ISLab at the NTCIR-18 AEOLLM: An Evaluator for Machine-Generated Text based on Data Augmentation and ORPO

Chia-Hui Lin
Department of Computer Science and Information Engineering National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
c111151162@nkust.edu.tw

Cen-Chieh Chen
Department of Computer Science and Information Engineering, National Taiwan Normal University
Taipei, Taiwan, R.O.C
81147003s@ntnu.edu.tw

Tao-Hsing Chang
Department of Computer Science and Information Engineering National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
changth@nkust.edu.tw

Fu-Yuan Hsu
Research Center for Psychological and Educational Testing National Taiwan Normal University
Taipei, Taiwan, R.O.C
kevin@rcpet.ntnu.edu.tw

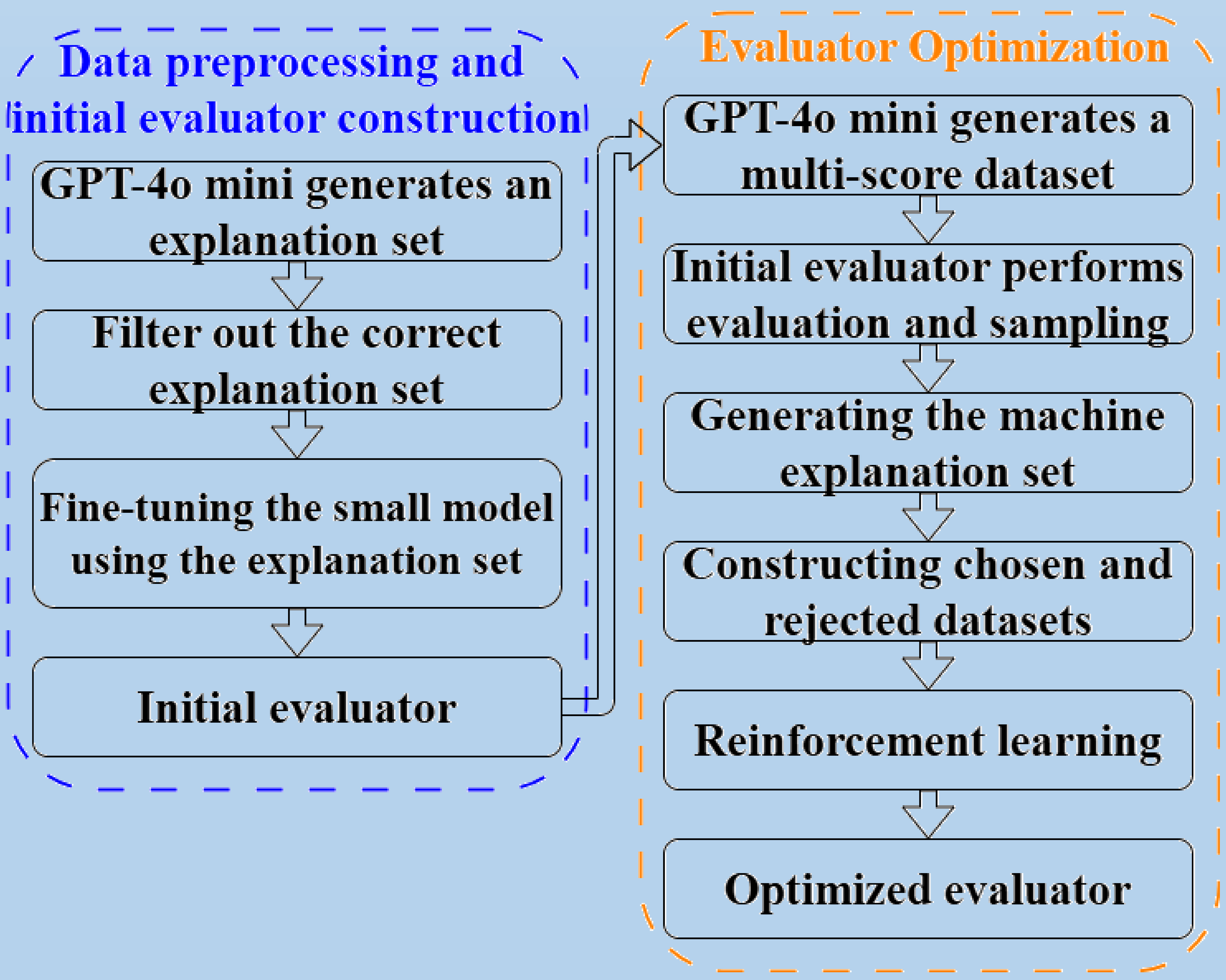
