Literature Review: Evaluating Large Language Models as a Judge in Code Generation Tasks

Tim Boleslawsky

I. Introduction

Large Language Models (LLMs) have become increasingly central to automated code generation and evaluation tasks. Their ability to both produce solutions and assess responses has opened new research avenues, but also raised questions about reliability, bias, and alignment with human judgment. A crucial research strand focuses on *LLMs as a judge*, where a model evaluates whether a generated response meets some qualitative criteria or the requirements of an input request. This literature review surveys the current status of research concerning the concept of LLM as a judge. Test Li et al. [2024]

References

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024. URL https://arxiv.org/abs/2412.05579.