Using AI to Grade Student Tests in SUI

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Authors: Bc. Martin Hemza
Bc. Petr Kaška
Bc. Jakub Magda

Supervisor: Ing. Michal Hradiš Ph.D.

Task

The goal is to automatically grade short written answers on a 0–4 scale. As a baseline, we fine-tuned a multilingual BERT model in a classification setup, using 17 classes corresponding to 0.0–4.0 scores in 0.25 steps. The model processed concatenated question-answer pairs and predicted a score label, which was then evaluated as a regression output. This simple method achieved a solid baseline with a test MAE of 1.64. The results confirmed that even a basic model can capture useful patterns in human grading and serve as a meaningful point of comparison for more advanced approaches.

Dataset

The dataset contains student exam answers digitized through OCR and corrected for mistakes. We selected 10 questions with varied score distributions for our experiments.

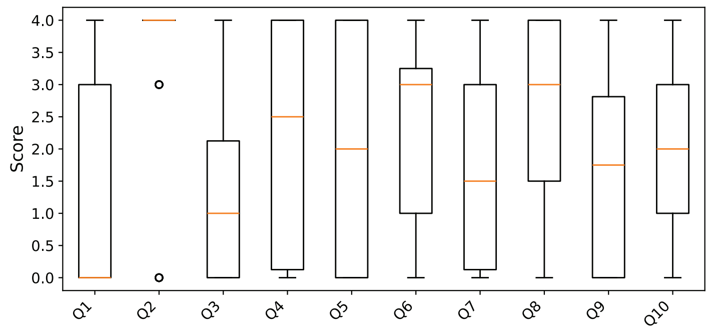

Figure. 1. Selected dataset for experiments.

Fine tuning models

We fine-tuned three Czech-compatible language models for automatic short-answer grading using parameter-efficient LoRA. Despite limited compute, Czech-GPT-2-XL-133k achieved the best performance (MAE: 0.994, MSE: 2.046), surpassing larger models like Mistral-7B. These results highlight the effectiveness of smaller, Czech-trained models for this task.

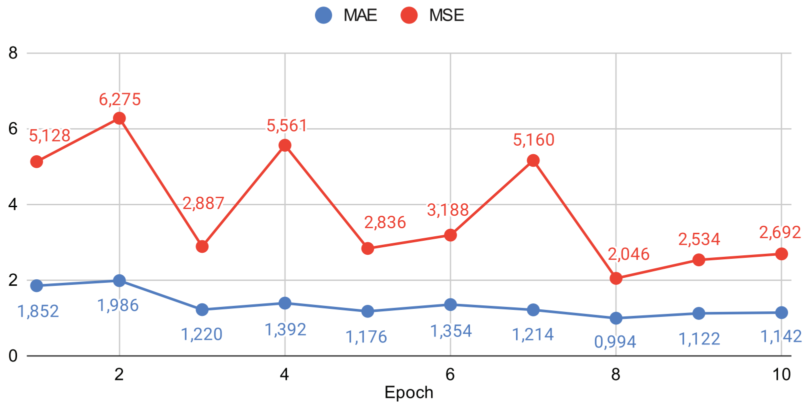

Figure. 2. Fine-Tuning Czech-GPT-2-XL-133k

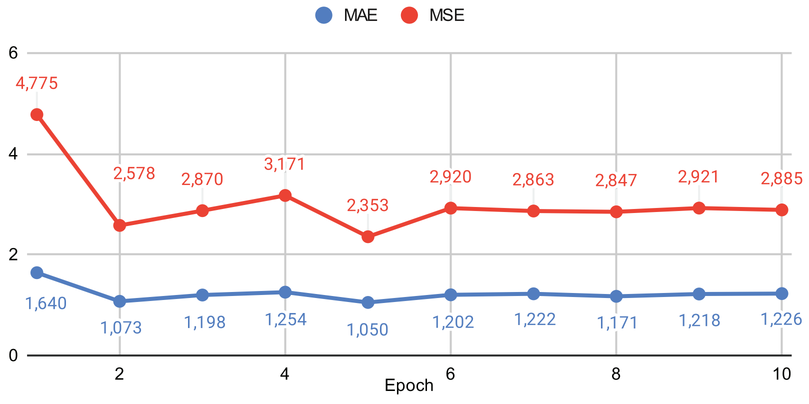

Figure. 3. Fine-Tuning Mistral-7B-Instruct-v0.3

GPT API

By default, the o4-mini model systematically underestimated students across all five questions, with median absolute errors lying around 1-2 points and isolated fluctuations ranging up to 4 points as shown on Figure 4.

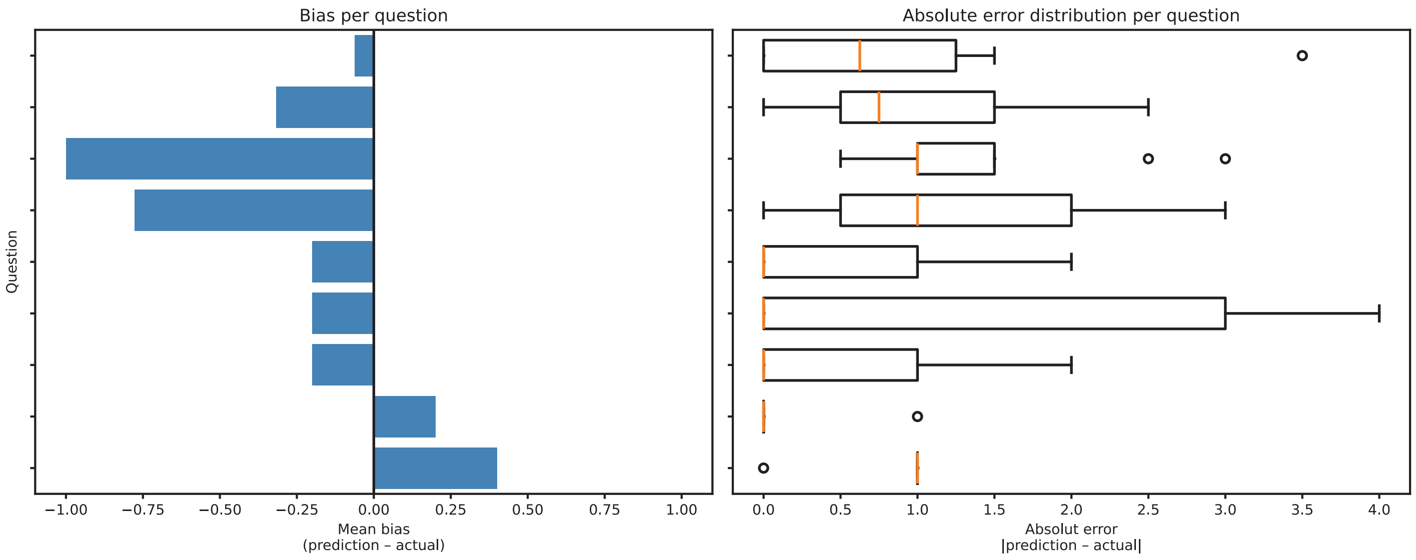

Figure. 4. Prompting GPT model o4-mini