A Performance Study of LLM-Generated Code on LeetCode

Coignion Tristan, Quinton Clément, Rouvoy Romain

Green Days 2024 - Toulouse
# New Shiny Things

## ChatGPT

<table>
<thead>
<tr>
<th>Examples</th>
<th>Capabilities</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Explain quantum computing in simple terms&quot; →</td>
<td>Remembers what user said earlier in the conversation</td>
<td>May occasionally generate incorrect information</td>
</tr>
<tr>
<td>&quot;Got any creative ideas for a 10 year old's birthday?&quot; →</td>
<td>Allows user to provide follow-up corrections</td>
<td>May occasionally produce harmful instructions or biased content</td>
</tr>
<tr>
<td>&quot;How do I make an HTTP request in Javascript?&quot; →</td>
<td>Trained to decline inappropriate requests</td>
<td>Limited knowledge of world and events after 2021</td>
</tr>
</tbody>
</table>

---

*ChatGPT Mar 23 Version. Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts*
import datetime

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

Large Language Model (LLM): An artificial intelligence capable of generating text
Some definitions

**Large Language Model (LLM):** An artificial intelligence capable of generating text

**Code LLM:** LLMs specialized in writing code
Large Language Model (LLM) :
An artificial 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
=> 100,000 kWh
=> 30,000 kg CO2eq

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

<table>
<thead>
<tr>
<th>LLM Model</th>
<th>Model family</th>
<th>Size</th>
<th>RQ1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GitHub Copilot</td>
<td>Codex</td>
<td>11</td>
<td>✔</td>
</tr>
<tr>
<td>CodeGen-Mono 6B</td>
<td>CodeGen</td>
<td>6</td>
<td>✔</td>
</tr>
<tr>
<td>CodeGen-Mono 2B</td>
<td>CodeGen</td>
<td>2</td>
<td>✔</td>
</tr>
<tr>
<td>CodeGen-Mono 350M</td>
<td>CodeGen</td>
<td>0.35</td>
<td>✔</td>
</tr>
<tr>
<td>CodeGen2.5-7B-mono</td>
<td>CodeGen2.5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CodeGen2.5-7B-instruct</td>
<td>CodeGen2.5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-7B-instruct</td>
<td>CodeLlama</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-7B</td>
<td>CodeLlama</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-7B-python</td>
<td>CodeLlama</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-13B-instruct</td>
<td>CodeLlama</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-13B-python</td>
<td>CodeLlama</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>replit-code-v1-3b</td>
<td>replit-code</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WizardCoder-python</td>
<td>WizardCoder</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>SantaCoder</td>
<td>Santacoder</td>
<td>1.1</td>
<td>✔</td>
</tr>
<tr>
<td>StarCoder</td>
<td>StarCoder</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td>InCoder 6B</td>
<td>Incoder</td>
<td>6</td>
<td>✔</td>
</tr>
<tr>
<td>InCoder 1B</td>
<td>Incoder</td>
<td>1</td>
<td>✔</td>
</tr>
<tr>
<td>CodeParrot</td>
<td>Codeparrot</td>
<td>1.5</td>
<td>✔</td>
</tr>
</tbody>
</table>
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?

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?

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

Figure 8: Distribution of the ranking for the CodeGen-6B-mono model
RQ4: How fast is code generated by LLMs compared to humans*?

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?