A Simple Model of Inference Scaling Laws

Noam Levi

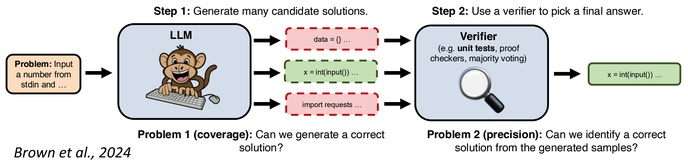
noam.levi@epfl.ch

Check out our paper:

Overview

TL; DR: We introduce a simple theoretical model explaining the observed behavior of LLM performance on reasoning tasks with increasing number of inference attempts.

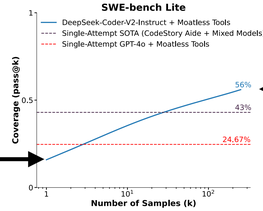
Models improve with more inference attempts



Brown et al., 2024

pass@k (aka Coverage) = the probability of at least one success in k trials averaged over the entire dataset of size n.

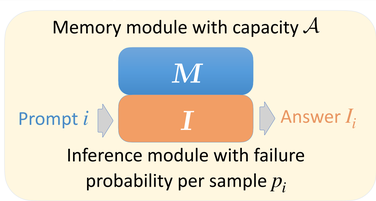
One shot weak models…



…Can become much better with repeated inference attempts!

An ansatz for an inference model that improves with repeated attempts

We assume a simple construction: A memory module M memorizes a set of "solutions" up to a certain capacity. Then the inference module I infers from M with some error probability for each solution given by p_i.



Under the following assumptions, the pass@k is given:

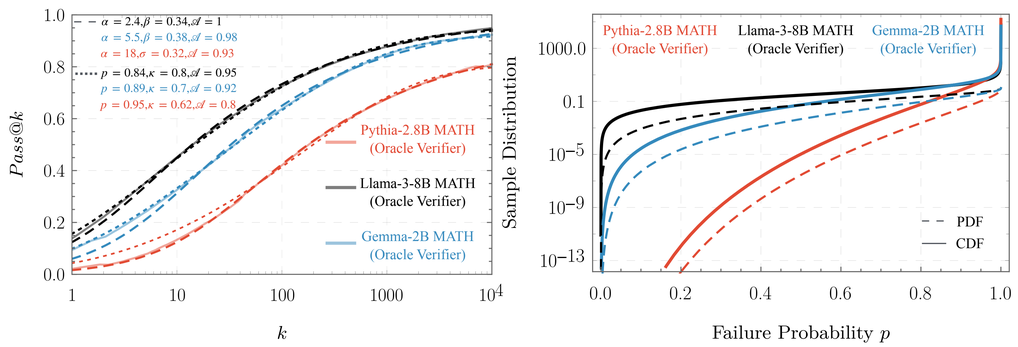
Assumption 3.1. For every sample i, we have access to a perfect verification method, which can determine if there exists a correct generated answer during inference $I(i) = y_i$, among k possible candidates $\{I_1(i), \ldots, I_k(i)\}$.

Assumption 3.2. Inference attempts $\{I_1(i), \ldots, I_k(i)\}$ are independent and identically distributed (i.i.d.) random variables.

$$\text{pass@k} = \frac{n_c}{n} \times \left(1 - \frac{1}{n_c}\sum_{i=1}^{n_c}\prod_{t=1}^{k}(1 - A_t(i))\right) = \mathcal{A} \times \left(1 - \frac{1}{n_c}\sum_{i=1}^{n_c} p_i^k\right)$$

Diversity in *"perceived difficulty"* of tasks determines the inference scaling of a model

***Dashed curves indicate the theory given here, dotted curves are an alternative, dual theory described in the paper**



Left: Empirical pass@k for several LLMs on MATH questions, for varying k repeated inference attempts.
Right: Probability distribution of inference failures required to match theory predictions to the empirical results.

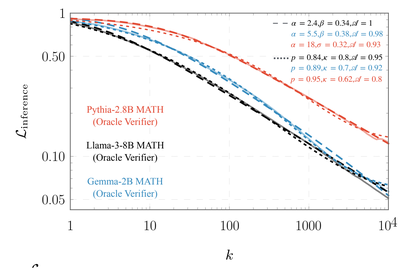
How and why does it work?

