Reasoning with Sampling: Your Base Model is Smarter Than You Think

Presented by: Aayush Karan

Aayush Karan

Yilun Du

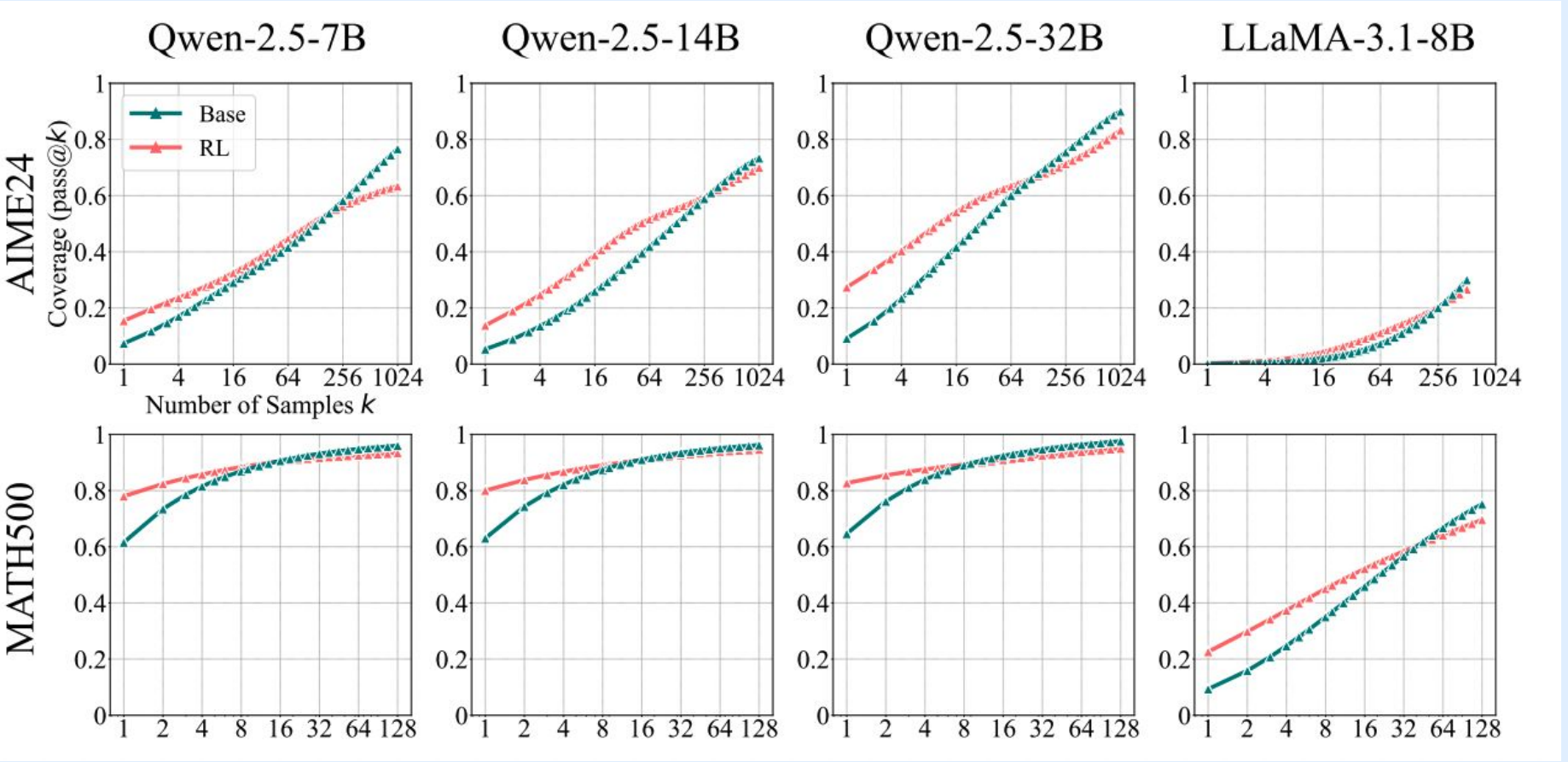
Frontier Reasoning Models

- Reinforcement learning for LLMs:
  - Policy gradient with a (verifiable) reward signal on the base model (GRPO/RLVR)
  - Massive boosts in math, coding, and other STEM reasoning tasks

