reasoning step. For each method and dataset, we report the average accuracies over two different random seeds.

Methods. We compare GSI against RSD (Liao et al., 2025), S-BoN with π_S, and S-BoN with π_B.

Hyperparameters. We use $\beta = 20$, $u = 0.5$ (selected empirically amongst a range of values based on acceptance rate vs. accuracy trade-off), `temperature` = 0.7, and `top_p` = 1.0. We set the threshold in RSD to 0.7, which is the same as in the RSD paper.

4.1. Performance on Reasoning Benchmarks

In Table 1, we compare the average accuracies of GSI, RSD, and S-BoN on the small and big model. We see that GSI clearly outperforms RSD and S-BoN on the small model across the datasets. While RSD is slightly better on OlympiadBench, this might be the case because the small model is better on this dataset than the big model, and RSD is closer to the small model due to its high overall acceptance rate, cmp. Table 2.

Figure 2 compares GSI without the rejection sampling step (i.e., without lines 6 to 11 in Algorithm 1) to regular GSI, S-BoN with π_B and π_S, and RSD. We see that GSI clearly outperforms GSI without rejection step; however, this difference becomes less significant as n increases, hinting at the fact that with larger n, the samples from the small model

reach better coverage of the support of π_B. As the accuracy of GSI without the rejection step, i.e. $\pi_{\beta,S}^{\tilde{r},n}$, approaches the accuracy of S-BoN with the big model, Figure 2 also empirically verifies Theorem 1. In future research, we plan to investigate this behavior as we scale n beyond 256.

In Table 2, we report the inference time per sample (in seconds) across methods (averaged over datasets), as well as the average percentage of samples accepted by GSI and RSD. We see that RSD generally tends to accept almost all samples, which explains why its performance is comparable to S-BoN with the small model, compare Table 1, while being slightly worse in terms of inference speed. GSI accepts less samples, thus is slower than RSD, while still outperforming S-BoN on the large model in terms of inference speed.

5. Discussion

We introduced *Guided Speculative Inference* (GSI), a novel inference-time algorithm for efficient reward-guided decoding from language models. GSI leverages speculative samples from a small auxiliary model to approximate the optimal tilted policy of a base model with respect to a given reward function. We showed that unlike previous approaches, GSI provably approaches the optimal policy as the number of samples generated at each step n increases. Empirical results on various reasoning datasets show that GSI significantly outperforms reward-guided speculative decoding (Liao et al., 2025) and soft best-of-n using the small model—and, perhaps surprisingly, even surpasses soft best-of-n with the base model in some cases. Future work will explore extending GSI beyond reasoning tasks (e.g., alignment to safety rewards), studying its scaling behavior with respect to n, and analyzing its sensitivity to different values of β and n across both models. Deriving tighter theoretical bounds is also an important aspect in better understanding the behavior of GSI.

Acknowledgements

We would like to thank Nick Hill and Guangxuan (GX) Xu from Red Hat for their help with vLLM. JG and DAM acknowledge support from the Kempner Institute, the Aramont Fellowship Fund, and the FAS Dean's Competitive Fund for Promising Scholarship.

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.

References

Beirami, A., Agarwal, A., Berant, J., D'Amour, A., Eisenstein, J., Nagpal, C., and Suresh, A. T. Theoretical guarantees on the best-of-n alignment policy, 2025. URL https://arxiv.org/abs/2401.01879.

Bhendawade, N., Belousova, I., Fu, Q., Mason, H., Rastegari, M., and Najibi, M. Speculative Streaming: Fast LLM Inference without Auxiliary Models, 2024. URL https://arxiv.org/abs/2402.11131.

Gao, L., Schulman, J., and Hilton, J. Scaling Laws for Reward Model Overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., and Sun, M. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems, 2024. URL https://arxiv.org/abs/2402.14008.

Huang, A., Block, A., Liu, Q., Jiang, N., Krishnamurthy, A., and Foster, D. J. Is Best-of-N the Best of Them? Coverage, Scaling, and Optimality in Inference-Time Alignment, 2025. URL https://arxiv.org/abs/2503.21878.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models, 2020. URL https://arxiv.org/abs/2001.08361.

