UFMG

lab-soft
software engineering laboratory

Evaluation of the use of LLMs in Software Maintenance and Evolution

Larisse Amorim, Caíque Fortunato, Gustavo Vale, Eduardo Figueiredo

UFMG

lab-soft
software engineering laboratory

Performance evaluation of LLMs as code judges

Motivation

- Code Review takes a long time
  - Developers typically spend over six hours per week on code reviews
- LLM models, like GPT, understand and generate code
- They perform well in software engineering tasks

Introduction

- CROP – Code Review Open Platform
  - Open platform with real code review data
  - Projects used: Eclipse and Couchbase

- Compare the classifications made by LLMs with those of human experts.

- Analyze cases of agreement and disagreement between LLMs and humans.

Goal

- Evaluate the ability of large-scale language models (LLMs) to identify refactorings and judge the quality of code releases in real-world Pull Requests.
- Assess whether LLMs are able to compare two versions of code (pre and post-PR), indicating which one presents better:
  - Clarity
  - Readability
  - Maintainability
- Compare the classifications made by LLMs with those of human experts.
- Analyze cases of agreement and disagreement between LLMs and humans.

Research Questions

- **RQ1:** How accurate are LLMs in judging code quality compared to human evaluators?

- **RQ2:** What factors contribute to discrepancies between LLMs' and human evaluators' judgments of code quality?

Dataset

- MaRV Scripts and Dataset
  - 126 GitHub Java repositories
  - 693 manually evaluated code pairs extracted from
  - 321 refactor instances that reviewers agreed
    - Rename Variable
    - Extract Method
      - 146 instances

LLM used

OpenAI
GPT-5 mini

chatgptdeutsch.info

Prompt

Can you please rate whether the code 1 is better than the code 2 in terms of clarity, readability, and maintainability?

I would like to know if:

1️⃣ The code is easy to understand

2️⃣ There are improvements in naming, structure, or style

3️⃣ It is easy to maintain and extend

Code 1:

[no_refactor_version]

Code 2:

[refactor_version]

🧠 **Prompt Goal:**

Assess whether, for LLMs, the refactored version of the code is superior to the previous version, based on software quality criteria.

Results

How accurate are LLMs in judging code quality compared to human evaluators?

Both LLMs chose code_2	115
ChatGPT code_1	9
ChatGPT Unable to choose between code versions	3
Llama code_1	5
Llama Unable to choose between code versions	20

What factors contribute to discrepancies between LLMs' and human evaluators' judgments of code quality?

In progress…

UFMG

lab-soft
software engineering laboratory

Future Works

Future Works

- Expand our evaluation to include more diverse projects
  - Bug fixes
  - New feature

UFMG

lab-soft
software engineering laboratory

Thank you!