Frontier Reasoning Models

- Reinforcement learning for LLMs:
  - Policy gradient with a (verifiable) reward signal on the base model (GRPO/RLVR)
  - Massive boosts in math, coding, and other STEM reasoning tasks

- Downsides to posttraining:
  - Requires: **verifiability**, curated **dataset**, **training** (susceptible to instabilities)
  - Generation diversity at multiple samples collapses
  - Multi-shot (pass@k) reasoning performance can be **worse** than the base model
  - Posttrained models are generally **uncalibrated** and **overconfident**

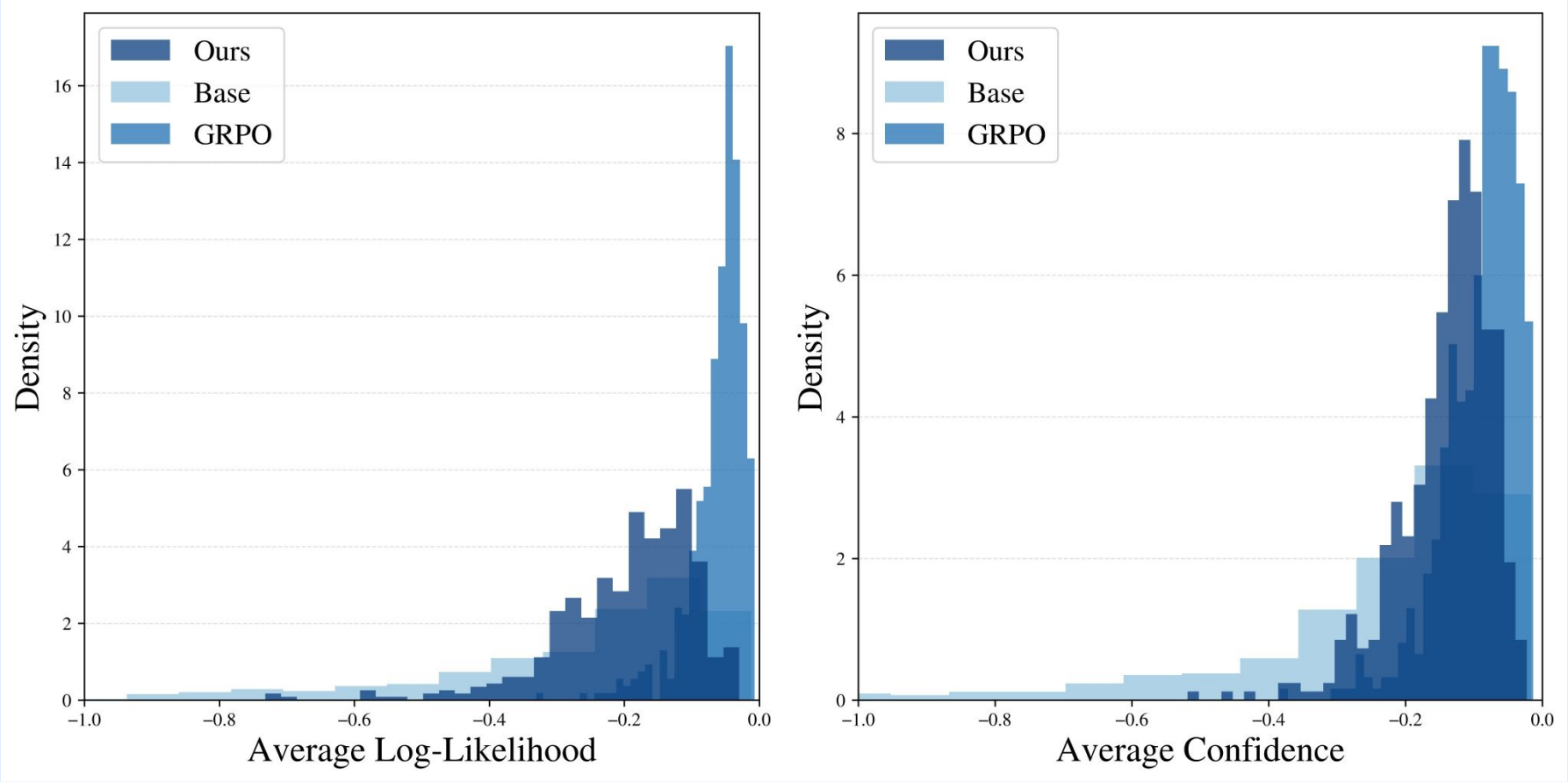


Yue et. al. 2025

Distribution Sharpening

- Is RL just “sharpening” the base model distribution*?
    - Pass@k upper bounded by the base model implies generations are within base model capabilities

