Supervisor's review of doctoral thesis

Student:	Mgr. Tomasz Limisiewicz
Thesis Title:	Interpreting and Controlling Linguistic Features in Multilingual Language Models
Supervisor:	RNDr. David Mareček, PhD.
Department:	Institute of Formal and applied Linguistics, MFF UK

Tomasz's thesis on the interpretation of Large Language Models (LLMs) addresses a highly relevant topic. While LLMs become increasingly integrated into a wide range of applications, the research on understanding of their inner workings is still in its early stages. Tomasz approached this topic from several points of view. He examined different parts of Transformer (word embeddings, attentions, hidden states, and feed-forward layers), evaluated on various NLP tasks (syntactic parsing, lexical semantics, question answering, gender bias), and also focused on multilingual aspects.

During Tomasz's PhD studies, his work evolved in line with the current trends in the field of Natural Language Processing.

In 2020, he started his work by analysing self-attention matrices in the BERT model and comparing them to syntactic annotations in the Universal Dependencies framework. His findings revealed that while syntactic relations are indeed captured in the attention matrices, they do not often match one-to-one. He also proposed a method for relation identification and syntactic tree construction. (Chapter 5)

In 2021, Tomasz began investigating the contextual embeddings of the BERT model. He introduced a method called "Orthogonal probing" for transforming hidden states into representations and showed its ability to separate syntactic and lexical aspects of words. He further extended his research to multilingual BERT, demonstrating that typologically distinct languages are encoded in mutually isomorphic subspaces. (Chapter 4)

In 2022, Tomasz turned his focus to input embeddings of multilingual models (XLM-Roberta). He proposed metrics for assessing the quality of lexical representations and vocabulary overlap in sub-word tokenizers. Among his findings, he observes that the coverage of the language-specific tokens in the multilingual vocabulary significantly impacts the word-level tasks. (Chapter 7)

In 2023, as the number of parameters in Transformer models continued to grow, Tomasz shifted his focus to feed-forward layers, which became the most parameter-heavy components. Employing a causal tracing method, he was able to detect which part of the network is responsible for learning particular phenomena. He centered his research on mitigating gender bias, and proposed DAMA tool (Debiasing algorithm through model adaptation), which is able to update selected feed-forward parameters in LLMs and weaken the selected bias of the model. (Chapter 6)

The thesis itself is well-structured and thoughtfully organized. After the introduction showing the motivation and objectives, Chapter 2 provides background on Transformer language models. Chapter 3 introduces the basic metrics used and describes the tasks employed throughout the research. The following four chapters present Tomasz's original research. Chapter 8 offers overall discussion and comparison with other related works. The final chapter concludes the thesis, addressing its limitations and outlining future challenges.

Tomasz worked very independently. I never needed to guide him in any particular direction. He always came with new ideas, with the new research proposals and actively studied related existing papers. His publication record is excellent. Over the course of his four-year studies, he published 12 papers, many of which appeared in top-tier conferences, including ACL (twice), EMNLP (twice), ICLR, NAACL, AACL, and ACL Findings (twice). Tomasz also broadened his academic experience by visiting two other universities. He completed a four-month internship at the Hebrew University of Jerusalem and another four-month internship at the University of Washington. Both internships led to ongoing collaborations and resulted in interesting scientific papers. In the last year, Tomasz supervised two master's students, both of them defended their theses with excellent grades. He also contributed to teaching a new course on Large Language Models.

Overall, I consider Tomasz's doctoral thesis to be of very high quality, and I fully recommend it to be approved for a PhD.

Prague, 21st August, 2024

David Mareček
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University