QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

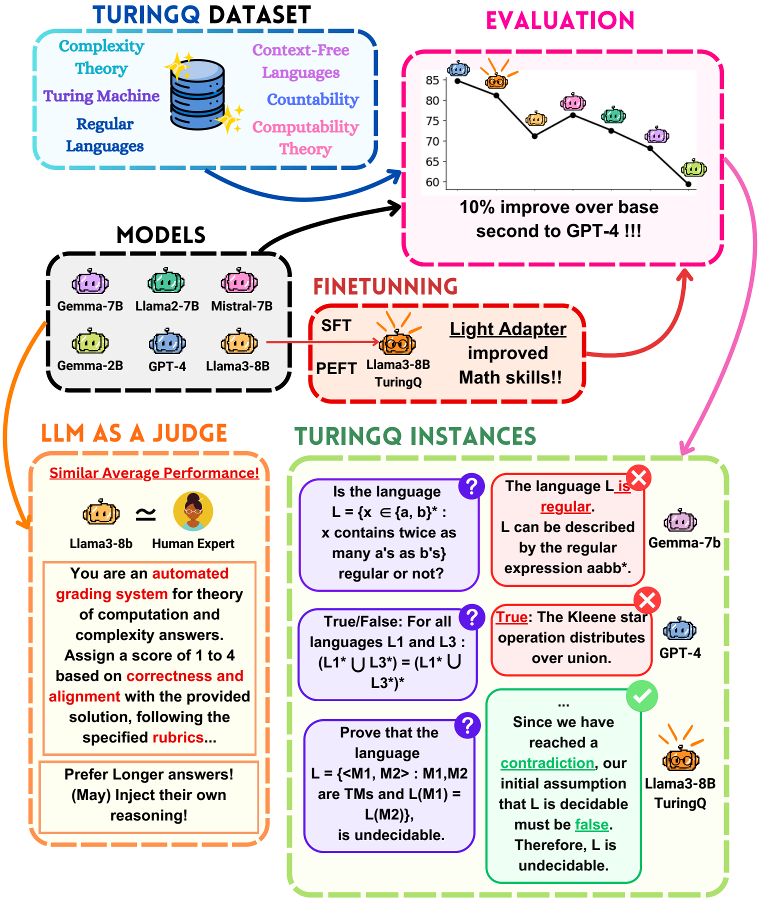
EMNLP
2024

Paper HuggingFace Github

TuringQ: Benchmarking AI Comprehension in Theory of Computation

Pardis Sadat Zahraei[♣,♦] (Paradisez2001@gmail.com) Sharif University of Technology[♣]
Ehsaneddin Asgari[♦] (easgari@hbku.edu.qa) Qatar Computing Research Institute - QCRI, Qatar[♦]

APPROACH



10% improve over base second to GPT-4 !!!

Light Adapter improved Math skills!!

Similar Average Performance!

You are an automated grading system for theory of computation and complexity answers. Assign a score of 1 to 4 based on correctness and alignment with the provided solution, following the specified rubrics...

Prefer Longer answers! (May) Inject their own reasoning!

Is the language L = {x ∈ {a, b}* : x contains twice as many a's as b's} regular or not?

The language L is regular. L can be described by the regular expression aabb*.

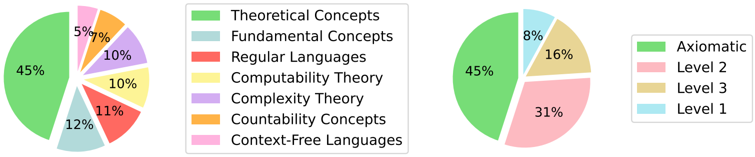
True/False: For all languages L1 and L3 : (L1* ⋃ L3*) = (L1* ⋃ L3*)*

True: The Kleene star operation distributes over union.

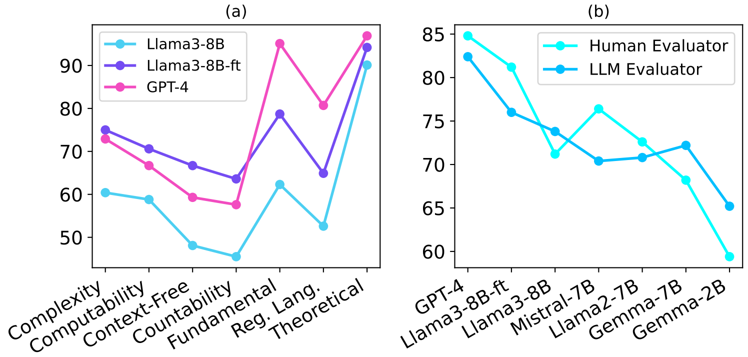
Prove that the language L = {<M1, M2> : M1,M2 are TMs and L(M1) = L(M2)}, is undecidable.

... Since we have reached a contradiction, our initial assumption that L is decidable must be false. Therefore, L is undecidable.

TuringQ Overview



Category and Difficulty Level Distribution in the TuringQ Dataset



a) LLM Performance Across TuringQ Categories, b) Performance on TuringQ: Evaluated by Human vs. LLM

SUMMARY

TuringQ is the **first benchmark** designed to evaluate the **reasoning capabilities** of Large Language Models (LLMs) in the field of **Computation Theory.**

- Sourced from **top university exams,** it comprises **4,006** question-answer pairs.
- Employs **Chain of Thought** prompting to assess the performance of **open-source LLMs** and **GPT-4.**

Automated Evaluation by LLMs

- Generates results **similar to human** judgment.
- Exhibits **surprising differences** compared to human evaluations.

Fine-Tuning Llama3-8B

- Finetuned using **SFT** and **PEFT.**
- **Substantial improvements** in reasoning accuracy and **out-of-domain tasks.**

Why Choose TuringQ?

- Evaluating LLM performance in **formal languages.**
- Advancing **complex computational reasoning** capabilities.

RESULTS

- Our fine-tuned model, **Llama3-8B-ft-TuringQ,** achieved a binary accuracy of **81.2%** on the TuringQ test set, indicating a **10%** improvement over the base model. Notably, it **enhanced performance across all categories,** approaching **GPT-4's** accuracy of **84.8%.**
- The LLM evaluator shows the **opposite trend** compared to human evaluators; **giving higher ratings to answers for harder questions.**
- LLM evaluators tend to **overrate weaker models** and **underrate stronger ones** compared to humans.
- LLMs prefer **longer, more complex answers, even if they are incorrect,** while humans evaluate more holistically.
- Evaluation on the **MATH dataset** showed a **0.6% accuracy increase** for the fine-tuned model.
- **Performance varied** across TuringQ categories, with the models performing best in **Theoretical Concepts** and worst in **Countability Concepts.**