Can we scale small LMs to o1 performance?
A Probabilistic Inference Approach to LLM Inference-Time Scaling

Isha Puri, MIT CSAIL

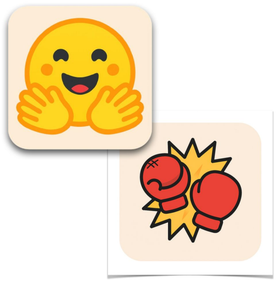
MIT EECS

Intro to Inference Scaling

Closed source models, through high performing, have a variety of privacy and cost (monetary and energy-wise) restrictions.

What if I told you that you could take a random model off of Huggingface, and with 16 or 32 samples, scale its accuracy by up to 40 percentage points? Without any effort from you?

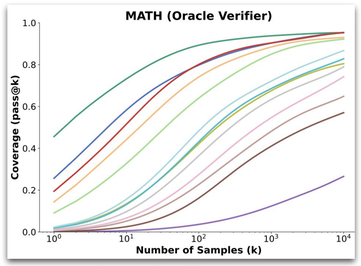
Inference scaling allows you to do exactly this. We want to be able to take models and get them to **punch above their weight class**!



If we can get small models to perform at the level of large models, this will open up a world of possibilities.

One can now get the **same performance for a lot cheaper**, unlocking access to language models for those who are in academia/startups/resource constrained environments. Those with privacy restraints can also unlock LM potential.

Now, studies have shown that smaller, much cheaper, open models - even those as small as 1B models - when queried several times, will often eventually correctly answer challenging reasoning questions.



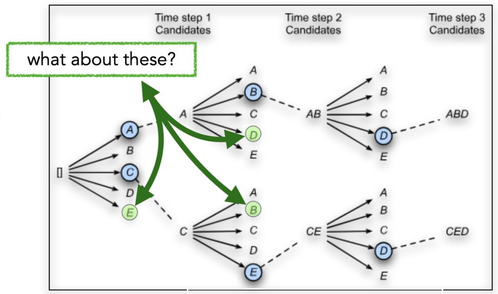
(from Large Language Monkeys: Scaling Inference Compute with Repeated Sampling, Brown et al., 2024)

so... if we know that small language models have it in them to answer difficult questions,
we just have to find a way to squeeze it out of them!

Background and Motivation

following our PRM **blindly** to determine which partial answers we want to continue expanding during our reasoning process can lead to **reward hacking**

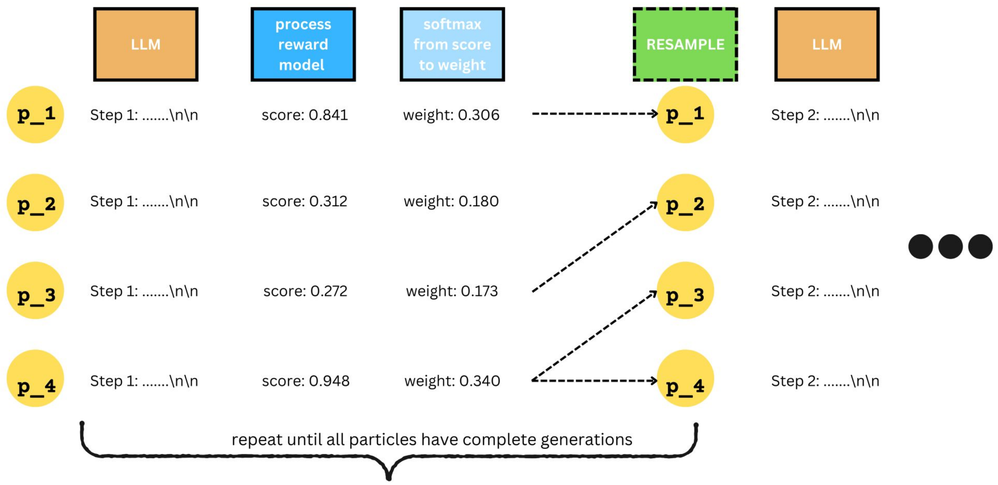
where the final output is optimized to score well according to the reward model but fails to be useful and/or correct



