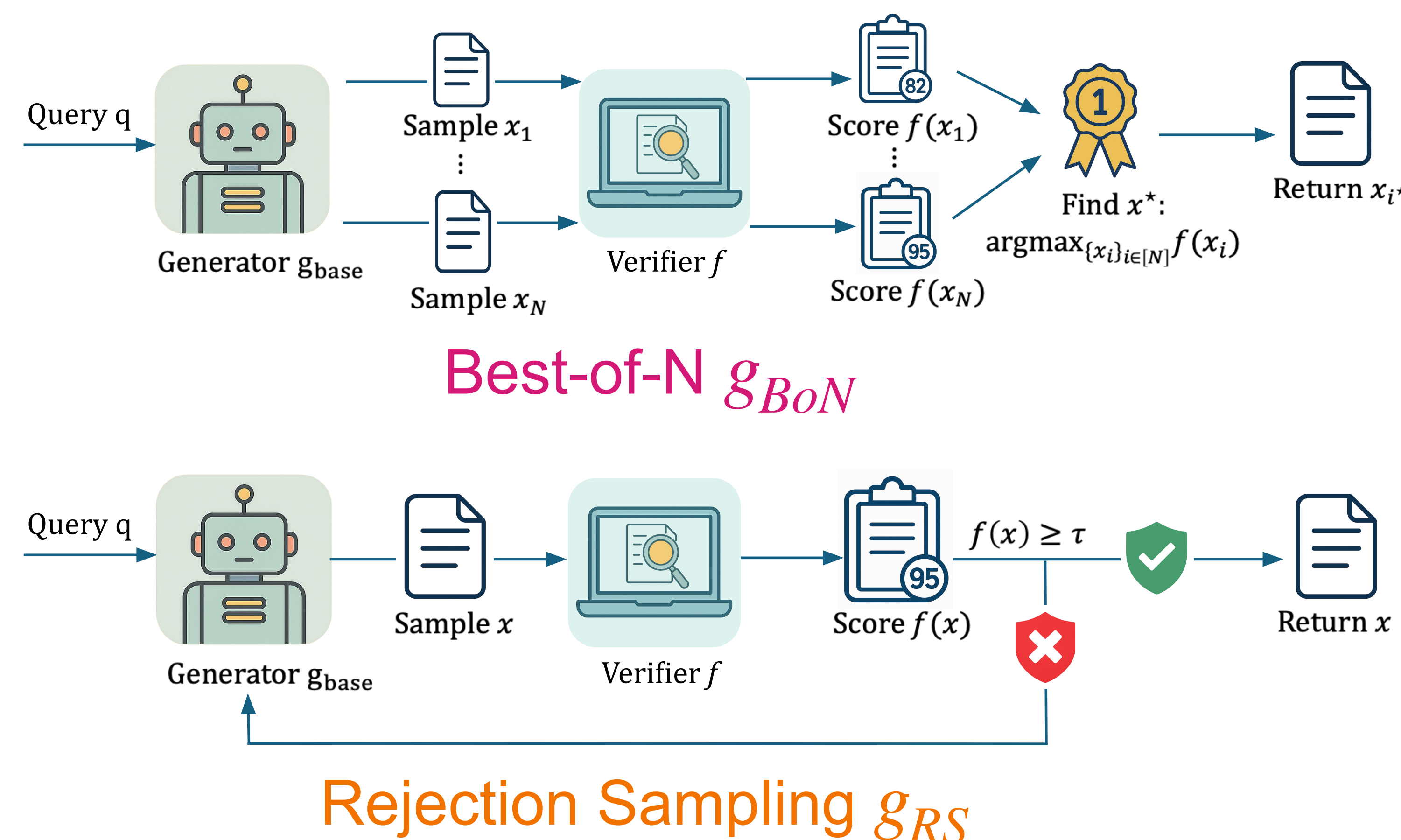


- **Test-time scaling:** Spend additional resources during inference to improve LLM performance
 - Simple approach: Sample multiple answers, select correct ones
 - Selection requires a **verifier** — but real verifiers are imperfect proxies for correctness
- **We characterise how verifier quality shapes test-time scaling**

Resampling Approaches



Key Quantities

- Ground truth score $y(x) \in \{0,1\}$
- Proxy score $f(x)$
- TPR/FPR: $P_{x \sim g}[f(x) \geq \tau | y = 0/1]$
- Generative accuracy $\text{ACC}(g) = P_{x \sim g}[y(x)]$
- Initial accuracy $\pi = \text{ACC}(g_{base})$
- ROC curve $\text{TPR}(\text{FPR})$: Plot TPR against FPR

How does Test-Time Accuracy $\text{Acc}(g_{RS})/\text{Acc}(g_{BoN})$ scale with #samples?