Korbak, T., Perez, E., and Buckley, C. L. RL with KL penalties is better viewed as Bayesian inference, 2022. URL https://arxiv.org/abs/2205.11275.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention, 2023. URL https://arxiv.org/abs/2309.06180.

Leviathan, Y., Kalman, M., and Matias, Y. Fast Inference from Transformers via Speculative Decoding, 2023. URL https://arxiv.org/abs/2211.17192.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving Quantitative Reasoning Problems with Language Models, 2022. URL https://arxiv.org/abs/2206.14858.

Liao, B., Xu, Y., Dong, H., Li, J., Monz, C., Savarese, S., Sahoo, D., and Xiong, C. Reward-Guided Speculative

Decoding for Efficient LLM Reasoning, 2025. URL https://arxiv.org/abs/2501.19324.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023.

Liu, J., Wang, Q., Wang, J., and Cai, X. Speculative decoding via early-exiting for faster llm inference with thompson sampling control mechanism, 2024. URL https://arxiv.org/abs/2406.03853.

Mroueh, Y. Information theoretic guarantees for policy alignment in large language models, 2024. URL https://arxiv.org/abs/2406.05883.

Mudgal, S., Lee, J., Ganapathy, H., Li, Y., Wang, T., Huang, Y., Chen, Z., Cheng, H.-T., Collins, M., Strohman, T., Chen, J., Beutel, A., and Beirami, A. Controlled Decoding from Language Models, 2024. URL https://arxiv.org/abs/2310.17022.

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

OpenAI. GPT-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Skalse, J., Howe, N. H. R., Krasheninnikov, D., and Krueger, D. Defining and Characterizing Reward Hacking, 2025. URL https://arxiv.org/abs/2209.13085.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Sun, Z., Suresh, A. T., Ro, J. H., Beirami, A., Jain, H., and Yu, F. SpecTr: Fast Speculative Decoding via Optimal Transport, 2024. URL https://arxiv.org/abs/2310.15141.

Sun, Z., Mendlovic, U., Leviathan, Y., Aharoni, A., Ro, J. H., Beirami, A., and Suresh, A. T. Block Verification Accelerates Speculative Decoding, 2025. URL https://arxiv.org/abs/2403.10444.

Team, G. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Touvron, H., Martin, L., Stone, K., Albert, P., and et al. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL https://arxiv.org/abs/2307.09288.

Verdun, C. M., Oesterling, A., Lakkaraju, H., and Calmon, F. P. Soft Best-of-n Sampling for Model Alignment, 2025. URL https://arxiv.org/abs/2505.03156.

Yang, K. and Klein, D. FUDGE: Controlled Text Generation With Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.276. URL http://dx.doi.org/10.18653/v1/2021.naacl-main.276.

Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Hua, W., Wu, H., Guo, Z., Wang, Y., Muennighoff, N., King, I., Liu, X., and Ma, C. A survey on test-time scaling in large language models: What, how, where, and how well?, 2025. URL https://arxiv.org/abs/2503.24235.

A. Code

The code is available at: https://github.com/j-geuter/GSI.

B. Proofs

Theorem 1. *Let $x \in \mathcal{X}$. Assume that the coverage assumption (3) holds. Let $\epsilon > 0$ be arbitrary, and*

$$n \geq \frac{\left(\chi^2\big(\pi_B(\cdot \mid x) \,\|\, \pi_S(\cdot \mid x)\big) + 1\right) e^{2\beta\|r\|_\infty} - 1}{e^\epsilon - 1}.$$

Then,

$$\mathrm{KL}\big(\pi_{\beta,B}(\cdot \mid x) \,\|\, \pi_{\beta,S}^{\tilde{r},n}(\cdot \mid x)\big) \;\leq\; \epsilon.$$

