Paper Overview 2

Ivo Melse, s1088677

March 7, 2025

Summary

Large Language models (LLMs) are used to generate code from prompts in natural language. A common problem is that the user-provided requirements are unclear or ambiguous. This often leads to code that does not do what the user wanted. The authors introduce a new framework called `ClarifyGPT`, which aims to enable LLMs to ask follow-up questions. To evaluate `ClarifyGPT`, the authors run a trial with 10 human participants and a large-scale automated simulation. They then use quantitative measures to compare performance of `ChatGPT` and `GPT-4` against `ClarifyGPT`.

Evidence

Emperical. The researchers introduce four sets of coding problems. Each problem has a set of unit tests attached.

1. They have 10 users use ChatGPT and ClarifyGPT to generate code and solve the problems.
2. They use ChatGPT and ClarifyGPT with simulated feedback to generate code and solve the problems.

In both cases, the code generated by ClarifyGPT is significantly more likely to pass the unit tests.

Furthermore, for the human test subjects, the researchers asked a number of questions about their experience using ClarifyGPT. The results were quite positive, but they did not really compare it to the baseline.

Strengths

The researchers have provided a new framework for LLM-based code generation, and showed that it was more effective than existing methods.

Weaknesses

- ClarifyGPT cannot be applied to all LLMs.
- ClarifyGPT adds more overhead to the code-generation process. This raises the energy cost.
- A threat to validity is the fact that the problem sets might have been in the LLM training data.

Evaluation

The paper makes a useful contribution to development using LLMs. It should be accepted in a peer-reviewed journal.

Quality of writing

The writing is generally clear and well-structured. I think it would have been better if the authors had been more brief. I felt like some sentences were generated by a LLM.

Questions

- The researchers claim that the ClarifyGPT framework cannot be used on all LLMs. Why is this?
- In Table 7, on page 17. The researchers provide human evaluation metrics for ClarifyGPT, but they don't provide the same figures for the base LLMs. They conclude that the evaluations are positive. Do you think that this could be a threat to validity of this particular result?