Distribution Sharpening



Distribution Sharpening

- Is RL just “sharpening” the base model distribution*?
    - Pass@k upper bounded by the base model implies generations are within base model capabilities
    - Likelihoods of RL outputs have low perplexity/high confidence under the base model distribution

Distribution Sharpening

- Is RL just “sharpening” the base model distribution*?
    - Pass@k upper bounded by the base model suggests generations are within base model capabilities
    - Likelihoods of RL outputs have low perplexity/high confidence under the base model distribution

**Sharpening dominance can depend on the setting*

Sampling directly from the base model can achieve single-shot reasoning on par with RL.

Sampling directly from the base model can achieve single-shot reasoning on par with RL.

This **does not** come at the expense of generation diversity or multi-shot performance.

Sampling directly from the base model can achieve single-shot reasoning on par with RL.

This **does not** come at the expense of generation diversity or multi-shot performance.

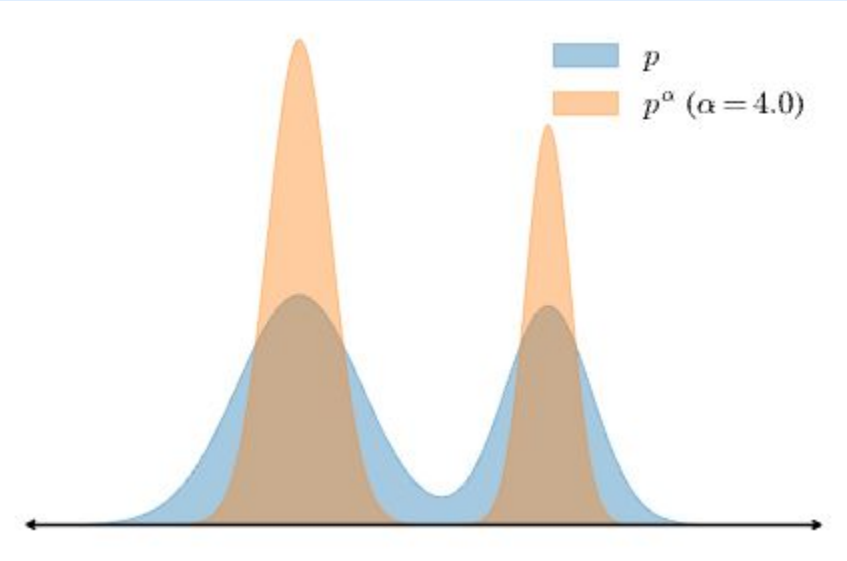
Pure sampling approach: no training, no verification, and no datasets.

Power Distributions

If RL induces a sharpening effect on the base model distribution, we should independently be able to write down an explicit sharpening and sample from it.

Power Distributions

- A natural way to sharpen any given distribution is to *exponentiate:*
    - Increases relative weight on high-likelihood sequences and decreases weight on low-likelihood sequences



$$p(\mathbf{x}) > p(\mathbf{x}') \implies \frac{p(\mathbf{x})^\alpha}{p(\mathbf{x}')^\alpha} > \frac{p(\mathbf{x})}{p(\mathbf{x}')}$$

Power Distributions

This is NOT token-level low-temperature sampling!

Power Distributions

This is NOT token-level low-temperature sampling!

$$p_{\text{temp}}(x_t|x_{<t}) \propto \left(\sum_{x_{>t}} p(x_0, \ldots, x_t, \ldots, x_T)\right)^{\alpha}$$

$$p_{\text{pow}}(x_t|x_{<t}) \propto \sum_{x_{>t}} p(x_0, \ldots, x_t, \ldots, x_T)^{\alpha}$$

Power Distributions

- Low-temperature sampling averages all future paths *before* sharpening
  - Favors sampling tokens with many low-likelihood future paths instead of few high-likelihood ones

Power Distributions

- Low-temperature sampling averages all future paths *before* sharpening
  - Favors sampling tokens with many low-likelihood future paths instead of few high-likelihood ones

- Power sampling sharpens future paths *and then averages*
  - Downweights low-likelihood completions first
  - Favors sampling tokens with few but high likelihood future completions

Power Distributions

- Low-temperature sampling averages all future paths *before* sharpening
  - Favors sampling tokens with many low-likelihood future paths instead of few high-likelihood ones

- Power sampling sharpens future paths *and then averages*
  - Downweights low-likelihood completions first
  - Favors sampling tokens with few but high likelihood future completions

- **Power distributions are better for reasoning**
  - Inductive bias for “planning” by avoiding sampling tokens* that trap generations in low-likelihood futures

*Li, **Karan**, Chen (2025). Blink of an eye: a simple theory for feature localization in generative models*

Power Sampling

- Power distributions are motivated targets, but how do we sample from them?

