A Performance Study of LLM-Generated Code on Leetcode

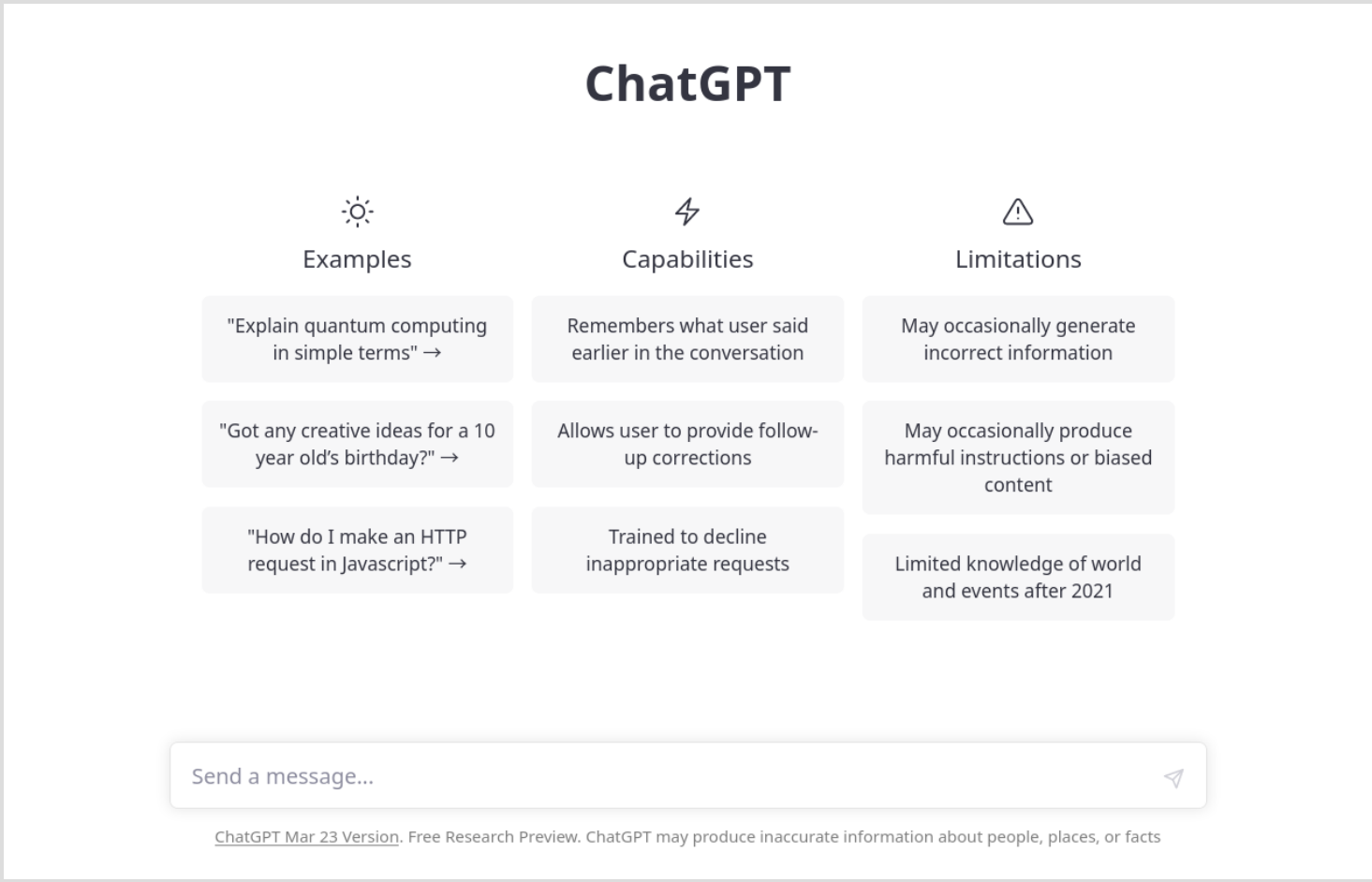
Coignion Tristan, Quinton Clément, Rouvoy Romain

Green Days 2024 - Toulouse

Université de Lille

Inria

New Shiny Things



New Shiny Things

```python
import datetime

def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date, va
```

GitHub Copilot

Some definitions

Some definitions

Large Language Model (LLM) :

