Easy2Hard-Bench: Standardized Difficulty Labels for Profiling LLM Performance and Generalization

Mucong Ding*[1], Chenghao Deng*[1], Jocelyn Choo[1], Zichu Wu[2], Aakriti Agrawal[1], Avi Schwarzschild[3], Tianyi Zhou[1], Tom Goldstein[1], John Langford[4], Anima Anadkumar[5], Furong Huang[1]

[1]University of Maryland, College Park [2]University of Waterloo [3]Carnegie Mellon University [4]Microsoft [5]California Institute of Technology

*Equal contribution Correspondence: {mcding,dengch16,furongh}@umd.edu

Introduction

❖ Motivation: difficulty labels of problems
- Current benchmarks: categorical, low granularity
- Ideal benchmarks: numerical, high granularity

❖ Why difficulty labels matters?
- Model behavior within a wide difficulty range.
- Generalization ability from easy to hard.

➢ Difficulty labels in current benchmarks

Easy | Medium | Hard

➢ Difficulty labels in our new benchmark

❖ **Our contributions**
- Presenting the innovative datasets with fine-grained difficulty labels
- Broadening the difficulty range and diversifying the task domains
- Thoroughly evaluating the SoTA LLMs on our new benchmark
- Detailed Profiling LLMs' easy-to-hard generalization ability

Easy2Hard-Bench Overview

E2H-AMC

Find the number of pairs of integers (a, b) with $1 \leq a < b \leq 57$ such that a^2 has a smaller remainder than b^2 when divided by 57 .

Difficulty: 0.587 ± 0.085 Percentile: 70.3%

E2H-GSM8K

openai/grade-school-math

Jerry is twice as old as he was 5 years ago. How old will Jerry be in 3 years?

Difficulty: 0.301 ± 0.157 Percentile: 30.3%

Mathematics

E2H-Codeforces

```
You're given a string of lower-
case Latin letters. Your task is
to find the length of its longest
substring that can be met in the
string at least twice. These
occurrences can overlap.
```

Difficulty: 0.204 ± 0.025 Percentile: 29.1%

Programming

E2H-ARC

Which of the following takes place during fertilization in animals?
(A) production of sperm and egg (B) joining of sperm and egg
(C) division of egg (D) development of embryo

Difficulty: 0.425 ± 0.066 Percentile: 30.4%

Reasoning

E2H-Winogrande

A very active Randy injured their elbow playing basketball against Lawrence, so _ said sorry.
(A) Randy (B) Lawrence

Difficulty: 0.124 ± 0.164 Percentile: 10.3%

Chess

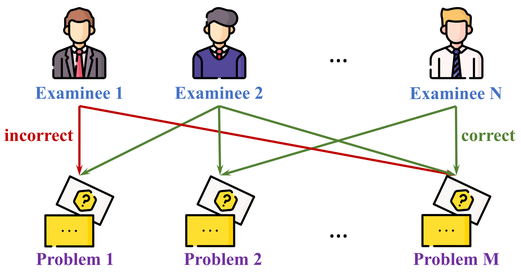
Difficulty: 0.163 ± 0.038 Percentile: 30.9%

Lichess.org

E2H-Lichess

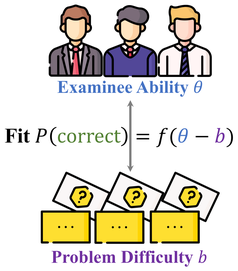
Fine-grained Difficulty Estimation

❖ Data source: examinees' attempts on problems (correct/incorrect)

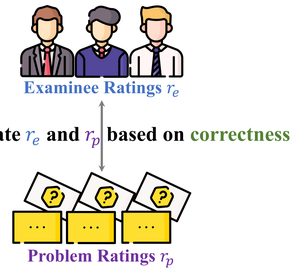


❖ The Method depends on whether examinee's ability is static or not

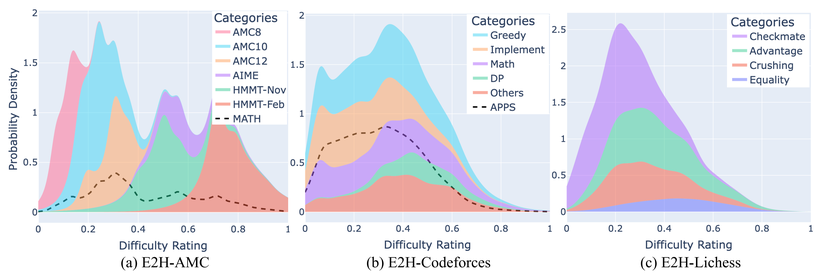
IRT Model
(examinee ability is static)