Power Sampling

- Power distributions are motivated targets, but how do we sample from them?
  - Generally intractable: normalizing over all future paths is computationally hard

Power Sampling

- Power distributions are motivated targets, but how do we sample from them?
  - Generally intractable: normalizing over all future paths is computationally hard

- Metropolis-Hastings
  - Approximate sampling algorithm from unnormalized distributions
  - Builds chain of sequences that eventually converges to samples from the power distribution
  - Uses a proposal distribution to generate new candidate samples, and then accepts or rejects them:

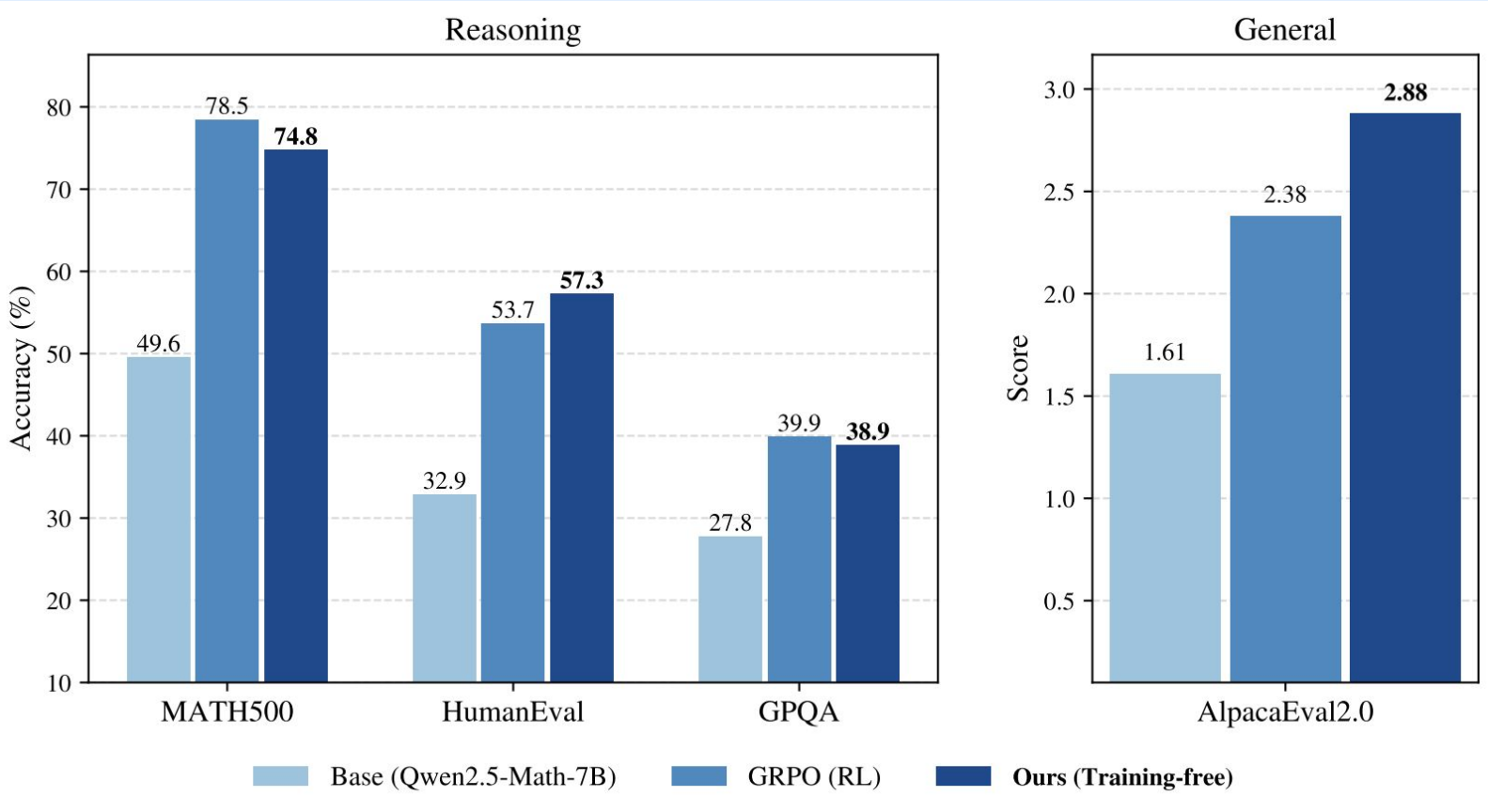
$$A(\mathbf{x}, \mathbf{x}^i) = \min\left\{1, \frac{p^\alpha(\mathbf{x}) \cdot q(\mathbf{x}^i|\mathbf{x})}{p^\alpha(\mathbf{x}^i) \cdot q(\mathbf{x}|\mathbf{x}^i)}\right\}$$