Scaling Laws

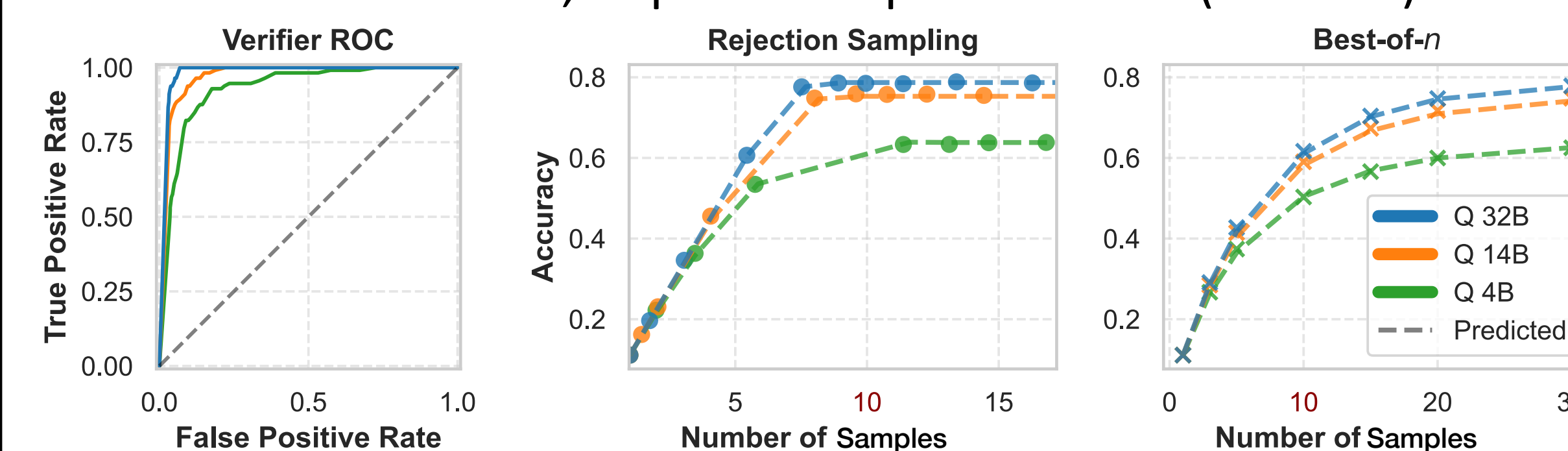
- Knowing the ROC curve, we can calculate Best-of-N and Rejection Sampling accuracy!
- **Infinite sample limit:** For $\#samples \rightarrow \infty$, $\text{Acc}(g_{RS})$ and $\text{Acc}(g_{BoN})$ both converge to $\frac{\pi \text{TPR}'(0)}{\pi \text{TPR}'(0) + 1 - \pi}$
 - The same goes for the optimal RL policy as KL penalty $\beta \rightarrow 0$.
- **Concave ROC $\Rightarrow$ monotonous scaling:** Whenever the ROC curve is concave, both $\text{Acc}(g_{RS})$ and $\text{Acc}(g_{BoN})$ increase in #samples

Key Takeaways

- **Verifier quality as key bottleneck:** $\text{Acc}(g_{RS})$ and $\text{Acc}(g_{BoN})$ usually < 1 .
- **Extrapolation is hard:** For known accuracy for $\#samples \leq S$, there are ϵ -consistent ROC curves with arbitrary accuracy as $\#samples \rightarrow \infty$
- **Rejection Sampling beats Best-of-N:** For any number of (expected) samples, $\text{Acc}(g_{RS}) > \text{Acc}(g_{BoN})$ under mild assumptions. But Best-of-N is more practical!

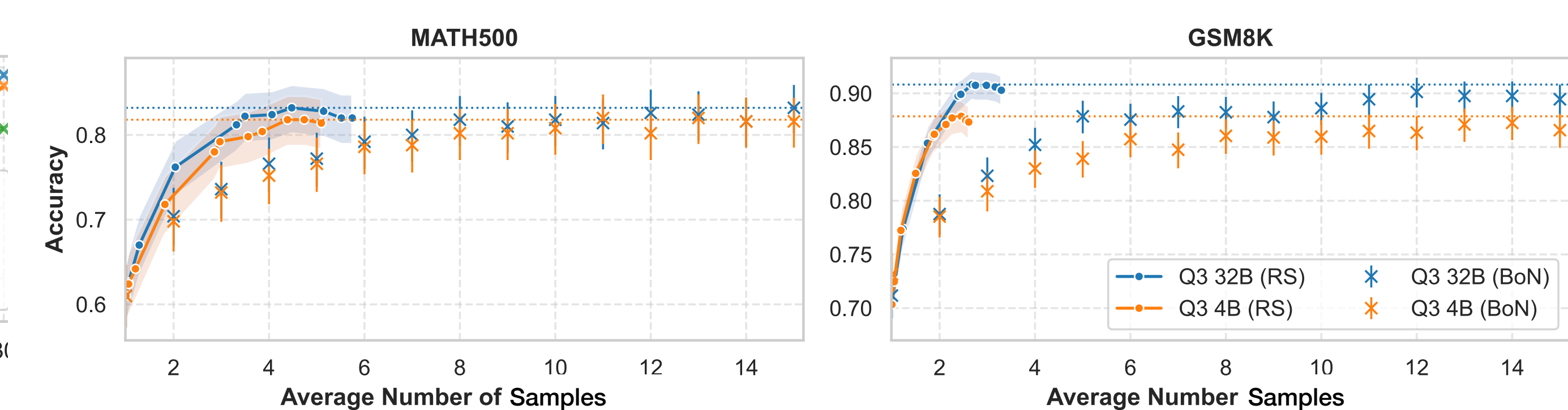
Empirical Results

Accuracy (lines) of RS (middle) and BoN (right) for Qwen3-1.7B on GSM8K item 58, vs predicted performance (markers).



Performance extrapolation is hard!

Aggregate Accuracy of RS (lines) and BoN (markers) for Qwen3-1.7B on MATH500 (left) and GSM8K (right)



Rejection Sampling is more efficient!