An artifical intelligence capable of generating text

Some definitions

Large Language Model (LLM) :

An artifical intelligence capable of generating text

Code LLM : LLMs specialized in writing code

starcoder

Some definitions

Large Language Model (LLM) :

An artifical intelligence capable of generating text

Code LLM : LLMs specialized in writing code

Code Assistant : Code LLMs integrated in the IDE

LLM + Green = 💔

LLM + Green = 💔

LLMs need a lot of computing resources

Training StarCoder2-7B

=> 100,000kWh

=> 30,000kgCO2eq

LLM + Green = 💔

LLMs need a lot of computing resources

Training StarCoder2-7B

=> 100,000kWh

=> 30,000kgCO2eq

Is it really worth the cost?

Is it worth it?

- Measure the impact of the LLM
- Measure the time gained for the developer
- Measure the energy saved on the software

Is it worth it?

- Measure the impact of the LLM
- Measure the time gained for the developer
- Measure the energy saved on the software

How fast is the code generated by LLMs ?

The task

The task

The task

A competitive programming platform hosting **algorithmic** problems

The task

A competitive programming platform hosting **algorithmic** problems

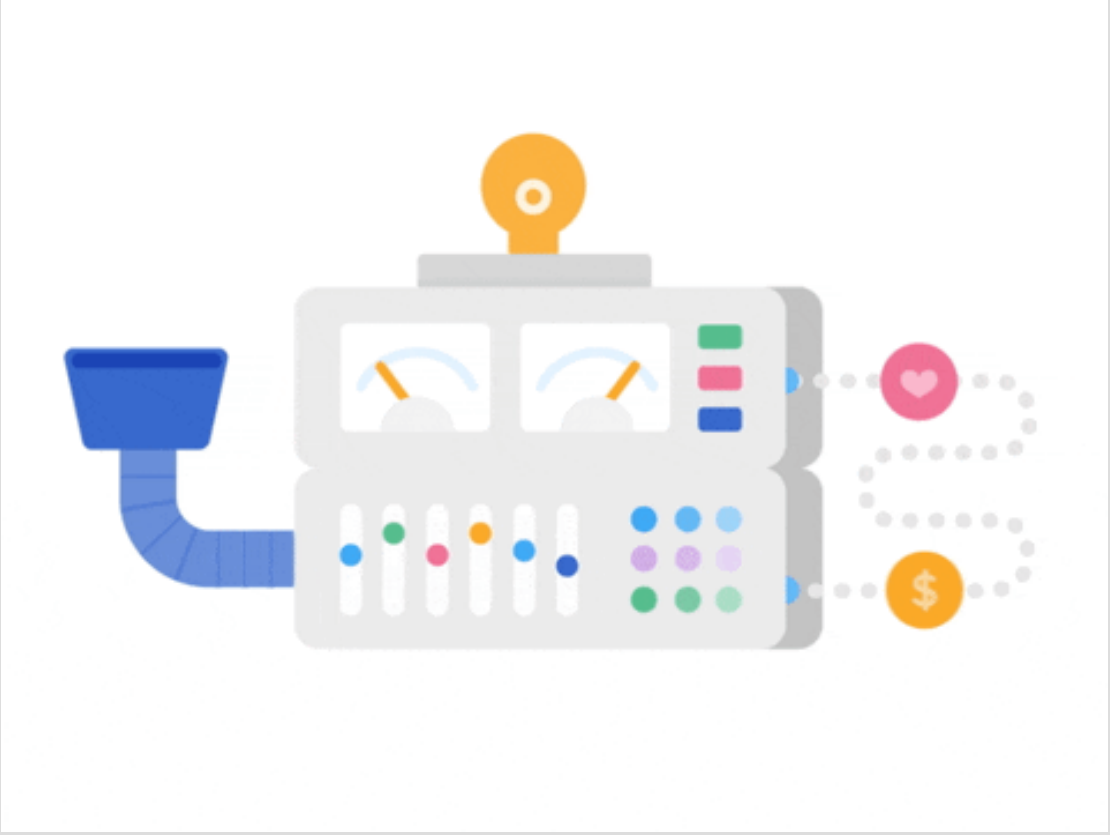
+ Practical for performance testing

+ Practical for evaluating LLMs

LLMs under study

LLM Model	Model family	Size	RQ1
GitHub Copilot	Codex	11	✓
CodeGen-Mono 6B	CodeGen	6	✓
CodeGen-Mono 2B	CodeGen	2	✓
CodeGen-Mono 350M	CodeGen	0.35	✓
CodeGen2.5-7B-mono	CodeGen2.5	7	
CodeGen2.5-7B-instruct	CodeGen2.5	7	
CodeLlama-7B-instruct	CodeLlama	7	
CodeLlama-7B	CodeLlama	7	
CodeLlama-7B-python	CodeLlama	7	
CodeLlama-13B-instruct	CodeLlama	13	
CodeLlama-13B-python	CodeLlama	13	
replit-code-v1-3b	replit-code	3	
WizardCoder-pythin	WizardCoder	7	
SantaCoder	Santacoder	1.1	✓
StarCoder	StarCoder	15.5	
InCoder 6B	Incoder	6	✓
InCoder 1B	Incoder	1	✓
CodeParrot	Codeparrot	1.5	✓

Results



RQ1: Can Leetcode be used as a **dataset** and a benchmark platform for evaluating LLMs?

LLMs success rate on :

RQ1: Can Leetcode be used as a **dataset** and a benchmark platform for evaluating LLMs?

LLMs success rate on :

- old problems : **37%** of valid solutions