Beam Search ↑

Method

PARTICLE FILTERING FOR INFERENCE SCALING



Algorithm 1 Particle Filtering for Inference-Time Scaling

Input: the number of particles N, a reward model $\hat{r}$, a LLM p_M and the prompt c
Initialize N particles $\{x_1^{(i)} \sim p_M(\cdot \mid c)\}_{i=1}^N$
$t \leftarrow 1$
while not all particles stop **do**
  Update rewards $\mathbf{w} = [\hat{r}(x_{1:t}^{(1)}), \ldots, \hat{r}(x_{1:t}^{(N)})]$
  Compute softmax distribution $\theta = \text{softmax}(\mathbf{w})$
  Sample indices $\{j_t^{(i)}\}_{i=1}^N \sim \mathbb{P}_t(j = i) = \theta_i$
  Update the set of particles as $\{x_{1:t}^{(j_t^{(i)})}\}_{i=1}^N$
  Transition $\{x_{t+1}^{(i)} \sim p_M(\cdot \mid c, x_{1:t}^{(i)})\}_{i=1}^N$
  $t \leftarrow t + 1$
end while
Return: the set of particles in the end

"translated"

now, we do this entire thing again!

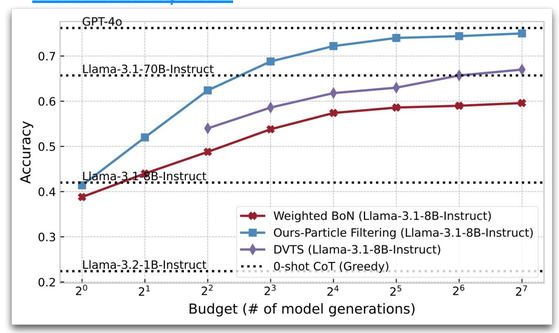
1. each particle generates its "next step"
2. we use the PRM (Process Reward Model) to calculate a score using the question and the entire answer generated by that particle so far
3. we convert that score to a weight with softmax
4. we then resample the particles according to those weights

we continue doing this until every single particle has generated an "end of sequence" token, and thus, finished its answer!

Results

Model	Method	MATH500	AIME 2024
Closed-Source LLMs			
GPT-4o	-	76.2	13.3
o1-preview	-	***87.0***	40.0
Claude3.5-Sonnet	-	78.3	16.0
Open-Source LLMs			
Llama-3.1-70B-Instruct	-	65.7	16.6
Qwen2.5-Math-72B-Instruct	-	82.0	30.0
Open-Source SLMs			
Llama-3.2-1B-Instruct	Pass@1	26.8	0.0
	Ours - PF	59.6	10.0
Llama-3.1-8B-Instruct	Pass@1	49.9	6.6
	Ours - PF	74.4	16.6
phi-4	Pass@1	79.8	16.6
	Ours - PF	83.6	26.6
Mistral-Small-24B-Instruct-2501	Pass@1	69.2	10.0
	Ours - PF	83.4	23.3
Qwen2.5-32B-Instruct	Pass@1	82.8	16.6
	Ours - PF*	***89.9***	***43.3***
Open-Source Math SLMs			
Qwen2.5-Math-1.5B-Instruct	Pass@1	70.0	10.0
	Ours - PF	85.4	23.3
Qwen2.5-Math-7B-Instruct	Pass@1	79.6	16.6
	Ours - PF	87.0	23.3

- A 7B model surpasses o1 accuracy in 32 rollouts
- A 1.5B model surpasses GPT4o in only 4 rollouts
- Our method (PF) has a 4-16x better scaling rate than other inference scaling methods / deterministic counterparts
- All of this is done *without any training at all! Just by intelligently / probabilistically navigating the search space.*



- This inference scaling method allows any *off the shelf* model to punch way above its weight class - just by sampling it a few times!

Coauthors: Shiv Sudalairaj, GX Xu, Kai Xu, Akash Srivastava

Scan for more information! →