Proof. By Lemma 1 in (Verdun et al., 2025) (which equally holds for countable instead of finite spaces), we have

$$\begin{aligned}
\pi_{\beta,S}^{\tilde{r},n}(y \mid x) \;&\geq\; \frac{\pi_S(y \mid x) \exp\Big[\beta\, r(x,y) + \log \frac{\pi_B(y \mid x)}{\pi_S(y \mid x)}\Big]}{\frac{1}{n}\,\exp\Big[\beta\, r(x,y) + \log \frac{\pi_B(y|x)}{\pi_S(y|x)}\Big] \;+\; \frac{n-1}{n}\,\mathbb{E}_{y' \sim \pi_S(\cdot|x)}\Big[\frac{\pi_B(y'|x)}{\pi_S(y'|x)} e^{\beta\, r(x,y')}\Big]} \\
&= \frac{\pi_B(y \mid x)\, e^{\beta\, r(x,y)}}{\frac{1}{n}\,\frac{\pi_B(y \mid x)}{\pi_S(y \mid x)}\, e^{\beta\, r(x,y)} \;+\; \frac{n-1}{n}\,\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta\, r(x,y')}\big]}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{KL}\big(\pi_{\beta,B} \| \pi_S^{n,t}\big) &= \sum_y \pi_{\beta,B}(y \mid x)\, \log \frac{\pi_{\beta,B}(y \mid x)}{\pi_S^{n,t}(y \mid x)} \\
&\leq \sum_y \frac{\pi_B(y \mid x) e^{\beta r(x,y)}}{\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta r(x,y')}\big]}\, \log \left(\frac{\pi_B(y \mid x)\, e^{\beta r(x,y)} \left[\frac{1}{n}\frac{\pi_B(y \mid x)}{\pi_S(y \mid x)} e^{\beta r(x,y)} + \frac{n-1}{n}\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta r(x,y')}\big]\right]}{\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta r(x,y')}\big]\pi_B(y \mid x)\, e^{\beta r(x,y)}} \right) \\
&= \sum_y \frac{\pi_B(y \mid x) e^{\beta r(x,y)}}{\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta r(x,y')}\big]}\, \log \left(\frac{1}{n}\,\frac{\pi_B(y \mid x)}{\pi_S(y \mid x)}\,\frac{e^{\beta\, r(x,y)}}{\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta\, r(x,y')}\big]} + \frac{n-1}{n} \right) \\
&\leq \log \left(\frac{1}{n} \left(\sum_y \frac{\pi_B(y \mid x)^2}{\pi_S(y \mid x)}\, \frac{e^{2\beta\, r(x,y)}}{\big(\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta\, r(x,y')}\big]\big)^2} \right) + \frac{n-1}{n} \right) \qquad \text{(Jensen's inequality)} \\
&\leq \log \left(\frac{1}{n}\, e^{2\beta\,\|r\|_\infty} \frac{\chi^2\big(\pi_B(\cdot \mid x) \,\|\, \pi_S(\cdot \mid x)\big) + 1}{\big(\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta\, r(x,y')}\big]\big)^2} + \frac{n-1}{n} \right) \\
&\leq \log \left(\frac{\left(\chi^2\big(\pi_B(\cdot|x) \,\|\, \pi_S(\cdot|x)\big)+1\right) e^{2\beta\|r\|_\infty}}{n} + \frac{n-1}{n} \right),
\end{aligned}$$

