Paper for Discussion: 03/26/2026

Cesar Cervera

Paper: Dan Hendrycks et al., *Measuring Mathematical Problem Solving With the MATH Dataset*, 2021.

The general topic of this paper is evaluating mathematical reasoning in large language models. The authors introduce a new benchmark called the MATH dataset, which contains 12,500 competition-level mathematics problems.

The specific activity studied is step-by-step mathematical problem solving. The dataset includes problems from algebra, geometry, counting and probability, number theory, and precalculus. The authors state that they aim to create "a dataset that measures mathematical reasoning ability rather than memorization." The main research question is how well modern transformer models can solve advanced math problems and whether increasing model size improves reasoning ability.

The problem addressed by the paper is that many existing NLP benchmarks test pattern recognition rather than deep reasoning. Because of this, it is difficult to know whether language models truly understand mathematics. The authors claim that current models perform far below human competitors on challenging math problems, even when models are scaled up.

The evidence comes from evaluating multiple language models on the MATH dataset. The primary metric used is exact-answer accuracy. Results are reported across subject areas and difficulty levels. The paper also compares model performance to human competition baselines.

The statistical analysis consists of reporting overall accuracy, category-level breakdowns, and performance changes as model size increases. These comparisons are used to support the conclusion that large language models still struggle with advanced mathematical reasoning.