Abstract

In recent years, large language models (LLMs) excel in natural language processing (NLP) tasks, and many studies use one LLM to evaluate others, performing well on public benchmarks. However, their effectiveness on unpublished data is limited. Fine-tuning improves performance but requires extensive labeled data, making it costly and less practical for widespread use.

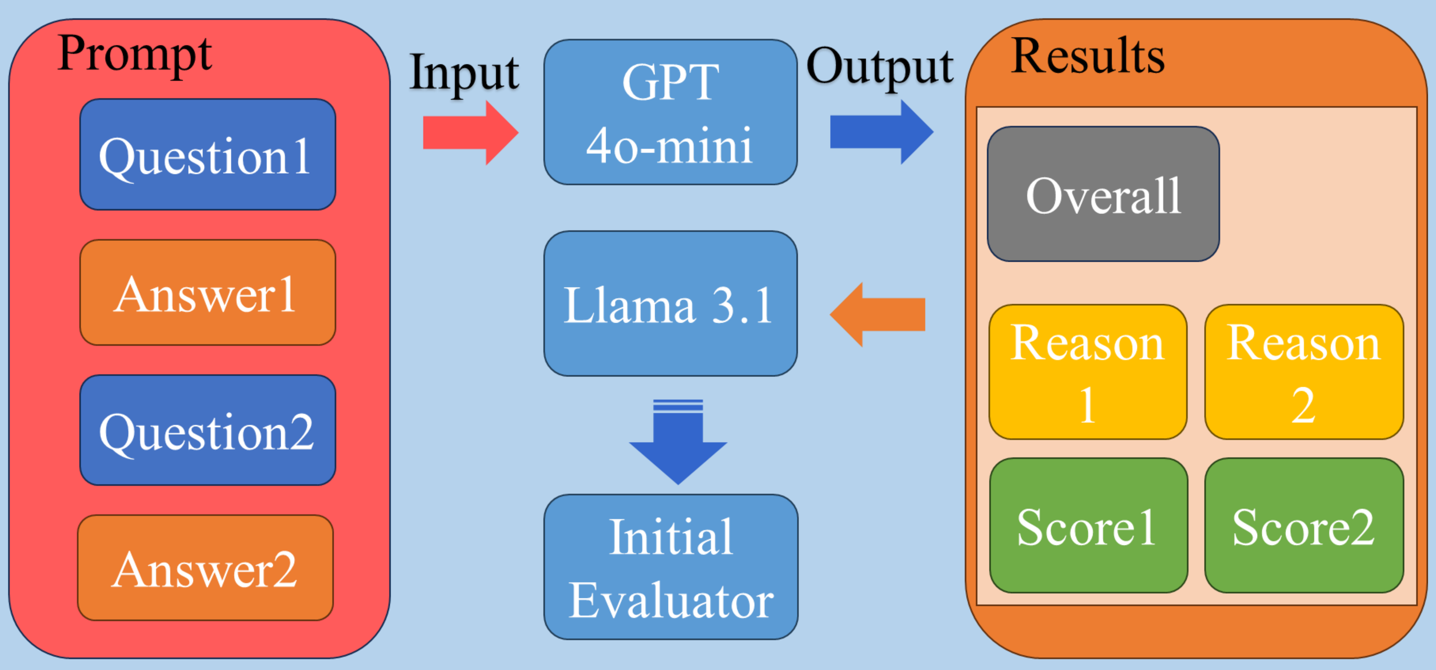
Therefore, our study leverages data augmentation to increase the volume of training data and employs the odds ratio preference optimization (ORPO) algorithm for reinforcement learning to optimize the evaluator. This study uses the dataset provided by NTCIR-18's Automatic Evaluation of LLMs (AEOLLM) task for training and testing.

The proposed method achieves an accuracy of 0.7658 on the summary generation subtask of AEOLLM, the highest among all compared models. Additionally, it yields the second-highest performance in both Kendall's tau and Spearman correlation coefficient on the summary generation and text expansion subtasks among all compared models.

Architecture of our proposed method

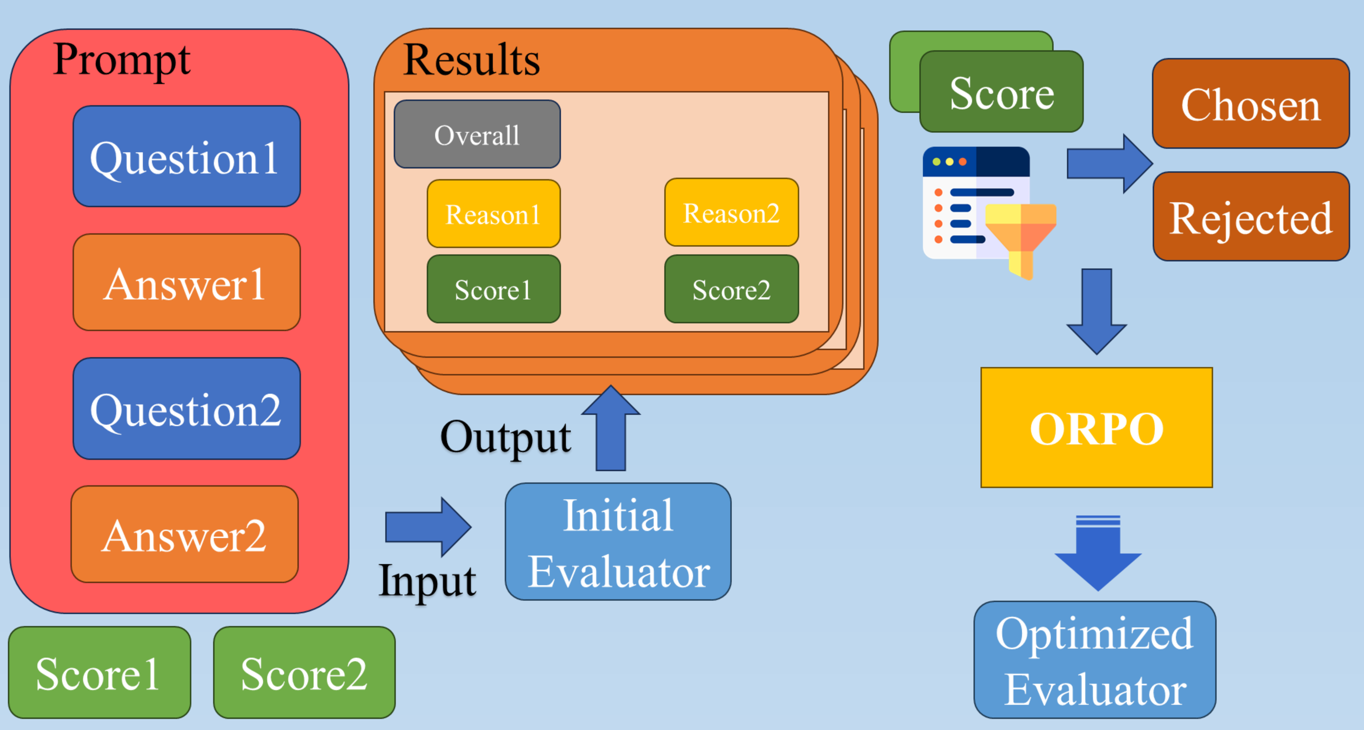


Knowledge Distillation



Leveraging a smaller model that mimics GPT-4o mini's scoring and explanation abilities using knowledge distillation. By fine-tuning with explanation sets and applying supervised learning with cross-entropy loss, the model achieves efficiency performance, offering a lightweight alternative to the larger initial evaluator.

