Demystifying LLM-as-a-Judge

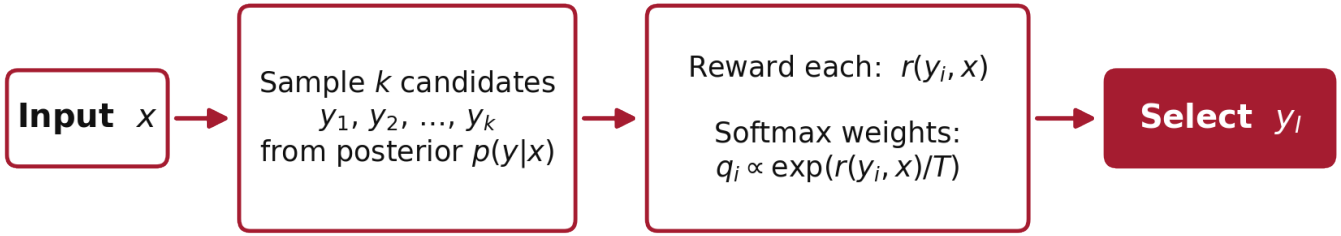
Analytically Tractable Model for Inference-Time Scaling

Indranil Halder · **Cengiz Pehlevan**

Harvard John A. Paulson School of Engineering And Applied Sciences

ICML 2026

Inference-time scaling: three open questions



At test-time, judge LLM reward-weighted reranking of candidate answers are now standard. Open questions remain regarding the number of inference time samples k, and the reranking softmax temperature T depending on the quality of the judge LLM:

Q1 Imperfect judge: How large should k be?

Does more sampling always help or is there a finite optimal k ?

Q2 Imperfect judge: How to choose T?

best-of-k *(T=0)* vs uniform (T=∞)

Is there an optimal T, how does that change with quality of the judge?

Q3 Perfect judge: Best-of-k scaling law when reasoning is taken into account

Even with very high quality judge how fast can inference time scaling improve generalization?

General framework: any predictive, any reward

Setup. Arbitrary predictive $p_x(y)$, reward $r(y,x)$, task loss $l(y,x)$. Define the reward-weighted reranked loss:

$$\delta_k^{(\ell,r)}(\mathbf{x},T) = \mathbb{E}_{Y_{1:k}\sim p_{\mathbf{x}}}\left[\frac{\sum_{i=1}^{k}\ell(Y_i,\mathbf{x})\,e^{r(Y_i,\mathbf{x})/T}}{\sum_{j=1}^{k}e^{r(Y_j,\mathbf{x})/T}}\right]$$

Result. Expanding in $1/T$:

$$\delta_k^{(\ell,r)}(\mathbf{x},T) = \mu_\ell(\mathbf{x}) + \frac{k-1}{k}\frac{1}{T}\,\mathrm{Cov}_x(\ell,r) + \frac{(k-1)(k-2)}{2k^2}\frac{1}{T^2}\mathrm{Cov}_x\big(\ell,(r-\mu_r(\mathbf{x}))^2\big) \quad \dots$$

where $\mu_\ell(\mathbf{x}) := \mathbb{E}_{p_{\mathbf{x}}}[\ell(Y,\mathbf{x})]$, $\mu_r(\mathbf{x}) := \mathbb{E}_{p_{\mathbf{x}}}[r(Y,\mathbf{x})]$, and Cov_x denotes covariance under $Y\sim p_{\mathbf{x}}$.

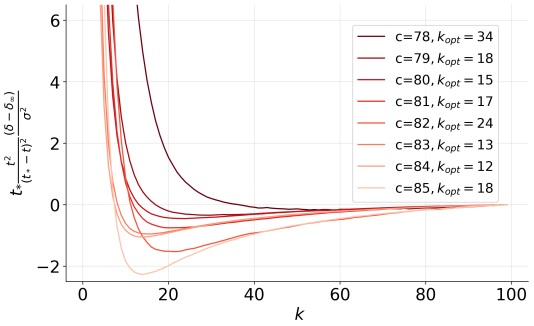
- Reranking is beneficial precisely when $Cov_x(l,r) < 0$ that is, when a larger reward is correlated with smaller task loss.
- If $Cov_x\big(l,(r-\mu_r)^2\big) > 0$, higher-loss draws tend to occur where the reward fluctuates more strongly around its mean.
  In that case, aggressive reranking can over-amplify reward noise or reward misspecification, that can generate a finite optimal temperature or a finite optimal sample budget.