Glicko-2 Model
(examinee rating is not static)

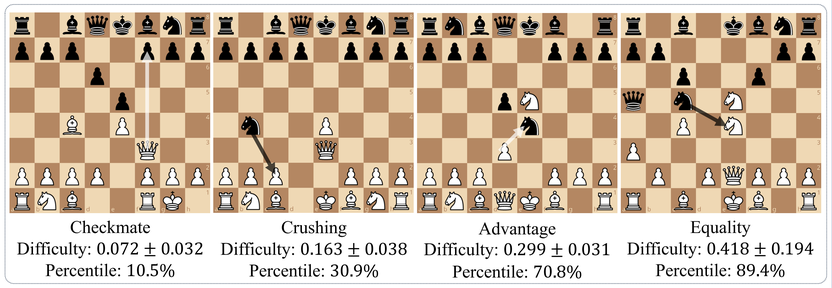


❖ The fine-grained standardized difficulty labels
- Shows that the problems within each domain cover a wide range of difficulties
- Verifies that our datasets include more challenging problems



(a) E2H-AMC (b) E2H-Codeforces (c) E2H-Lichess

- Corresponds with human recognition



Checkmate
Difficulty: 0.072 ± 0.032
Percentile: 10.5%

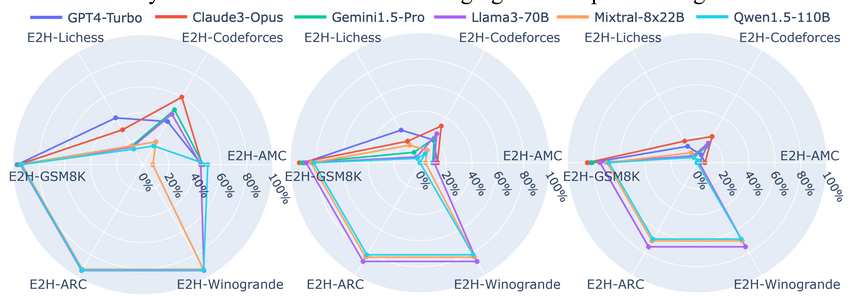
Crushing
Difficulty: 0.163 ± 0.038
Percentile: 30.9%

Advantage
Difficulty: 0.299 ± 0.031
Percentile: 70.8%

Equality
Difficulty: 0.418 ± 0.194
Percentile: 89.4%

Benchmarking SoTA LLMs

❖ 6 SoTA on all datasets, segmented into easy, medium, and hard levels
- The performance notably decreases as difficulty increases
- The newly curated are much more challenging than the pre-existing ones

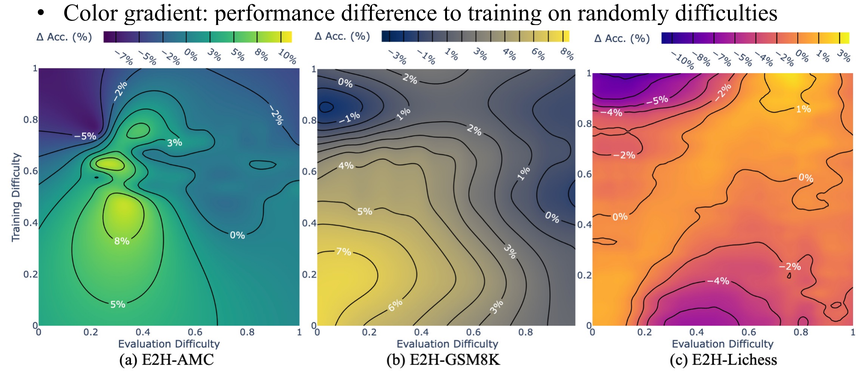


(a) Easy (b) Medium (c) Hard

❖ Model behavior against increasing difficulty levels for each dataset
- Most models show monotonic decreasing accuracies
- The extent of this decline varies significantly among models and datasets

Profiling Easy2Hard Generalizations

❖ The first to deliver detailed easy-to-hard generalization results across continuous, wide-range of difficulties on LLMs.
- Y-axis: trained on subsets of varying difficulty via Supervised Fine-Tuning (SFT)
- X-axis: evaluated across all evaluation difficulties
- Color gradient: performance difference to training on randomly difficulties



❖ Our preliminary observation
- Generalization benefits when training and evaluation difficulties are similar
- Training on more challenging samples poses increased generalization difficulties.