Question: If a square has side length four, what is its area?

Results

Power sampling nearly matches, if not outperforms, GRPO across various domains and base models.

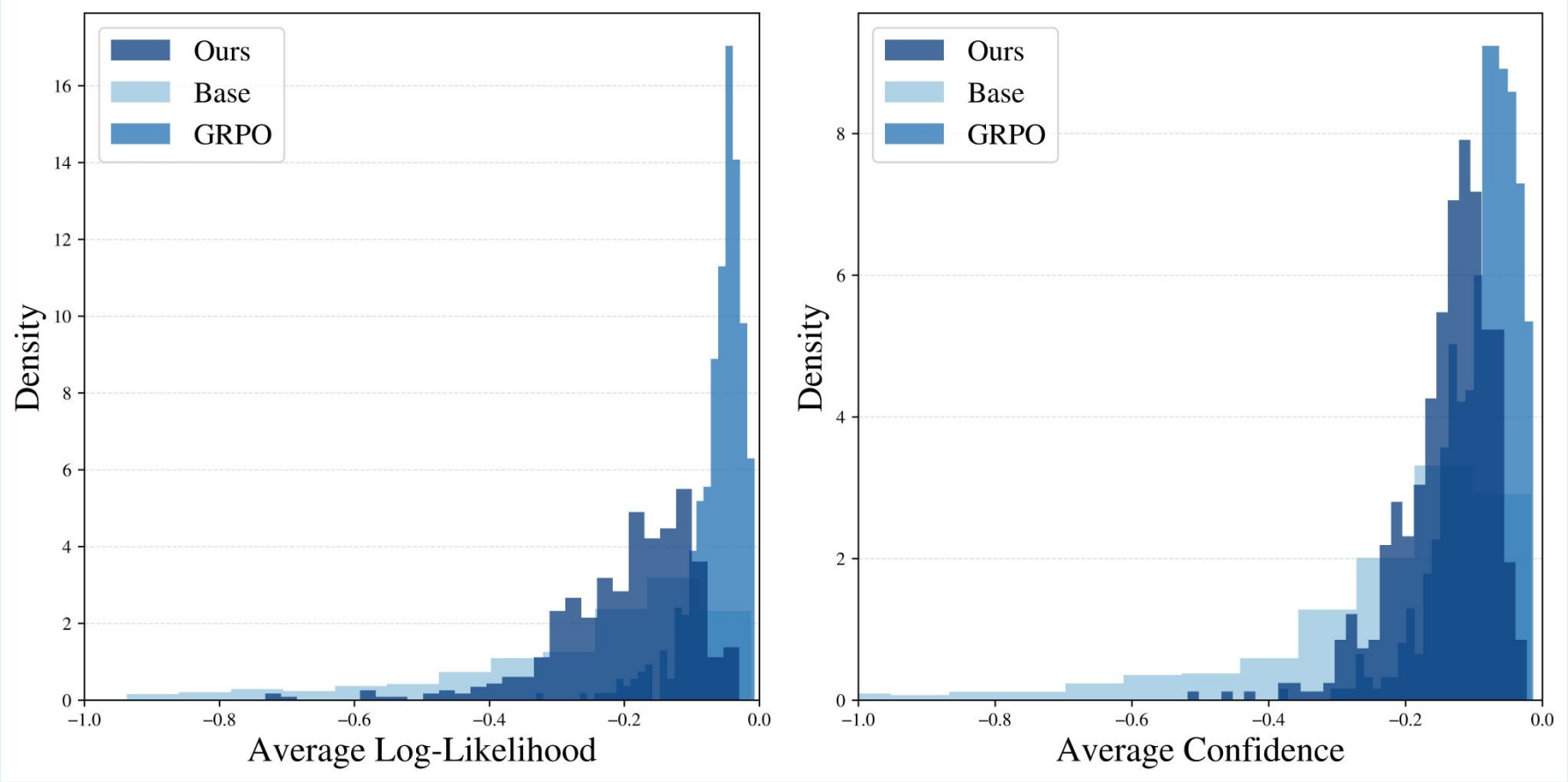
Single-shot Reasoning



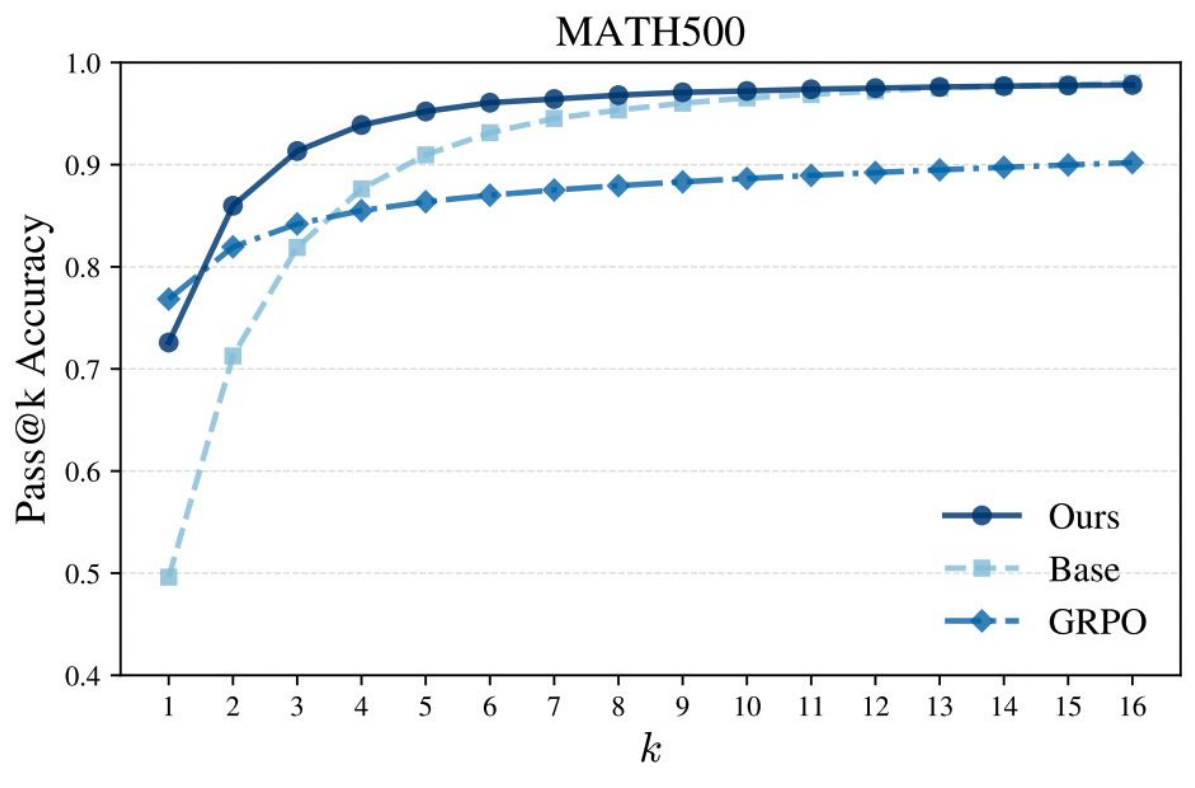
Results

Power sampling avoids the diversity collapse of RL, and, in fact, exceeds GRPO and the base model in pass@k performance.

Diversity and Multi-shot Reasoning



Diversity and Multi-shot Reasoning



Takeaways

Sampling as a Primitive: By treating LLMs as adaptable distributions, we can elicit strong yet apparently hidden capabilities directly from the base model, *without training.*

Takeaways

Sampling as a Primitive: By treating LLMs as adaptable distributions, we can elicit strong yet apparently hidden capabilities directly from the base model, *without training.*

Rich Distributional Reasoning: Sampling demonstrates the existence of an ideal reasoning distribution: one with both strong single-shot performance as well as multi-sample diversity.