RQ1: Can Leetcode be used as a **dataset** and a benchmark platform for evaluating LLMs?

LLMs success rate on :

- old problems : **37%** of valid solutions
- new problems (published after training) : **3%** of valid solutions

RQ1: Can Leetcode be used as a **dataset** and a benchmark platform for evaluating LLMs?

LLMs success rate on :

- old problems : **37%** of valid solutions
- new problems (published after training) : **3%** of valid solutions

Why are the LLMs 10x worse on newer questions?

RQ1: Can Leetcode be used as a **dataset** and a benchmark platform for evaluating LLMs?

LLMs success rate on :

- old problems : 37% of valid solutions
- new problems (after January 2023) : 3% of valid solutions

Why are they 10x worse on newer questions ?

Data contamination

RQ1: Can Leetcode be used as a **dataset** and a benchmark platform for evaluating LLMs?

LLMs success rate on :

- old problems : 37% of valid solutions
- new problems (after January 2023) : 3% of valid solutions

Why are they 10x worse on newer questions ?

Data contamination

=> Harder to reproduce and generalize research

=> Questions the previous research done using Leetcode

RQ1: Can Leetcode be used as a dataset and a **benchmark** platform for evaluating LLMs?

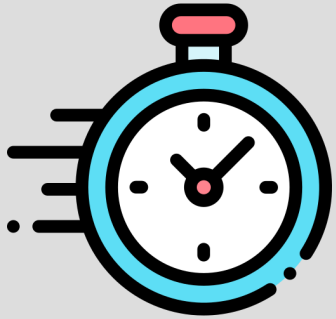
RQ1: Can Leetcode be used as a dataset and a **benchmark** platform for evaluating LLMs?

Leetcode provides useful measures :

run time

memory usage

ranking (based on run time)



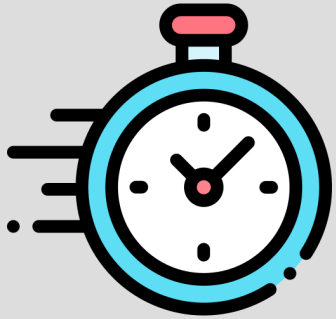
RQ1: Can Leetcode be used as a dataset and a **benchmark** platform for evaluating LLMs?

Leetcode provides useful measures :

run time

memory usage

ranking (based on run time)



BUT

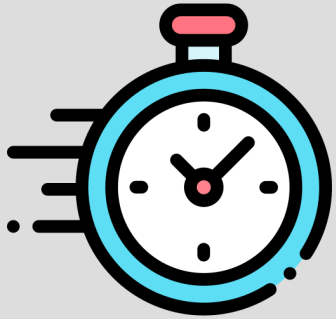
RQ1: Can Leetcode be used as a dataset and a **benchmark** platform for evaluating LLMs?

Leetcode provides useful measures like :

run time

memory usage

ranking (based on run time)



BUT

Very **high variance** (inability to differentiate solutions of different time complexities)

Ranking evolves over time, thus is **unreliable**

RQ2: Are there notable differences in performances between LLMs?

RQ2: Are there notable differences in performances between LLMs?

Almost (<5%) no problems where one LLM is consistently better than another.

RQ2: Are there notable differences in performances between LLMs?

Almost (<5%) no problems where one LLM is consistently better than another.

Very small differences (Cohen's d < 0.05), thus **negligible**.

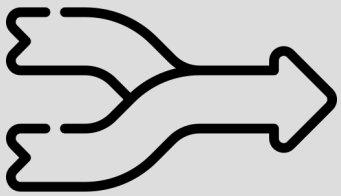
RQ2: Are there notable differences in performances between LLMs?

Almost (<5%) no problems where one LLM is consistently better than another.

Very small differences (Cohen's d < 0.05), thus **negligible**.

LLMs seem to **converge** towards the same kinds of solutions

(not necessarily the best ones)



RQ2: Are there notable differences in performances between LLMs?

Better LLMs

Faster code

RQ3: Is there an effect of the temperature on the code's performance?

RQ3: Is there an effect of the temperature on the code's performance?

Temperature : Parameter controlling the "**creativity**" of the model

RQ3: Is there an effect of the temperature on the code's performance?

Temperature : Parameter controlling the "**creativity**" of the model

Higher temperatures => higher variance of the performance of the code

=> Higher temperatures can help in searching for faster solutions.

RQ4: How fast is code generated by LLMs compared to humans?

RQ4: How fast is code generated by LLMs compared to humans?

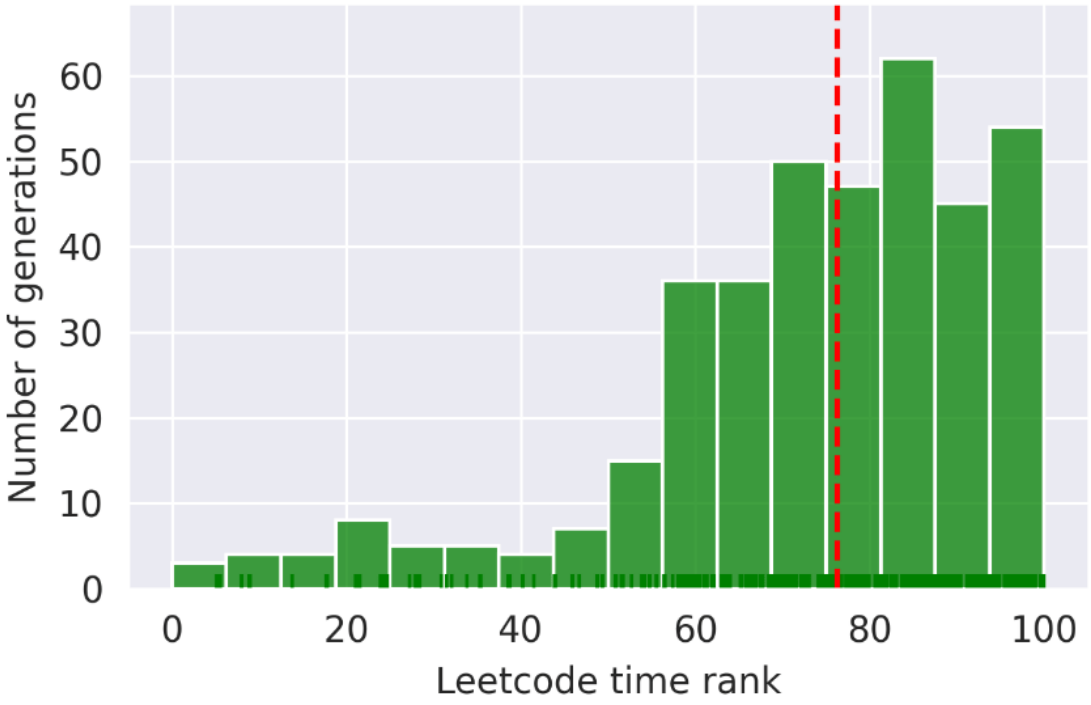


Figure 8: Distribution of the ranking for the CodeGen-6B-mono model

On average, the generated solutions are faster than **73%** of the other submissions on Leetcode

RQ4: How fast is code generated by LLMs compared to humans*?

