Benchmarking the Communication Competence of Code Generation for LLMs and LLM Agent

Jie JW Wu, Fatemeh H. Fard

University of British Columbia

Sponsored by:

videns

Introduction

Problem Statement: there is still a gap between LLMs being capable coders and being top-tier software engineers: Top-level software engineers often ask clarifying questions to reduce ambiguity in both requirements and coding solutions.

Our Vision: For code generation task, we argue that the AI system should proactively recognize which information is missing, and find these missing pieces to be able to complete the task with high quality.

Our Approach:

We conducted an empirical study on the benchmark and analysis of the communication skills of models for code generation.

- We define communication skills of a model as "being able to ask clarifying questions when the description of the code generation problem has issues".
- We created a new benchmark, *HumanEvalComm*, by modifying problem descriptions according to three issues: inconsistency, ambiguity, incompleteness.
- We proposed a new LLM agent approach, *Okanagan*, to identify and ask questions in ambiguous parts from code and descriptions for further refining the generated code.

HumanEvalComm Benchmark

Overview: To develop HumanEvalComm, we changed each problem description in HumanEval manually, using a taxonomy of clarification types: *Ambiguity, Inconsistency, Incompleteness*.

Clarification Category	*Ambiguity*	*Inconsistency*	*Incompleteness*
1a	✓		
1c		✓	
1p			✓
2ac	✓	✓	
2cp		✓	✓
2ap	✓		✓

Table 1. Problem descriptions with different combinations of clarification types being applied in HumanEvalComm.

Empirical Study

Research Questions:

- RQ1: How do Code LLMs perform in communication competency when requirements in the problem descriptions are incomplete, inconsistent, ambiguous?
- RQ2: How does Okanagan perform compared with Code LLMs in terms of communication skills?

Methodology:

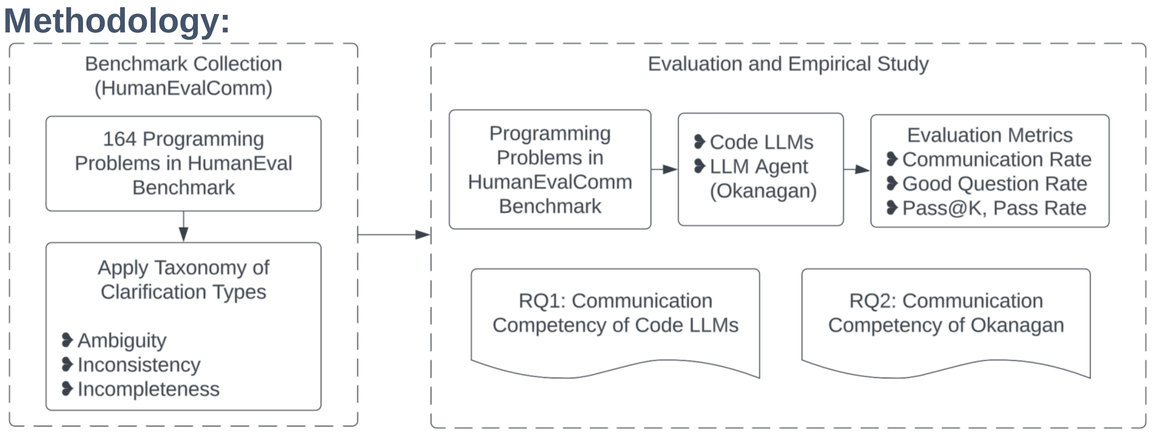

Fig. 1. The visual illustration of the methodology on the evaluation of communication skills for Large Language Models of code.