using the fact that

$$\mathbb{E}_{y' \sim \pi_B(\cdot|x)}\big[e^{\beta\, r(x,y')}\big] \geq 1$$

since $r(x, y') \geq 0$. Now for $\epsilon > 0$, we have

$$\begin{aligned}
&\log\left(\frac{\left(\chi^2\left(\pi_B(\cdot|x)\,\|\,\pi_S(\cdot|x)\right)+1\right)e^{2\beta\|r\|_\infty}}{n} + \tfrac{n-1}{n}\right) \;\leq\; \epsilon \\
\Leftrightarrow\quad &\frac{\left(\chi^2\left(\pi_B(\cdot|x)\,\|\,\pi_S(\cdot|x)\right)+1\right)e^{2\beta\|r\|_\infty}}{n} + 1 - \tfrac{1}{n} \;\leq\; e^\epsilon, \\
\Leftrightarrow\quad &1 + \frac{\left(\chi^2\left(\pi_B(\cdot|x)\,\|\,\pi_S(\cdot|x)\right)+1\right)e^{2\beta\|r\|_\infty - 1}}{n} \;\leq\; e^\epsilon, \\
\Leftrightarrow\quad &\frac{\left(\chi^2\left(\pi_B(\cdot|x)\,\|\,\pi_S(\cdot|x)\right)+1\right)e^{2\beta\|r\|_\infty - 1}}{n} \;\leq\; e^\epsilon - 1, \\
\Leftrightarrow\quad &\frac{\left(\chi^2\big(\pi_B(\cdot\mid x)\,\|\,\pi_S(\cdot\mid x)\big)+1\right)e^{2\beta\|r\|_\infty} - 1}{e^\epsilon - 1} \;\leq\; n,
\end{aligned}$$

which finishes the proof. □

Theorem 2. *Let $x \in \mathcal{X}$. Assume that $\mathbb{E}_{\pi_{\mathrm{GSI}}}[r^*] < \infty$ and $\mathbb{E}_{\pi_{\beta,B}}[r^*] < \infty$ (here we implicitly assume that distributions and rewards are conditioned on x, which we omit for ease of notation). Furthermore, assume the coverage assumption (3) holds. Denote by $p(u)$ the acceptance probability of GSI. Then*

$$\mathbb{E}_{\pi_{\beta,B}}[r^*] - \mathbb{E}_{\pi_{\mathrm{GSI}}}[r^*] \leq \frac{\|r^*\|_\infty}{\sqrt{n}}\left[p(u)^{\frac{1}{2}} e^{\beta\|r\|_\infty}\left(\chi^2(\pi_B\|\pi_S)+1\right)^{\frac{1}{2}} + (1-p(u))\left(\mathrm{CV}(e^{\beta r})^2+1\right)^{\frac{1}{2}}\right],$$

where $\mathrm{CV}(e^{\beta r}) = \sqrt{\frac{\mathrm{Var}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]}{\left(\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]\right)^2}}$. *In particular, we have* $\mathbb{E}_{\pi_{\mathrm{GSI}}}[r^*] - \mathbb{E}_{\pi_{\beta,B}}[r^*] \xrightarrow{n\to\infty} 0$.

Proof. Denote by $Y_{\geq} \subset Y$ the set where $\tilde{r}_t \geq u$, i.e. where $\pi_{\mathrm{GSI}} = \pi_{\beta,S}^{\tilde{r},n}$, and let $Y_< = Y \setminus Y_{\geq}$, hence $\pi_{\mathrm{GSI}} = \pi_{\beta,B}^{r,n}$ on $Y_<$. We write

$$\mathbb{E}_{\pi_{\beta,B}}[r^*] - \mathbb{E}_{\pi_{\mathrm{GSI}}}[r^*] = \underbrace{\mathbb{E}_{\pi_{\beta,B}}[\mathbb{1}_{Y_{\geq}} r^*] - \mathbb{E}_{\pi_{\mathrm{GSI}}}[\mathbb{1}_{Y_{\geq}} r^*]}_{\text{(a)}} + \underbrace{\mathbb{E}_{\pi_{\beta,B}}[\mathbb{1}_{Y_<} r^*] - \mathbb{E}_{\pi_{\mathrm{GSI}}}[\mathbb{1}_{Y_<} r^*]}_{\text{(b)}}.$$

Step 1: Bounding (a). We have by Cauchy-Schwarz:

$$\begin{aligned}
\text{(a)} &= \mathbb{E}_{y\sim\pi_{\beta,B}(\cdot|x)}\left[\mathbb{1}_{Y_{\geq}}(y)\, r^*(x,y)\right] - \mathbb{E}_{y\sim\pi_{\beta,S}^{\tilde{r},n}(\cdot|x)}\left[\mathbb{1}_{Y_{\geq}}(y)\, r^*(x,y)\right] \\
&\leq \|r^*\|_\infty \left(\int \mathbb{1}_{Y_{\geq}}(y)\,\mathrm{d}\pi_{\beta,S}^{\tilde{r},n}(y\mid x)\right)^{\frac{1}{2}} \left(\int\left(\frac{\pi_{\beta,B}(y\mid x) - \pi_{\beta,S}^{\tilde{r},n}(y\mid x)}{\pi_{\beta,S}^{\tilde{r},n}(y\mid x)}\right)^2 \pi_{\beta,S}^{\tilde{r},n}(\mathrm{d}y\mid x)\right)^{\frac{1}{2}} \\
&= \|r^*\|_\infty \left(\pi_{\beta,S}^{\tilde{r},n}(Y_{\geq}\mid x)\right)^{\frac{1}{2}} \left(\chi^2\big(\pi_{\beta,B}(\cdot\mid x)\,\big\|\,\pi_{\beta,S}^{\tilde{r},n}(\cdot\mid x)\big)\right)^{1/2}. \qquad (4)
\end{aligned}$$

By Lemma 1 from (Verdun et al., 2025) we have

$$
\begin{aligned}
&\chi^2\big(\pi_{\beta,B}(\cdot \mid x) \,\big\|\, \pi_{\beta,S}^{\tilde{r},n}(\cdot \mid x)\big) \qquad (5)\\
&= \int \frac{\pi_{\beta,B}(y \mid x)^2}{\pi_{\beta,S}^{\tilde{r},n}(y \mid x)}\,\mathrm{d}y \;-\; 1\\
&= \int \frac{\big(\pi_B(y \mid x)\, e^{\beta\, r(x,y)}\big)^2}{\big(\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2\, \pi_{\beta,S}^{\tilde{r},n}(y \mid x)}\,\mathrm{d}y \;-\; 1\\
&\overset{\text{Lemma 1}}{\leq} \int \frac{\big(\pi_B(y \mid x)\, e^{\beta r(x,y)}\big)^2}{\big(\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2}\, \frac{\frac{1}{n}\frac{\pi_B(y|x)}{\pi_S(y|x)} e^{\beta r(x,y)} + \frac{n-1}{n}\,\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]}{\pi_B(y \mid x)\, e^{\beta r(x,y)}}\,\mathrm{d}y \;-\; 1\\
&= \frac{1}{n\,\big(\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2} \int \frac{\pi_B(y \mid x)^2}{\pi_S(y \mid x)}\, e^{2\beta r}\,\mathrm{d}y + \frac{n-1}{n}\,\frac{\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]}{\big(\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2} \;-\; 1\\
&\leq \frac{e^{2\beta\|r\|_\infty}}{n\,\big(\mathbb{E}_{y'\sim\pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2}\Big(\chi^2\big(\pi_B(\cdot \mid x)\|\pi_S(\cdot \mid x)\big) + 1\Big) \;-\; \frac{1}{n}\\
&\leq \frac{1}{n}\, e^{2\beta\|r\|_\infty}\Big(\chi^2\big(\pi_B(\cdot \mid x)\|\pi_S(\cdot \mid x)\big) + 1\Big). \qquad (6)
\end{aligned}
$$

Plugging (6) into (4) yields

$$
\begin{aligned}
\text{(a)} \;&\leq\; \|r^*\|_\infty \,\Big(\pi_{\beta,S}^{\tilde{r},n}(Y_{\geq} \mid x)\Big)^{\frac{1}{2}} \left(\frac{1}{n}\, e^{2\beta\|r\|_\infty}\,\big(\chi^2(\pi_B\|\pi_S) + 1\big)\right)^{\frac{1}{2}}\\
&=\; \frac{\|r^*\|_\infty}{\sqrt{n}}\, p(u)^{\frac{1}{2}} e^{\beta\|r\|_\infty}\,\big(\chi^2(\pi_B\|\pi_S) + 1\big)^{\frac{1}{2}}. \qquad (7)
\end{aligned}
$$

