SUPERVISOR'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Czech Foundational Large Language Model Corpus**
Author's name:	**Bc. Tommaso Gargiani**
Type of thesis :	Master's thesis
Faculty/Institute:	Faculty of Electrical Engineering
Department:	Department of Computer Science
Thesis supervisor:	Ing. Jan Šedivý, CSc.
Supervisor's department:	CIIRC

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

Evaluation of thesis difficulty of assignment.

The thesis primary goal was to design robust procedures for collecting and preprocessing a data corpus essential for training a foundational Large Language Model (LLM). The key was exploring diverse sources of internet data, including but not limited to platforms such as Common Crawl and Wikipedia. The attention was paid to finding the most effective methods for cleaning and deduplicating raw text, trying to find a balance between efficiency and quality.

Satisfaction of assignment **fulfilled**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

The student has delivered a comprehensive study of the task at hand. He reviewed a significant amount of literature and summarized it in the initial sections of the thesis. He then briefly described different sources of Czech text. In the following sections, he outlined step-by-step the required text processing. He also used the collected data samples to train and test small LLMs, measuring their perplexity. This experiment extended beyond the initial thesis assignment.

Activity and independence when creating final thesis **A - excellent**

Assess that student had positive approach, time limits were met, conception was regularly consulted and was well prepared for consultations. Assess student's ability to work independently.

Tommaso had a very positive and enthusiastic approach. We met weekly to discuss further steps and potential problems. Our discussions were always fast-paced and on topic, as Tommaso consistently came well-prepared with numerous new suggestions. During the preparation of the thesis, he worked very independently.

Technical level **B - very good.**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

Tommaso has studied and identified the latest and most successful methods for downloading and processing text for training of foundational models in Czech. Based on this initial study, he wrote his own scripts specifically and downloaded and tokenized around 200 billions of Czech tokens. This effort required extensive use of AWS cloud. The next step was cleaning, deduplication, tokenization requiring programming skills. Tommaso used lot of open-source resources to optimize his approach. Most of the processing has been done on the CIIRC and AWS clouds. It was an elaborate task involving many trial-and-error cycles. Most of the required knowledge is not part of the standard curriculum.

Formal and language level, scope of thesis **A - excellent.**

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

The thesis is clearly written, some of the parts are very brief, but this does not affect the overall quality of the work. It is aesthetically pleasing and uses the recommended LaTeX template. It is written in English and the language is clear and understandable, adhering to standard notations and other formal requirements.

Selection of sources, citation correctness **A - excellent.**

Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.

Tommaso systematically collected and studied a vast array of Internet and literature resources. Through this ongoing process, he was steadily discovering the field's newest and most sophisticated papers. Notably, he maintained deep attention to detail, ensuring all statements were properly referenced in accordance with ethical citation practices. His citations adhere to the standard format.

Additional commentary and evaluation

Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

Please insert your commentary (voluntary evaluation).

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation.

Tommaso's thesis advances the field of large language models in the Czech Republic by providing the largest dataset for training purposes. This is a significant contribution that will benefit numerous researchers.

I evaluate handed thesis with classification grade **A - excellent.**

Date: **7.6.2024** Signature: