Contribution ID: **150** Type: **Research Presentation (30 minutes)**

Flavoring the Vanilla: Finetuning Large Language Models for Automated Evaluation of Argumentative Writing

Saturday, 18 May 2024 10:50 (30 minutes)

To address the long-standing challenge facing traditional automated writing evaluation (AWE) systems in assessing higher-order thinking, this study built an AWE system for scoring argumentative essays by finetuning the GPT-3.5 Large Language Model and compared the system's effectiveness with that of the non-finetuned GPT-3.5 and GPT-4 base models, or "vanilla" models, using zero-shot prompting methods. The dataset used was the TOEFL Public Writing Dataset provided by Education Testing Service, containing 480 argumentative essays with ground truth scores under two essay prompts. Three finetuned models were generated: two finetuned exclusively on either prompt and one on both. All finetuned and base models were used to score the remaining essays after finetuning and their scoring effectiveness was compared with ground truth scores as the benchmark. The impact of the variety of finetuning prompts and the robustness of finetuned models were also explored. Results showed a 100% consistency of all models in two scoring sessions. More importantly, the finetuned models significantly outperformed the base models in accuracy and reliability. The best-performing model, finetuned on prompt 1, showed an RMSE of 0.57, a percentage agreement (score discrepancy ≤ 0.5) of 84.72%, and a QWK of 0.78. Further, the model finetuned on both prompts did not exhibit enhanced performance, and the two models finetuned on one prompt remained robust when scoring essays from the alternative prompt. These results suggest 1) task-specific finetuning for AWE is beneficial; 2) finetuning does not require a large variety of essay prompts; and 3) fine-tuned models are robust to unseen essays.

Is this a sponsored session?

Keywords

Automated Writing Evaluation, Large Language Models, GPT-3.5, Finetuning, TOEFL Public Writing Dataset

Primary authors: WANG, Qiao (Waseda University); Dr GAYED, John (Waseda University)

Presenters: WANG, Qiao (Waseda University); Dr GAYED, John (Waseda University)

Session Classification: DN 411: Mixed Sessions

Track Classification: Artificial Intelligence CALL: AI for Teaching