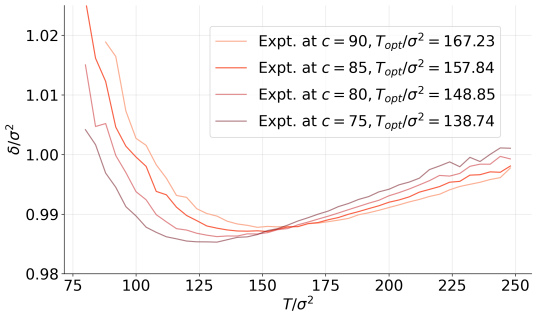
Predictions from the solvable model

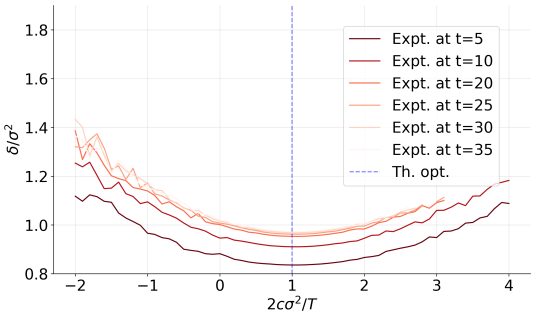
Optimal k exists
when the judge is misaligned

Optimal T exists
for the reward sampler

Best judge ≠ teacher
optimal reward differs from generator







$\exists\, k_{\text{opt}}$ $\exists\, T_{\text{opt}}$ $\mathbf{w}_R^{\text{opt}} \neq \mathbf{w}_T$

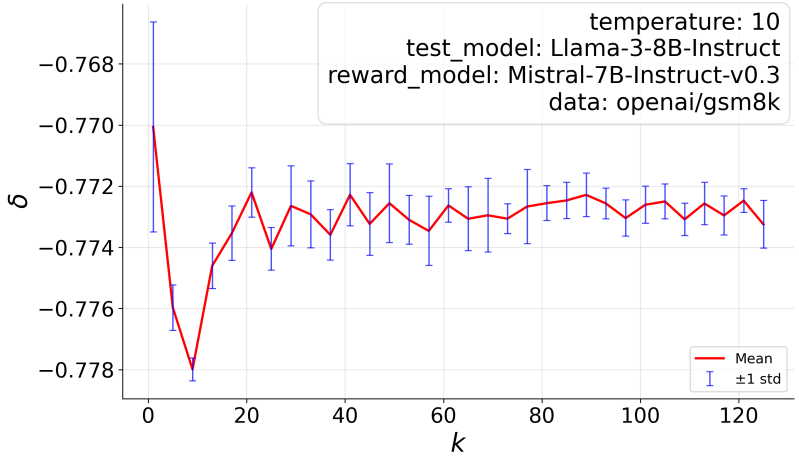
BLR: $\ell = (y-\mu_T)^2, \quad r = -(y-\mu_R)^2, \quad p_{\mathbf{x}} = \mathcal{N}(m, s^2)$

$$y = w_T \cdot \frac{x}{\sqrt{d}} + \eta, \quad x \sim \mathcal{N}(0, S^2 I), \quad \eta \sim \mathcal{N}(0, \sigma^2), \quad \mu_R = \frac{1}{\sqrt{d}} w_R \cdot x, \quad \mu_T = \frac{1}{\sqrt{d}} w_T \cdot x$$

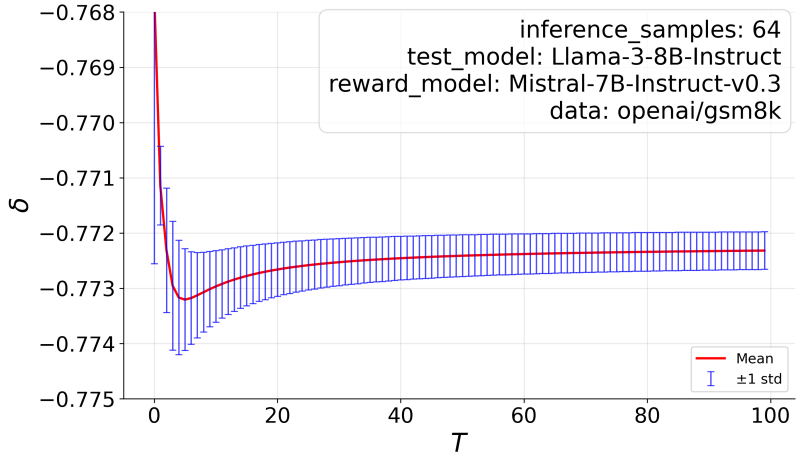
Predictions verified in Llama-3-8B-Instruct on GSM8K

Reward model: Mistral-7B-Instruct-v0.3 · Dataset: openai/gsm8k

δ vs k (fixed T)



δ vs T (fixed k)



