

Are Large Language Models a Threat to Programming Platforms? An Exploratory Study

MdMustakim Billah
University of
Saskatchewan
Canada

mustakim.billah@usask.ca

Palash Ranjan Roy
University of
Saskatchewan
Canada

palash.roy@usask.ca

Banani Roy
University of
Saskatchewan
Canada

Zadia Codabux
University of
Saskatchewan
Canada

発表者: 福井 遥陽

プログラミングプロットフォーム

プログラミングをするための環境を用意し、**オンライン**でコンテストを開催するサイト

- | 現在、プログラミングプラットフォームであるCodeforces, LeetCodeなどがプログラミングスキルを育成するための手段として利用.
- | 技術系の一流企業で、候補者を評価する指標として使用.
→しかし、近年発達しているChatGPTなどを使った不正行為を見過ごす可能性がある.

背景

LLMの登場により、様々なコード生成に特化したモデルが登場している。

本研究では、対話型のLLMに焦点を当てている。→ブラウザ上で専門的な知識なしで利用できるモデル。

LLMからより良い推論を引き出すために、「Chain-of-Thought」によるプロンプトを用い、エラー情報を提供している。

Chain-of-thought

ある問いに対する答えではなく，論理的な段階を経て結論に至るプロセスを重視し，思考過程を表現する技術.

つまり、、、

| どのようなエラーが起きたかだけでなく，どのような思考を経て，そのエラーにたどり着いたかの情報を重視.

例) 問題 : $1+1\times 2$

回答 : 4

思考過程 : $1+1=2 \rightarrow 2\times 2=4$

背景

Codeforces, LeetCodeは提出されたコードに対して自動で評価を行う。これらは、カテゴリごとに分類された様々なタイプの問題が用意されている。

正誤判定システムは

- Time Limit Exceeded（時間制限超過）
- Memory Limit Exceeded（メモリ制限超過）
- Compilation Error（コンパイルエラー）
- Runtime Error（実行時エラー）

で結果进行分类している。

目的

LLMを用いてプログラミングプラットフォームに対して、
評価をすることを目的としている。

つまり、、、

LLMを用いてプログラミングプラットフォームに用意されて
る問題を解き、正答率の度合いで脆弱性を判断する。

リサーチクエスチョン

｜ RQ1...LLMは多様な課題をどれだけ解けるか？

成功率で脆弱性を判断.

｜ RQ2...LLMの性能はどのような要因で変化するか？

カテゴリ・難易度・使用言語の要素ごとの精度を評価.

｜ RQ3...LLMの問題解決能力は人間と比べてどうか？

LLMと人間のプログラマーの正答率を判断.

データ収集

CodeforcesやLeetCodeで分類されるカテゴリの内，共通している15のカテゴリを選定.

例) 文字列・ツープインタ・数学・貪欲法など...

そこから，難易度が3段階(Easy, Medium, Hard)に，正答率も3段階(高, 中, 低)に分類されているため，それぞれ1問ずつ収集.

そのほかにオンラインコンテストに参加し，LLMのリアルタイム性能を調査.

問題解決手法

使用したLLM：

- ChatGPT-3.5
- Gemini 1.0 Pro
- Meta AI

以上のLLMを使用して得たコードをプログラミングプラットフォームに提出し，正誤判定を得る．

エラー時は，エラー内容をLLMにフィードバックし，再度回答を生成．

実験結果・RQ1

| 成功率

LeetCodeで選定した98問で実施

- ChatGPT 71.43%
- Meta AI 58.16%
- Gemini 68.37%

Codeforcesで選定した126問で実施

- ChatGPT 26.98%
- Meta AI 16.67%
- Gemini 7.94%

実験結果・RQ1

成功率(オンラインコンテスト)

Table 3: Codeforces (Cf) and LeetCode (LC) virtual contests' results

Contests	Solved	Standing
Cf - Educational	0 / 6	
Cf - Division 4	1 / 7	22134 / 60343
Cf - Division 3	1 / 8	16324 / 53980
Cf - Division 2	0 / 6	
Cf - Division 1	0 / 6	
LC - Weekly 1	1 / 4	10383 / 22688
LC - Biweekly 1	2 / 4	9999 / 25287
LC - Weekly 2	1 / 4	15801 / 29736
LC - Biweekly 2	2 / 4	12834 / 26241

実験結果・RQ1

LLMsはLeetCodeで高い成功率を記録.

