KNUIR at the NTCIR-18 AEOLLM: Automatic Evaluation of LLMs

Yumi Kim[1], Meen Chul Kim[2], and Jongwook Lee[1,*]

[1] Department of Library & Information Science, Kyungpook National University (KNU), 80 Daehak-ro, Buk-gu, Daegu, Republic of Korea
[2] Children's Hospital of Philadelphia, 3401 Civic Center Blvd. Philadelphia, U.S.A.
*Corresponding author: jongwook@knu.ac.kr

Abstract

This study explores automated evaluation methods for large language models (LLMs) that approximate human judgment, comparing two distinct approaches: ***(1) LLM-based scoring*** using GPT models with prompt engineering, and ***(2) feature-based machine learning*** using transformer-based metrics such as BERTScore, semantic similarity, and keyword coverage. As part of this research, we participated in the NTCIR-18 Automatic Evaluation of LLMs (AEOLLM) task, submitting predictions for both the test and reserved datasets and analyzing the evaluation results. The results show that GPT-4o Mini with updated prompting achieved the highest performance. The feature-based approach performed competitively, outperforming GPT-3.5 Turbo and showing only a small gap with GPT-4o Mini. LLM-based methods offered scalability but lacked explainability, while feature-based approaches provided greater interpretability but required extensive tuning, illustrating the trade-offs between the two strategies. We hope our findings contribute to a deeper understanding of human judgment and support the development of more effective automated evaluation methods for LLMs.

Research Motivation & Objectives

- As large language models (LLMs) continue to evolve, there is a growing need for ***automated evaluation methods*** that are both ***scalable and aligned with human judgment***, the gold standard for quality assessment, yet ***costly and time-consuming*** to apply at scale.
- This study compares ***LLM-based scoring*** and ***feature-based machine learning*** approaches to understand:
  - How well each aligns with human evaluations across different task types
  - What trade-offs exist between scalability, interpretability, and performance
  - Which approach (or combination) may offer the most reliable, efficient evaluation framework for future LLMs

Methodology

LLM-based Approach

- Utilized GPT models (GPT-3.5 Turbo and GPT-4o Mini) to predict human evaluation scores.
- Designed task-specific prompts to instruct the models to score responses on a 5-point scale.
- Conducted prompt engineering to improve alignment with human judgment:
  - Incorporated ***dataset descriptions*** to provide contextual grounding.
  - Updated prompts included explicit ***evaluation criteria*** (Relevance, Conciseness, Clarity, Accuracy)
- Compared base and updated prompts to analyze the effectiveness of prompt engineering across tasks.

Feature-based Approach

- Utilizes machine learning models trained on features extracted from LLM-generated responses.
- Extracted features reflect quality dimensions such as:
  - Semantic similarity
  - Factual correctness
  - Fluency
  - Coherence
- A common set of core features, including BERTScore, semantic similarity, and keyword coverage, was applied across all tasks.
- Additionally, task-specific features were selected to capture the unique characteristics of each evaluation task.

1 Extracted features for each task

Task	Features
Summary Generation (SG)	**BERTScore, Semantic Similarity Score, Keyword Coverage Score**, Topic Similarity Score, Fact Extraction Score
Non-Factoid QA (NFQA)	**BERTScore, Semantic Similarity Score, Keyword Coverage Score**, Topic Similarity Score, Fact Extraction Score
Text Expansion (TE)	**BERTScore, Semantic Similarity Score, Keyword Coverage Score**, Grammar Error Rate, Coherence Score, Lexical Diversity
Dialogue Generation (DG)	**BERTScore, Semantic Similarity Score, Keyword Coverage Score**, DialoGPT Score, Sentiment Alignment Score, Intent Alignment Score

Results

Dry Run Results (Test Dataset)

No	Methods	Accuracy	Kendall's Tau	Spearman
0	GPT-3.5 Turbo with base prompt (LLM-based apporach)	0.5846	0.3554	0.3824
1	GPT-3.5 Turbo with updated prompt (LLM-based apporach)	0.6246	0.3717	0.3987
2	GPT-4o Mini with updated prompt (LLM-based apporach)	0.6710	0.4412	0.4730
3	Feature-based approach	0.6416	0.4294	0.4535

- GPT-4o Mini(LLM-based approach) achieved the highest average scores.
- Feature-based approach performed comparably to GPT-4o Mini and ***outperformed*** GPT-3.5 Turbo.