Matches the predictions of the theory qualitatively

Best-of-k limit: a coverage-driven general power law

T → 0: reward-weighted sampler → best-of-*k* : $Y_{(R)}$

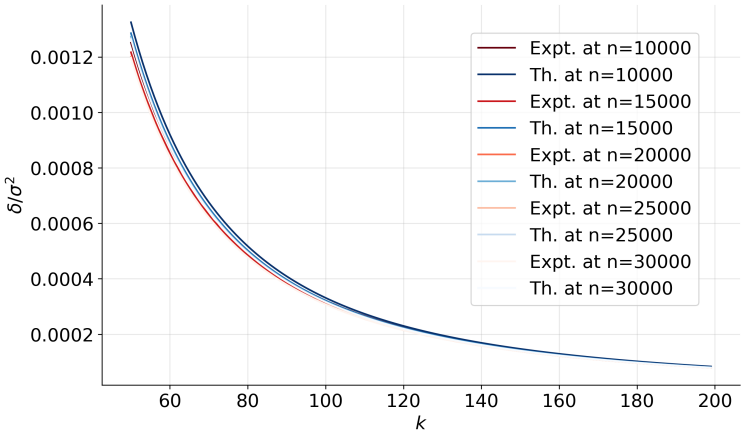
Local behavior of the predictive density near the reward target $\mu_T = \mu_R$:

$$p_{\mathbf{x}}(\mu_R + u) \sim c_{R,\mathbf{x}} |u|^{\beta} \quad \text{as } u \to 0$$

Then via extreme-value theory:

$$\mathbb{E}\left[\left(Y_{(R)} - \mu_T(\mathbf{x}) \right)^2 \mid \mathbf{x} \right] \asymp k^{-\frac{2}{1+\beta}}$$

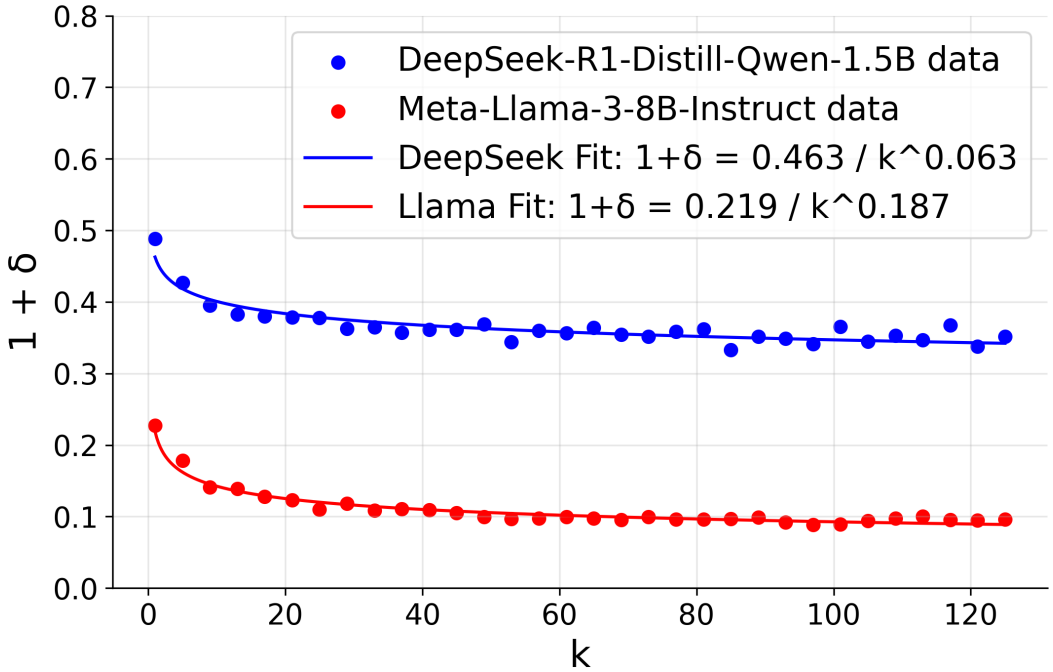
BLR experiment (β = 0)



Better coverage (small β), faster decay of best-of-k generalization error (δ) for the ideal reward model

Power law fits LLM best-of-k data

Same functional form as theory fits real LLM best-of-k experiments when reasoning is taken into account



Fit form

$$1 + \delta = \frac{A}{k^{\,2/(1+\beta)}}$$

Extracted exponents

Llama-3-8B: **0.187** *(β ≈ 10)*

DeepSeek-R1-1.5B: **0.063** *(β ≈ 31)*

Different from pass@k that
Only takes into account the final prediction accuracy

Takeaways

1. **A solvable model of LLM-as-a-Judge**
   Bayesian linear regression + quadratic reward + softmax sampler

2. **Imperfect judge theoretical predictions, verified in LLMs**
   Optimal k · Optimal T

3. **Best-of-k power law for nearly perfect judge, verified in LLMs**
   Coverage near the reward target sets the scaling exponent.

Thank you!