Codeforcesでは成功率が30%にも満たなかった.

HackerRankの認定試験では3つのLLMの全てが合格

オンラインコンテストでは精度に難が見られた.

実験結果・RQ2

問題カテゴリごとの成功率

- ChatGPTは、平均的に高い成功率を記録
- Meta AIは数論、確率・統計のカテゴリで最も高い成功率を記録
- Geminiは平均的にMeta AIよりも高い成功率を記録

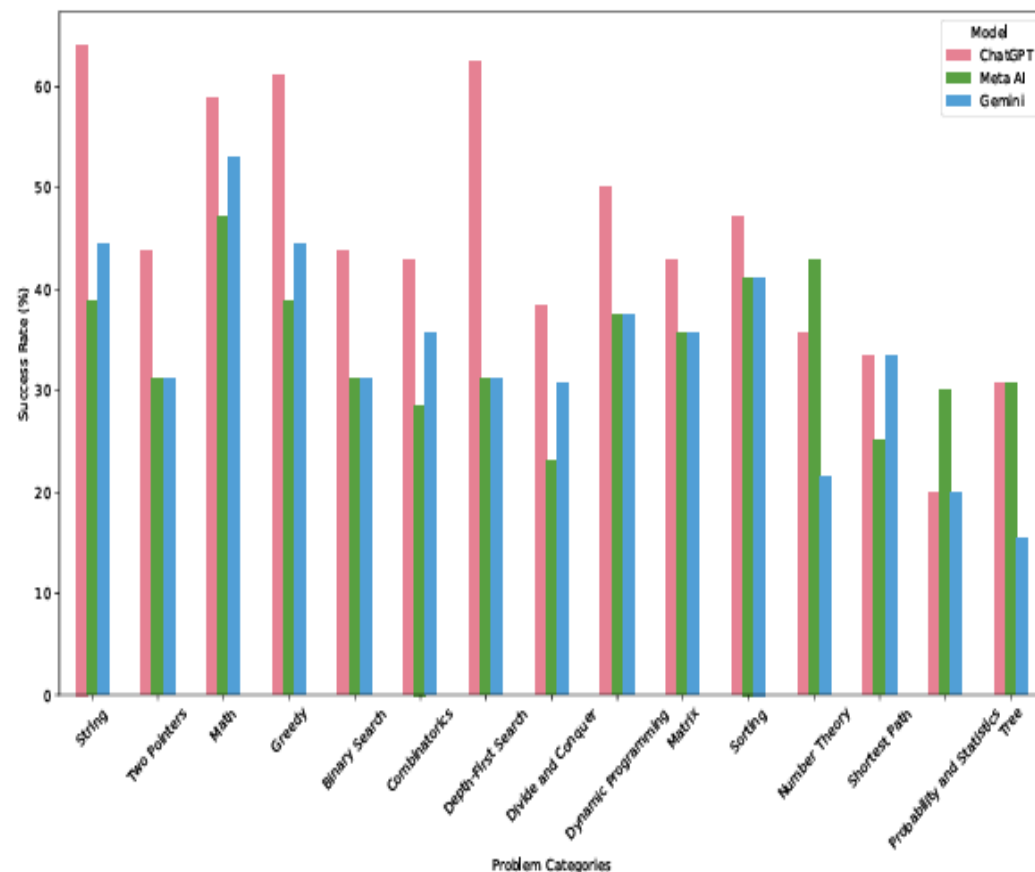


Figure 3: Success Rates Across Problem Categories

実験結果・RQ2

難易度・正答率ごとの成功率

- ・全てのLLMが“簡単で正答率が高い問題”に対しての成功率が高く, “難しくて正答率が低い問題”に対しての成功率が低い.