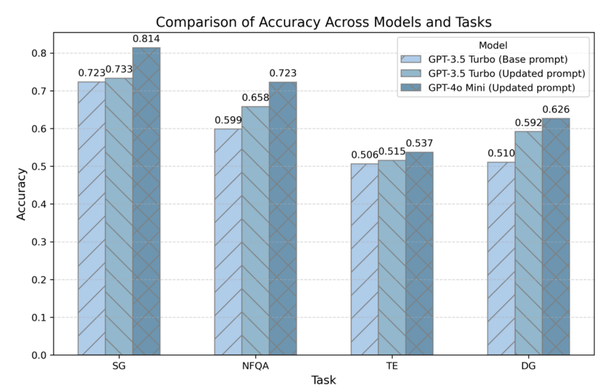
Formal Run Results (Reserved Dataset)

Task	Accuracy	Kendall's Tau	Spearman
Overall	0.6654	0.4043	0.4340
Dialogue Generation	0.6778	0.4404	0.4717
Text Expansion	0.5512	0.3141	0.3430
Summary Generation	0.7375	0.4524	0.4914
Non-Factoid QA	0.6951	0.4102	0.4297

- ***Structured tasks*** (e.g., SG, NFQA) are better handled by LLM-based evaluators.
- Creative tasks like TE remain more challenging for automated evaluation.

Key Findings

- ***LLM-based***
  - ***Prompt engineering*** improved performance in some tasks (e.g., NFQA, DG).
  - ***Model capability*** had a greater effect than prompt variation (GPT-4o Mini > GPT-3.5 Turbo).
  - GPT-4o Mini showed the highest alignment with human judgment, especially in SG.
  - LLMs still struggle with creative or dialogue-based tasks (TE, DG).
- ***Feature-based***
  - Performed well on ***structured tasks*** such as summarization and QA.
  - Weaker performance on open-ended or conversational tasks.
  - A viable alternative to LLMs for evaluating structured outputs.



Comparative Analysis and Insights

Advantages

- **[LLM-based scoring]**
  - Directly mimics human evaluation by generating scores in a similar way to human annotators.
  - ***Fast*** and scalable: LLMs can evaluate text without the need for manual feature extraction.
  - Adaptable via prompt engineering, allowing flexibility in evaluation criteria.
  - ***Requires no labeled training data***, making it useful in scenarios with limited annotations.
- **[Feature-based machine learning]**
  - ***Transparent and explainable***, as individual features can be analyzed.
  - More stable and reproducible.
  - More fine-grained control: different features can be weighted to improve performance.
  - Potentially ***more cost-effective*** after training, since it does not require querying an API like OpenAI's models

Disadvantages

- **[LLM-based scoring]**
  - Less explainable.
  - Prompt sensitivity.
  - Expensive for large-scale evaluation, as querying LLMs can be costly.
  - May not be fully aligned with human judgment.
- **[Feature-based machine learning]**
  - Feature engineering requires effort.
  - Limited adaptability.
  - Requires labeled training data (dependent on human-annotated scores for supervised learning).
  - Computational overhead: training models on extracted features requires additional resources.

Summary

- ***This study provides a comprehensive comparison*** of LLM-based and feature-based evaluation methods across both structured and open-ended tasks, highlighting their respective strengths and limitations.
- ***We also propose a feature-based evaluation framework*** as a cost-effective and interpretable alternative to LLM-based scoring—especially useful in resource-constrained or model-agnostic evaluation scenarios.