NPHardEval: Dynamic Benchmark on Reasoning Ability of Large Language Models via Complexity Classes

Lizhou Fan, et al

Rutgers University

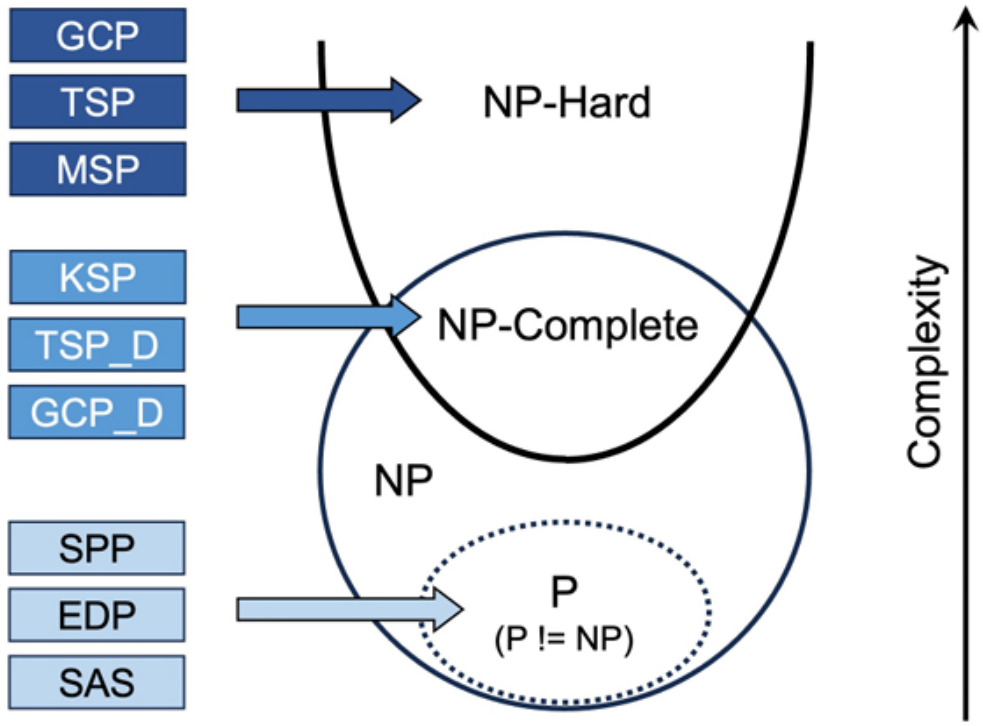
February 12, 2024

Abstract

- Current benchmarks are prone to the risk of overfitting
- We suggest benchmark with various complexity classes
  - P, NP-Complete, NP-Hard
- Dynamic update to avoid overfitting
  - Monthly Update datapoints

Introduction

- 3 Complexity classes, 3 Tasks, 100 Problems offered
  - Total $3 \times 3 \times 100 = 900$ problems
  - P: Solved in polynomial time
  - NP-Complete: Verified in polynomial time
  - NP-Hard: No guarantee for polynomial time



Introduction

- End-to-end automation
  - Generation, Verification are automated by using well-known task
- Excluded numerical computation
  - It is noticably difficult for LLMs
  - Focusing on pure logical reasoning ability
- Difficulty level system
  - For each task, there are 100 problems
  - Each problem has difficulty level from 1 to 10

P Tasks

- Sorted Array Search (SAS)
  - Finding the position of a target value after sorting a given array
  - Solved by sorting and binary search
- Edit Distance Problem (EDP)
  - Minimum number of operations to transform one string into another
  - Insertion, deletion, and substitution of a single character.
  - Solved by Dynamic programming
- Shortest Path Problem (SPP)
  - Finding the shortest path between two nodes in a weighted graph
  - Solved by Dijkstra's algorithm

NP-Complete Tasks

- Traveling Salesman Problem (TSP-D)
  - Can you complete a route, visiting each city at least once?
  - Total travel distance must be less than a specified value
- Graph Coloring Problem (GCP-D)
  - Can you color the vertices of a graph using a given number of colors?
  - No two adjacent vertices share the same color
- Knapsack Problem (KSP)
  - Fill a knapsack of fixed capacity without exceeding it
  - Maximize the total value of the selected items