Evaluator Optimization



Using temperature sampling and the initial evaluator to categorize generated data into chosen and rejected sets. Reinforcement learning with the ORPO algorithm to refine evaluator, encouraging accurate assessments from chosen data and discouraging errors, ultimately transforming it into a more optimized and reliable evaluator.

Motivation

Large language models (LLMs) have been widely applied to various natural language processing (NLP) tasks, demonstrating exceptional performance. To evaluate the output quality of these LLMs, numerous studies utilize one **LLM as an evaluator** to assess the quality of outputs from other LLMs.

However, the performance of LLMs as evaluators on many unpublished benchmarks still needs improvement. **To achieve better evaluation performance**, some studies have attempted **to fine-tune evaluators based on large amounts of data**, incurring significant **manual costs** and posing substantial limitations in practical applications.

Therefore, our study leverages **data augmentation** to increase the volume of training data and employs the odds ratio preference optimization (ORPO) algorithm for reinforcement learning to **optimize the evaluator**.

A series of crucial experiments

This study uses two datasets for each subtask. For summary generation, the AEOLLM training set serves as the human score set, and 200 unused samples from XSum, sourced from BBC news articles, form an additional original set. For text expansion, the AEOLLM training set is again used, along with 200 unused samples from WritingPrompts, derived from online forum content. The AEOLLM test set is used to evaluate our method's effectiveness using accuracy, Kendall's tau, and Spearman correlation.

Metrics / Prompt	Accuracy	Kendall's tau	Spearman
Scoring-Based	0.7685	0.5139	0.5611
Comparison-Based	0.7659	0.5636	0.6164

Table 1: Comparison of Prompt Design Methods

The experiment compares scoring-based and comparison-based prompts using GPT-4o mini as the evaluator. As a closed model, GPT-4o mini is limited in flexibility to fine-tune. Results show it performs better with comparison-based prompts, which are therefore used in all subsequent experiments.

Metrics / Fine-tuned	Accuracy	Kendall's tau	Spearman
Unused	0.7293	0.5010	0.5416
Used	0.7703	0.5790	0.6242

Table 2: Impact of Fine-Tuning on Performance

Fine-tuning significantly improves the Llama 3.1 evaluator's performance across all metrics. The fine-tuned Llama 3.1 evaluator, referred to as the initial evaluator, even outperforms GPT-4o mini. This aligns with prior studies, despite Llama 3.1 generally underperforming GPT-4o mini as a general-purpose model.

Metrics / Method	Accuracy	Kendall's tau	Spearman
Initial evaluator	0.7703	0.5790	0.6242
Optimized evaluator	0.8003	0.6065	0.6664

Table 3: Effectiveness of Data Augmentation & RL

This comparison between the initial and optimized evaluators shows that data augmentation and reinforcement learning significantly improve evaluation performance, highlighting the effectiveness of these techniques in enhancing model accuracy and reliability.

Evaluation Results

Metrics / Prompt	Accuracy	Kendall's tau	Spearman
Text Expansion	0.5241	0.3609	0.4035
Summary Generation	0.7658	0.5117	0.5632

Table 4: Optimized Evaluator Performance on Reserved Set

Conclusions

This study presents an evaluator built on Llama 3.1 to assess LLM-generated text quality. Optimized using data augmentation and ORPO-based reinforcement learning, it outperforms existing models on key evaluation metrics in the NTCIR-18 AEOLLM task, demonstrating the effectiveness of the proposed method.

Our proposed method holds the potential for further improvement and development in the future. Specifically, the current approach relies on human-authored generation and evaluation prompts, which is labor-intensive and may not yield optimal prompt design.

To address this, we intend to integrate automated prompt-generation techniques like TextGrad. This automation will replace manual prompt engineering, potentially enhancing the effectiveness of data augmentation and overall evaluator performance.

NKUST

NATIONAL TAIWAN NORMAL UNIVERSITY

NTCIR