→難易度・正答率で精度が大きく変化

Table 4: Difficulty and Acceptance Range Wise Solved Problems by ChatGPT (C), Gemini (G), and Meta AI (M)

Difficulty	Acceptance (C)			Acceptance (G)			Acceptance (M)		
	High	Medium	Low	High	Medium	Low	High	Medium	Low
Easy	23	17	2	18	16	3	22	17	3
Medium	18	18	5	8	12	7	10	14	2
Hard	8	11	1	3	8	2	4	5	1

実験結果・RQ3

人間との比較

- ・ 時間効率

全体で50%以上のユーザに勝利

- ・ メモリ効率

全体で50%以上のユーザに勝利

→LLMは多数の人間に対して優れていると言える。

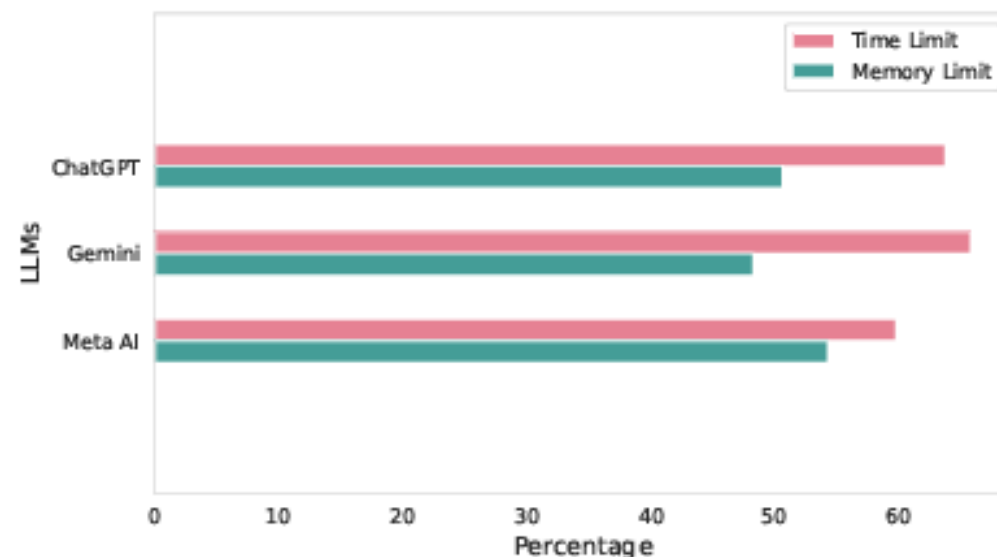


Figure 4: Comparison with users' performance with LLMs in terms of time and memory usage of LeetCode

議論

- | 今回の研究で, LLMが多様なプログラミング課題において高い成功率を収め, 多数の人間より優れた成績を残した.
→ プログラミングプラットフォームの信頼性の低下
- | フィードバックにより, LeetCodeでChatGPTとGeminiの両方で成功率が12.24%上昇 CodeforcesではChatGPTとMeta AIで4.76%上昇

議論

LLMはLeetCodeよりもCodeforcesでの問題解決に苦戦.

- Codeforcesの問題文の平均語数 : 274.14
- LeetCodeの問題文の平均語数 : 164.25

→LLMはン外問題文を処理する際に理解力が低下しやすい.

結論

- LLMはLeetCodeにおいて71.43%の高い成功率を記録し、LLMの性能は非常に優れていることが確認できた。
- CodeforcesではLLMは満足な成果を挙げられず、またライブコンテストでは明確に苦戦する傾向が見られた。
- HackerRankの認定試験では好成績を残し、形式化された課題では強さを発揮することが確認できた。

今後の展望

| 評価対象のプラットフォームを拡張

| さらに多くのLLMを使用して比較研究を行う

| 自然言語理解，問題文長，形式の違いなどがLLM性能に与える影響の分析