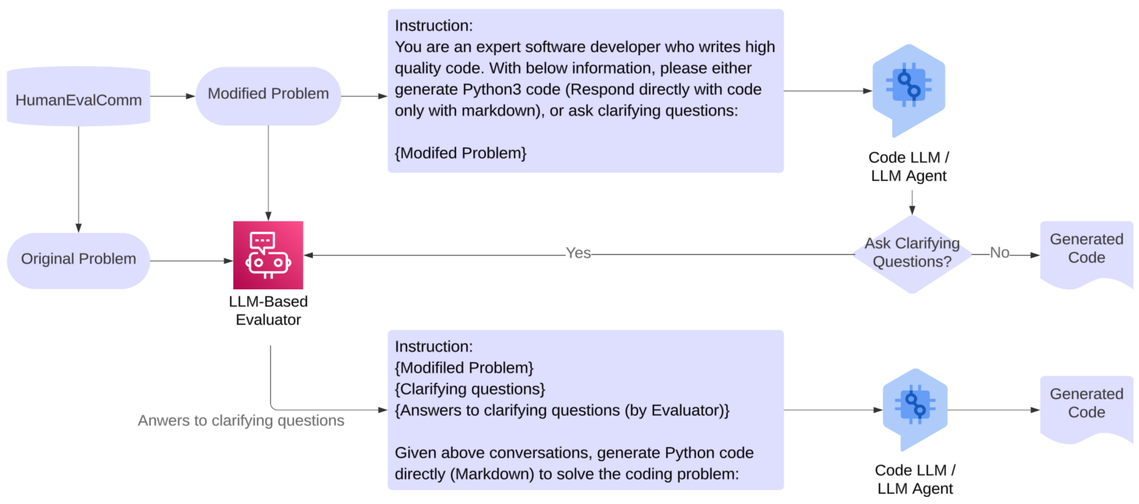

Fig. 2. Flowchart for the evaluation of models, either Code LLMs or Okanagan (LLM agent), in communication capability.

LLM Agent Approach (Okanagan)

Okanagan leverages multi-round structure and customized prompt format for asking clarifying questions in code generation tasks. We introduce 3 rounds in Okanagan:

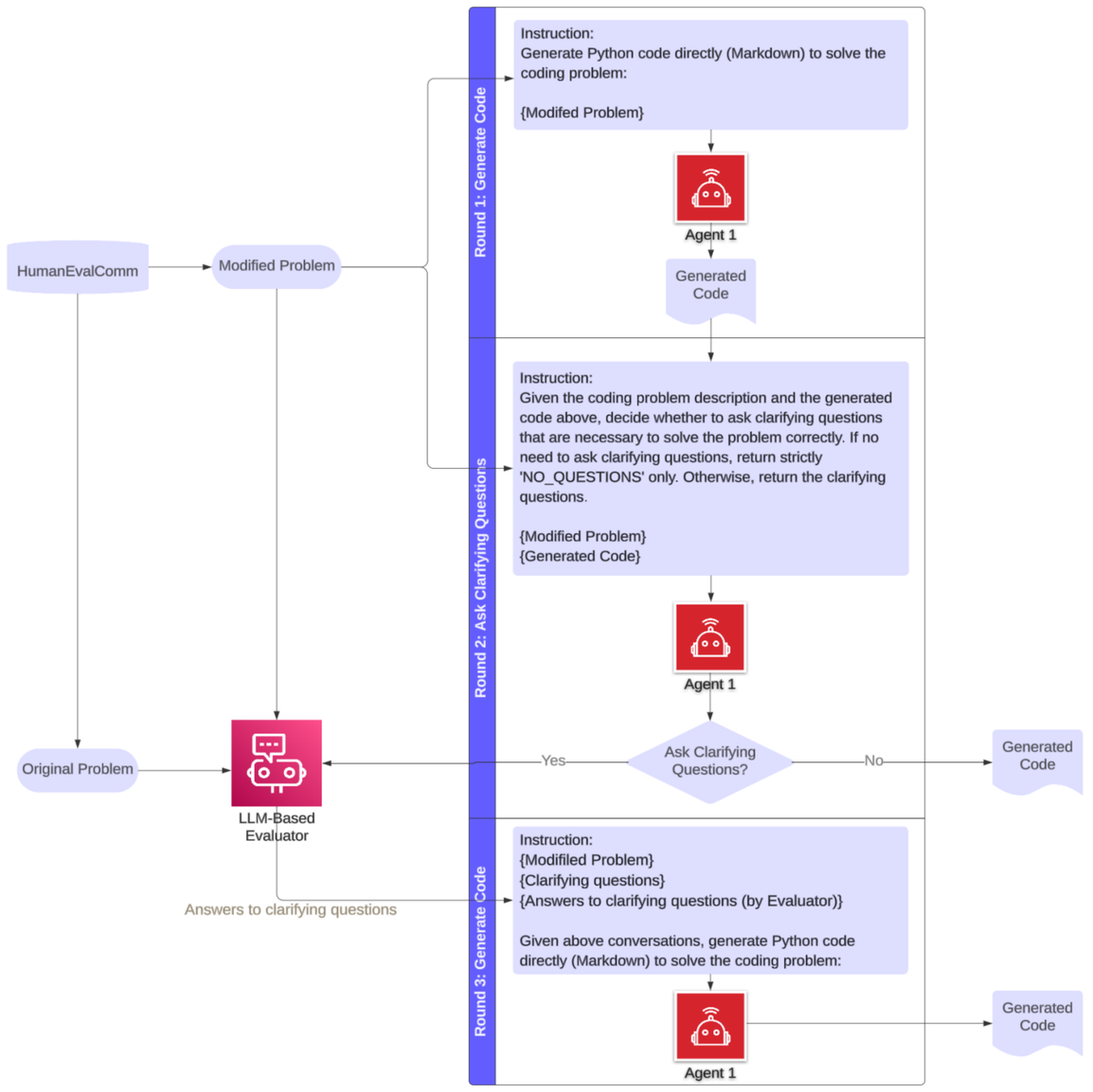

Fig. 3. An illustration of the process of Okanagan, an LLM agent approach.

Result and Summary

Model	Pass@1		Test Pass Rate		Comm. Rate	Good Question Rate
	HmEval	*HmEvalComm*	*HmEval*	*HmEvalComm*		
ChatGPT	65.58%	31.34%	76.42%	49.39%	14.21%	13.43%
CodeLlama	29.88%	19.35%	45.71%	37.79%	10.16%	37.55%
CodeQwen1.5 Chat	76.83%	**47.61%**	84.4%	**62.89%**	4.82%	41.68%
DeepSeek Coder	71.78%	**45.68%**	79.44%	**62.25%**	**30.76%**	**61.42%**
DeepSeek Chat	12.8%	26.32%	13.86%	44.52%	**37.93%**	**58.71%**
Okanagan	27.45%	**39.62%**	33.45%	**56.98%**	**72.73%**	**52.24%**

Table 3. Evaluation result across all clarification categories on Pass@1, Test Pass Rate, communication rate, and Good Question Rate with different models on HumanEvalComm (*HmEvalComm* in the table). Additionally, the Pass@1 and Test Pass Rate on the original problems in HumanEval (*HmEval* in the table) are also shown. Top 3 results are marked as **bold**.

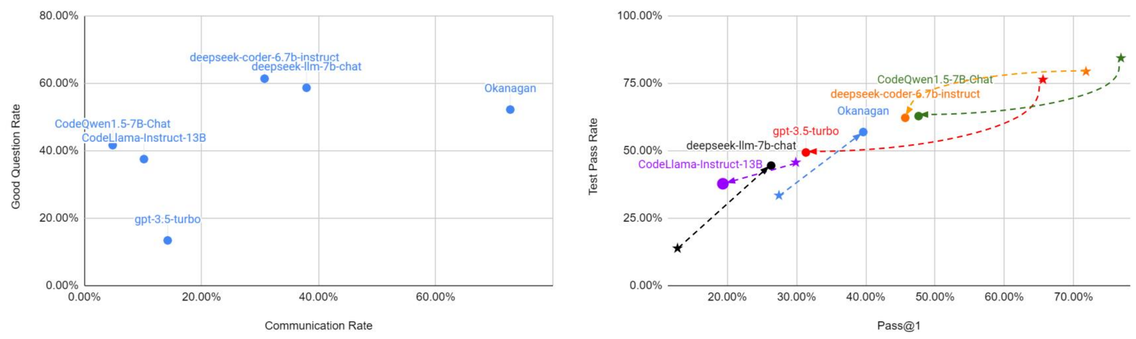

Fig. 4. Comparison of the effectiveness of the models in Communication Rate, Good Question Rate (left), and Pass@1, Test Pass Rate (right). Note that in the right figure, the stars represent the original performance of the corresponding model with the same color in the HumanEval benchmark. This shows visually how the performance has changed when the problem description is modified.

Answer to RQ1: More than 60% of responses from Code LLMs still generate code rather than ask questions when the problem descriptions are manually modified according to different clarification categories.

Incompleteness category results in higher communication rates and Good Question Rates, but lower Pass@1 and Test Pass Rate for Code LLMs.

Answer to RQ2: Okanagan, as a LLM agent approach that uses LLM (specifically ChatGPT), effectively increases all 4 metrics in HumanEvalComm. This indicates headroom for achieving more effective communication capability using LLM agent.

RESEARCH POSTER PRESENTATION DESIGN © 2019
www.PosterPresentations.com