UCLWI at the NTCIR-18 AEOLLM Task: A Low-Cost Comparison of RAGs

Xiao Fu[♠] Navdeep Singh Bedi[♡] Noriko Kando[♢] Fabio Crestani[♡] Aldo Lipani[♠]

University College London[♠] Università della Svizzera italiana[♡] National Institute of Informatics[♢]

[♠]{xiao.fu.20, aldo.lipani}@ucl.ac.uk [♡]{navdeep.singh.bedi, fabio.crestani}@usi.ch [♢]Noriko.Kando@nii.ac.jp

Abstract

The UCLWI team participated in the Automatic Evaluation of LLMs (AEOLLM) task of the NTCIR-18 [1]. We propose an efficient evaluation pipeline for Retrieval-Augmented Generation (RAG) systems tailored **for low-resource settings**. Our method uses ensemble similarity measures combined with a logistic regression classifier to assess answer quality from multiple system outputs using only the available queries and replies. Experiments across diverse tasks demonstrate competitive accuracy and reasonable correlation with ground truth rankings, establishing our approach as a reliable metric.

Introduction

In this study, we address the challenge of RAG evaluation in low-resource environments, where neither additional corpora nor GPUs are available. Our pipeline, developed within the NTCIR AEOLLM framework, processes single queries answered by multiple RAG systems [1]. By leveraging an ensemble strategy that analyzes the similarity between generated answers and their corresponding queries, our method not only achieves the highest agreement but also attains the second-best accuracy relative to the ground truth. These results underscore its potential as an efficient and effective evaluation strategy for RAG systems.

The Pipeline

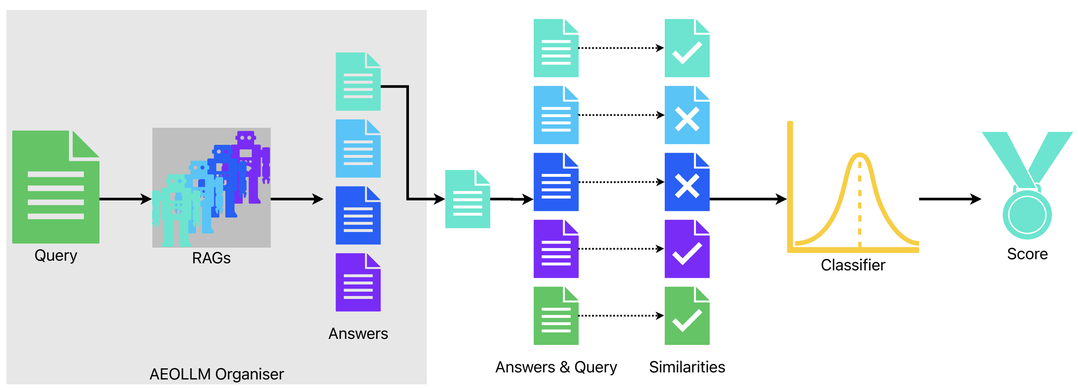

Figure 1: Pipeline based on an ensemble

Similarity Model To minimize computational cost while effectively calculating similarities, we utilize *static-similarity-mrl-multilingual-v1*[1] from *sentence-transformers* [2]. Unlike Transformer-based models (e.g., *all-MiniLM-L6-v2*[2]), this static model encodes text chunks into 1024-dimensional vectors, offering a balance between performance and computational efficiency. In our study, we compute similarities by obtaining embeddings for text chunks and measuring the cosine similarity between them.

Classifier The classifier employs a multinomial logistic regression model. For each prediction, the input feature vector is 8-dimensional, comprising seven components representing the similarity scores between the candidate reply and each of the other replies $(f_1, \ldots, f_7)$, along with one component capturing the similarity between the reply and the query (f_q). The feature matrix $F \in \mathbb{R}^{n\times 8}$ is standardized to have zero mean and unit variance. The model is configured for multi-class classification using the multinomial scheme with the 'lbfgs' optimizer, running for a maximum of 1000 iterations. The model parameters are optimized by maximizing the multinomial log-likelihood.

[1]https://huggingface.co/sentence-transformers/static-similarity-mrl-multilingual-v1

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Results

Task	Accuracy	Kendall's Tau	Spearman
Test Set			
Dialogue Generation	0.7044	0.4913	0.5527
Text Expansion	0.5340	0.2758	0.3019
Summary Generation	0.7494	0.6114	0.6418
Non-Factoid QA	0.6905	0.4090	0.4311
Overall	0.6696	0.4468	0.4819
Final Set			
Dialogue Generation	0.7756	0.5798	0.6426
Text Expansion	0.5266	0.3482	0.3815
Summary Generation	0.7273	0.5432	0.5763
Non-Factoid QA	0.6853	0.4105	0.4291
Overall	0.6787	0.4704	0.5074

Table 1: Performance of our pipeline.

Our proposed pipeline performed promisingly on both the test and final sets. On the test set, we obtained an overall accuracy of 0.6696, with moderate agreement levels (above 0.4) observed in three out of four tasks. Similarly, on the final set, we achieved an overall accuracy of 0.6787, with agreement levels similar to those on the test set.

At the task level, in particular, Dialogue Generation and Summary Generation yielded relatively higher performance, followed by Non-Factoid QA, while Text Expansion consistently demonstrated the lowest performance. The significant variance in performance across tasks, which appears to stem from inherent differences in each task's characteristics. A clear trend emerges: tasks with longer queries and shorter replies tend to yield higher accuracy and better agreement in our pipeline. This trend can be interpreted from two perspectives. On one hand, generation-based systems benefit from richer contextual information provided by longer queries, often leading to more accurate responses. On the other hand, the nature of the task itself plays a significant role. For instance, Summary Generation tasks require condensing text into concise, less diverse outputs, whereas Text Expansion tasks—where systems generate narratives from a given theme—tend to produce more varied responses. Consequently, the similarities among summaries are generally higher than those among expanded texts. This observation underscores a limitation of our pipeline: it relies on surface-level similarity metrics rather than a deeper semantic understanding to rank replies.

Conclusions

In summary, our study presents a comprehensive evaluation of RAG systems through a novel pipeline that leverages similarity metrics and classifier-based scoring. The analysis of task-level performance, classifier weights, and the role of query-to-reply similarities provides valuable insights into the strengths and limitations of our approach. While the reliance on similarity metrics offers computational efficiency, it also highlights the need for incorporating deeper semantic understanding in future work.

References

[1] Junjie Chen, Haitao Li, Zhumin Chu, Yiqun Liu, and Qingyao Ai. Overview of the ntcir-18 automatic evaluation of llms (aeollm) task. *arXiv preprint arXiv:2503.13038*, 2025.

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.