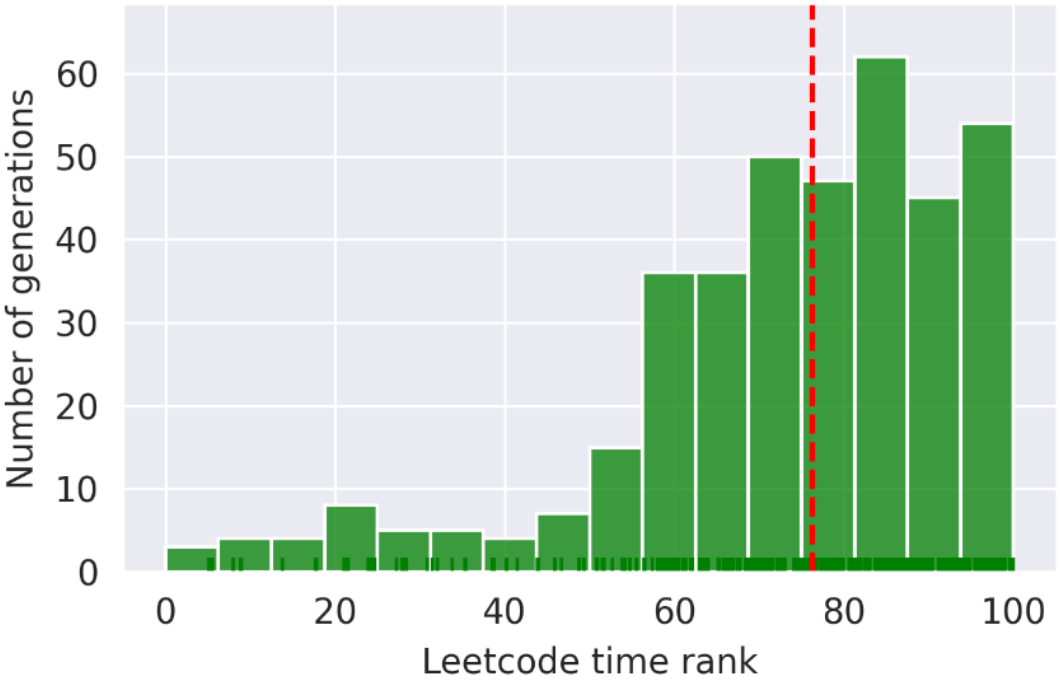


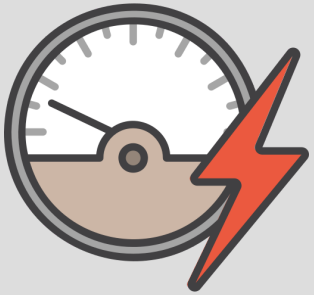
Figure 8: Distribution of the ranking for the CodeGen-6B-mono model

On average, the generated solutions are faster than **73%** of the other submissions on Leetcode

* assuming the other submissions on Leetcode were made by humans

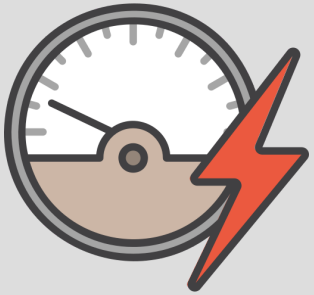
Conclusions

Conclusions



Performance of generated code is **largely similar** across different models regardless of their size, training data or architecture

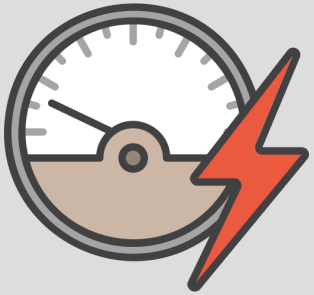
Conclusions



Performance of generated code is **largely similar** across different models regardless of their size, training data or architecture

Increasing the temperature parameter leads to a **greater variance** in performance

Conclusions



Performance of generated code is **largely similar** across different models regardless of their size, training data or architecture

Increasing the temperature parameter leads to a **greater variance** in performance

Leetcode should be used cautiously when evaluating LLMs because of issues of measure stability and **data contamination**

Perspectives

Perspectives

- Extend the study on other kinds of problems

Perspectives

- Extend the study on other kinds of problems
- How to make LLMs produce **greener code** ?

Perspectives

- Extend the study on other kinds of problems
- How to make LLMs produce **greener code** ?
- What is the energy consumption of a code assistant ?

Thanks for listening !

Any questions?