Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Teodora Stojcheska
Thesis title	Fine-tuning Code Generation Models with Compiler Feedback
Year submitted	2024
Study program	Computer Science
Specialization	Artificial Intelligence
Review author	Mgr. Gabriela Kadlecová — Reviewer
Department	Department of Theoretical Computer Science and Mathematical Logic

Overall

	good	OK	poor	insufficient
Assignment difficulty	X	X		
Assignment fulfilled		X		
Total size *... text and code, overall workload*		X		

The thesis topic is a very recent problem of improving the performance of relatively small code generation models using compiler feedback. The author has chosen fine-tuning methods that are more stable than standard reinforcement learning approaches. The student compared the performance of the methods on two target benchmarks.

The difficulty of the assignment is high, as it requires large amounts of GPU memory and a precise setup of input data. It is to some extent facilitated by available libraries for large language model inference or fine-tuning.

The student has successfully compared the fine-tuning methods, included an ablation study, and experimented with a wide range of method configurations. A great plus is that all created datasets and the models are publicly available in the Hugging Face hub. The work has some notable weaknesses in the form and structure of the thesis text, but the good implementation and results outweigh the negatives.

Thesis Text

	good	OK	poor	insufficient
Form *... language, typography, references*			X	
Structure *... context, goals, analysis, design, evaluation, level of detail*		X	X	
Problem analysis		X		
Developer documentation		X		
User Documentation		X	X	

Formally, the theoretical part (chapters 1 and 2) includes all necessary background information and related work. The sections on the proposed approach and results (chapters 3 and 4) are well written, and the results of the experiments are presented in a clear way.

However, the level of detail of some sections (chapter 1 and 2) is too low, and the text is sometimes unclear. The student also omitted important parts of the methodology that are found in the code but not in the text. The first case is the definition of the prompt variants, and the second case is how exactly the DPO negative examples were constructed.

A major weakness are citations – some passages in the theoretical part lack an appropriate citation (e.g. section 1.3.2 includes only 1 citation). The citation frequency of other sections could be improved, e.g. section 1.3.1 has a citation at the start of the text, but other paragraphs should also have a reference, ideally from multiple sources.

The documentation is in the form of a README and well-written. However, it is unclear which config files lead to the results presented in the thesis text.

Minor suggestions:

- Named paragraphs in the introduction are not commonly used
- Section 2.1 describes again what an MDP is, but could just reference section 1
- Section 2.3 lambdas are not defined
- Section 1 – ML is not defined as an abbreviation, and Mean square error and Cross entropy losses are not defined.
- The hyperparameter configuration and the exact size of input data of the best results should be added in the text.

Despite the mentioned flaws, the thesis text satisfies the formal requirements of a Bachelor thesis.

Thesis Code

	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*	X			
Implementation *... naming conventions, formatting, comments, testing*	X			
Stability		X		

The code quality is very good – it is well-structured and configurable, and provides good installation instructions. The student has a clean experimental setup, uses the hydra library for clean configuration file management, and properly logs the results in Weights and Biases. The finished experiments are saved on the Hugging Face hub and enable fast evaluation of different models and datasets.

The results are reproducibile, selected config files lead to the same results as reported in the text. However, the results of the fine-tuning methods differ notably across config files, the author should provide instructions for choosing the best settings, or include a hyperparameter optimization method.

Overall grade Very Good (worse)

Award level thesis No

Date: 26. 8. 2024 Signature