Supervisor's review of the dissertation
Title: Machine learning with Human in the Loop for textual augmentation in the era of LLM
Student: Ing. Ján Čegiň
Supervisor: doc. Ing. Jakub Šimko, PhD.

Jan's work focuses on the question of how to efficiently augment textual datasets. At the beginning of his 4 years study, he identified adversarial augmentation as a good research topic, as it fitted well the broader scope of the thesis: machine learning with human-in-the-loop. Humans, being creative, are very adept in exploring (and exploiting) loopholes and weak spots of machine learned models. Therefore, using the concept of games with a purpose, Jan created an approach for collecting adversarial examples for text classification tasks and investigated how these examples strengthen the models after re-training thanks to enhanced distribution of the newly augmented training sets.

This initial research was then met with a global event: the release of ChatGPT (later followed by the proliferation of other LLMs and LLM-based services). What immediately struck the dissertation concept at that time, was the question about the future of human computation in the era of LLMs. Jan, following the text augmentation theme, conducted a study, in which he replaced human workers with LLM in a previously known crowdsourcing task setup and investigated the quality of acquired augmentations. Jan then continued exploring the data augmentation capabilities of LLMs: first by investigating the effects of diversity incentives (originally used in crowdsourcing scenarios) and then, by comparing the LLMs with traditional NLP augmentation techniques. He then also explored the effects of various seed sample selection strategies in few shot augmentation scenarios.

Overall, Jan's work has contributed to the understanding of benefits and limits of LLMs in formerly human-dependent dataset creation and augmentation tasks. His contributions are underlined with solid publication activity, with especially valuable full papers published in top NLP research venues (ACL, EMNLP and NAACL).

Besides that, Jan has demonstrated his ability to effectively address research problems, conduct research, and collaborate within research teams. He is a proactive thinker, coming up with research ideas and decision options, which he then efficiently discusses. He is also a skilled implementer of research designs and very efficient in experiment execution. The cooperation with Jan was both excellent and inspiring. He consistently approaches his activities with a suitable balance of responsibility and enthusiasm for the subject matter. It was my pleasure to advise Jan on his PhD journey and I'm confident that he will put his acquired knowledge and skills to good use in further research projects.

In Bratislava, June 27th, 2025