Key: all the information is in the distribution of p_i.
To construct the failure distribution, we assume that different samples may have different inference complexity levels, incorporating some "easy" and some "difficult" samples with respect to the inference model.

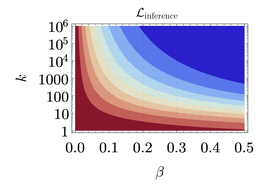
One way to model the different complexities is using the Beta distribution. We think of the failure probability across samples itself $p = p_i$ as a random variable, drawn from

$$\text{Beta}(\alpha, \beta; p) = \frac{p^{-1+\alpha}(1-p)^{-1+\beta}}{B(\alpha, \beta)}$$

Where α controls easy questions and β determines the hard tail.



Inference losses for different models and parameters



Result 1: analytical pass@k given by

$$\text{pass@}k \approx \mathscr{A} \times \left(1 - \langle p^k \rangle\right) = \mathscr{A} \times \left(1 - \int_0^1 dp\, p^k \frac{p^{-1+\alpha}(1-p)^{-1+\beta}}{B(\alpha, \beta)}\right) = \mathscr{A} \times \left(1 - \frac{\Gamma(\beta)\Gamma(k+\alpha)}{B(\alpha, \beta)\Gamma(k+\alpha+\beta)}\right)$$

Result 2: analytical *inference loss:*

$$\mathscr{L}_{\text{inference}}(k) \equiv \mathbb{E}(\text{Error in k trials}) = \mathbb{E}(\mathscr{A} \times p^k) \approx \mathscr{A} \times \frac{\Gamma(\beta)\Gamma(k+\alpha)}{B(\alpha, \beta)\Gamma(k+\alpha+\beta)} \underset{k\to\infty}{\approx} \mathscr{A} \times \frac{\Gamma(\beta)k^{-\beta}}{B(\alpha, \beta)}$$

Power law decay for large *k*!

Interpretability and inference costs

This model can easily connect the measured pass@k curves with the unknown failure probability distribution of different models!

Procedure: measure the pass@k, then define $\tilde{f}(k) = (\mathscr{A} - \text{pass@}k)/\mathscr{A}$. The Laplace transform relates the *"difficulty distribution"* with the pass@k metric, as:

$$\tilde{f}(k) = \langle p^k \rangle = \int_0^\infty d\sigma e^{-\alpha\sigma} \frac{e^{-\alpha\sigma}(1-e^{-\sigma})^{-1+\beta}}{B(\alpha, \beta)} = \int_0^\infty d\sigma e^{-\sigma k} f(\sigma)\,.$$

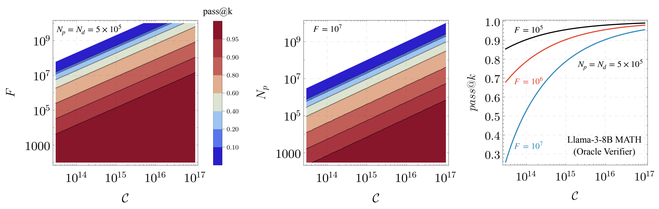
Cost of Compute: We can also relate our scaling laws to a *"compute pass@k"* with repeated inference attempts, given analytically by:

$$\text{Coverage}(\bar{\mathcal{C}}) \approx \mathcal{A} \cdot \left(1 - \frac{\Gamma(\beta)}{B(\alpha, \beta)}\left(\frac{\bar{\mathcal{C}} - N_p}{N_d}\right)^{-\beta}\right)$$

We relate attempts to compute cost through a simple formula:

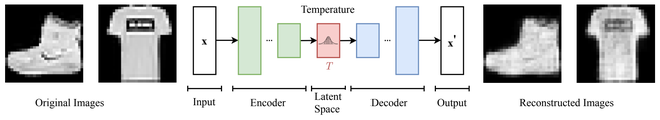
$$\mathcal{C} = \bar{\mathcal{C}} \times F = N_p \times F + N_d \times k \times F$$

where $\mathcal{C}$ is the inference cost per token, N_p, N_d are the number of prompt and decode tokens, respectively, and F is the number of FLOPS per token.



Inference cost for different parameter values

Bonus: Variational Autoencoder reconstruction experiment



Task: We train a VAE with a temperature parameter to reconstruct its training samples (FMNIST), with a failure threshold parameter ε.

Result: The controlled reconstruction task obeys the same type of pass@k scaling, indicating some universality of the simple model.