Step 2: Bounding (b). Similar to the bound for (a), we get

$$
\begin{aligned}
\text{(b)} &= \pi_{\beta,S}^{\tilde{r},n}(\mathbb{1}_{Y<}) \left(\int r^*(x,y)\frac{\pi_{\beta,B}(y \mid x) - \pi_{\beta,B}^{r,n}(y \mid x)}{\pi_{\beta,B}^{r,n}(y \mid x)}\pi_{\beta,B}^{r,n}(\mathrm{d}y \mid x)\right)\\
&\leq \pi_{\beta,S}^{\tilde{r},n}(\mathbb{1}_{Y<}) \left(\int r^*(x,y)^2 \pi_{\beta,B}^{r,n}(\mathrm{d}y \mid x)\right)^{\frac{1}{2}} \left(\int \left(\frac{\pi_{\beta,B}(y \mid x) - \pi_{\beta,B}^{r,n}(y \mid x)}{\pi_{\beta,B}^{r,n}(y \mid x)}\right)^2 \pi_{\beta,B}^{r,n}(\mathrm{d}y \mid x)\right)^{\frac{1}{2}}\\
&\leq \pi_{\beta,S}^{\tilde{r},n}(\mathbb{1}_{Y<})\, \|r^*\|_\infty \left(\int \left(\frac{\pi_{\beta,B}(y \mid x) - \pi_{\beta,B}^{r,n}(y \mid x)}{\pi_{\beta,B}^{r,n}(y \mid x)}\right)^2 \pi_{\beta,B}^{r,n}(\mathrm{d}y \mid x)\right)^{\frac{1}{2}}\\
&= (1 - p(u))\, \|r^*\|_\infty \left(\chi^2(\pi_{\beta,B}||\pi_{\beta,B}^{r,n})\right)^{\frac{1}{2}} \qquad (8)
\end{aligned}
$$

by independence of the event $Y_<$ and π_B^n resp. $\pi_{\beta,B}$, and applying Cauchy-Schwarz.

Again, using Lemma 1 from (Verdun et al., 2025) we get

$$\begin{aligned}
\chi^2(\pi_{\beta,B}||\pi^{r,n}_{\beta,B}) &= \int \frac{\pi_{\beta,B}(y \mid x)^2}{\pi^{r,n}_{\beta,B}(y \mid x)}\, dy \;-\; 1 \\
&\overset{\text{Lemma 1}}{\leq} \int \frac{\big(\pi_B(y \mid x)\, e^{\beta r(x,y)}\big)^2}{\big(\mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2} \frac{\frac{1}{n}\, e^{\beta r(x,y)} + \frac{n-1}{n}\, \mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]}{\pi_B(y \mid x)\, e^{\beta r(x,y)}}\, dy \;-\; 1 \\
&= \frac{1}{n} \frac{\mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{2\beta r(x,y')}]}{\big(\mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2} + \frac{n-1}{n} - 1 \\
&\leq \frac{1}{n} \frac{\mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{2\beta r(x,y')}]}{\big(\mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2} \\
&= \frac{1}{n} \left(\frac{\mathrm{Var}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]}{\big(\mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2} + 1 \right). \qquad (9)
\end{aligned}$$

Plugging equation (9) into (8) yields

$$\text{(b)} \leq \frac{\|r^*\|_\infty}{\sqrt{n}} (1 - p(u)) \left(\frac{\mathrm{Var}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]}{\big(\mathbb{E}_{y' \sim \pi_B(\cdot|x)}[e^{\beta r(x,y')}]\big)^2} + 1 \right)^{\frac{1}{2}} \qquad (10)$$

Combining equations (7) and (10) gives

$$\mathbb{E}_{\pi_{\beta,B}}[r^*] - \mathbb{E}_{\pi_{\text{GSI}}}[r^*] \leq \frac{\|r^*\|_\infty}{\sqrt{n}} \left[p(u)^{\frac{1}{2}} e^{\beta \|r\|_\infty} \left(\chi^2(\pi_B \| \pi_S) + 1\right)^{\frac{1}{2}} + (1 - p(u)) \left(\mathrm{CV}(e^{\beta r})^2 + 1\right)^{\frac{1}{2}} \right]$$

as desired.

□