NP-Hard Tasks

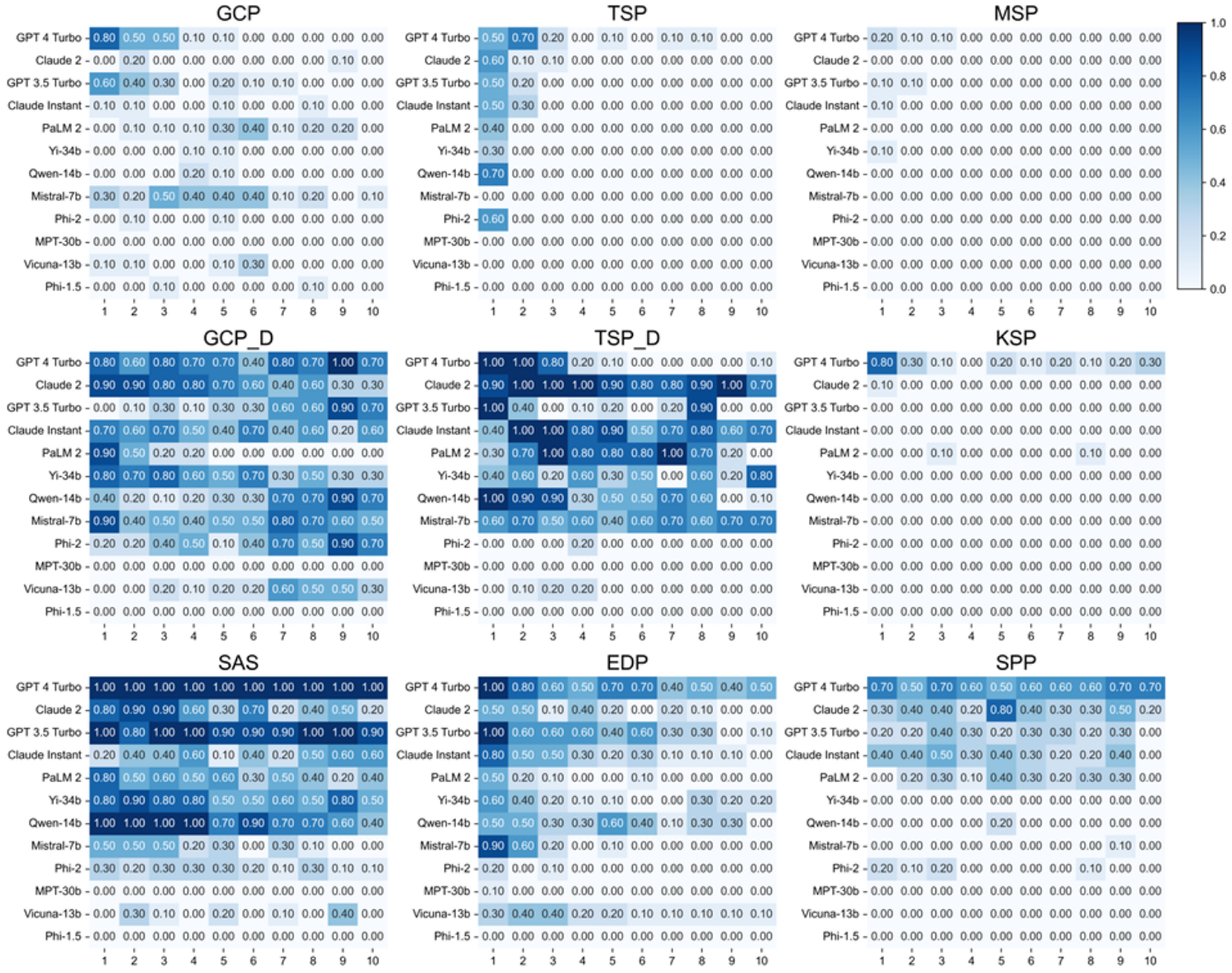
- Traveling Salesman Problem (TSP-O)
  - Find the shortest route, visiting each city at least once
  - Important for delivery services, maintenance operations, and sales
- Graph Coloring Problem (GCP-O)
  - Color the vertices of a graph with the constraint
  - Used in exam timetabling and register allocation in compilers
- Meeting Scheduling Problem (MSP)
  - Allocating time slots with participant availability and room capacity
  - Crucial in organizational management for scheduling meetings

Experiment

- Model Performance Comparison
  - Compare 5 closed-source models, 7 open-source models
  - Shed light on the relative strengths and weaknesses
- Robustness of Benchmark Assessments
  - Examines if the risk of “hacking” the benchmark is prevented
  - Examines if finetuning LLMs on benchmarks leads to overfitting
- Generalization through In-context Learning
  - Discern whether the model is “learning” or “mimicking”
  - Examine if performance is constant with varing difficulty levels

Result

- Increasing difficulty, performance significantly dropped



Limitations

- Task Complexity's Comparison
  - There are only 9 tasks
  - Difficulty is defined as linear increment of variables
- Randomness
  - In decision problem, LLMs may find correct answer by luck
- Model Update and Emergence
  - Fast-paced evolution of LLMs like Gemini Ultra
  - The analysis based on our benchmark may quickly become outdated