Overview of the NTCIR-18 Automatic Evaluation of LLMs (AEOLLM)

Junjie Chen[1], Haitao Li[1], Zhumin Chu[1], Yiqun Liu[1], Qingyao Ai[1],

1 Department of Computer Science and Technology, Tsinghua University,
Zhongguancun Laboratory, Beijing 100084, China

❖ Introduction

The Automatic Evaluation of LLMs (AEOLLM) task is a core task in NTCIR-18

- In recent times, the persistent advancement of LLMs has sparked a lot of interest.
- However, the rapid advancement of LLMs has introduced a key challenge in the progression of these models—— **efficiently and effectively evaluating their performance.**
- The existing LLM evaluation methods could be categorized into two groups: manual evaluation and **automatic evaluation**.
- However, existing automatic evaluation methods for LLMs still have the following limitations: (1) **Limited task format**. (2) **Limited evaluation criteria.**
- Based on these considerations, we propose the NTCIR-18 Automatic Evaluation of LLMs (AEOLLM) task, which: (1) **concentrates on generative tasks,** (2) **encourages reference-free evaluation methods.**
- To make our task more comprehensive, we set up multiple types of tasks including **dialogue generation, text expansion, summary generation and non-factoid question answering**.

❖ Task Framework

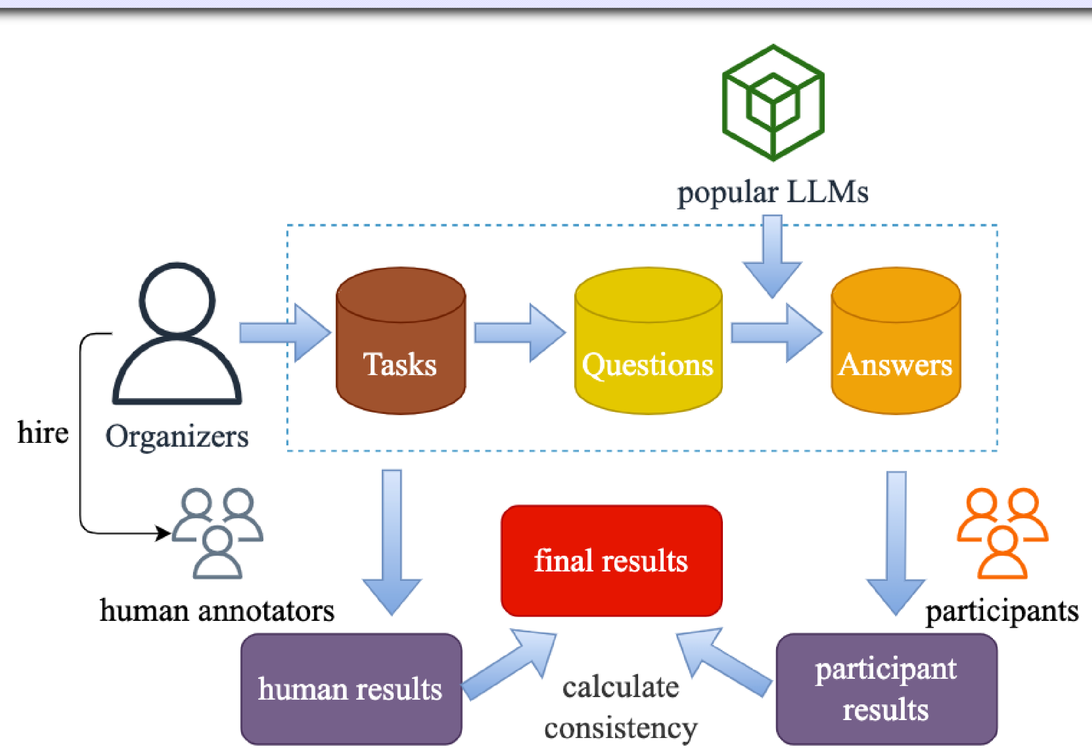

Figure 1: The overall framework of the AEOLLM task.

Table 2: NTCIR-18 AEOLLM run statistics.

Team	Dry run	Formal run	Total
KNUIR	7	2	9
ISLab	18	3	21
UCLWI	1	1	2
PanguIR	10	6	16
Total	36	12	48

Evaluation Metrics

- Accuracy (acc)
- Kendall's tau (τ)
- Spearman's Rank Correlation Coefficient (ρ)

❖ Official Results

Table 3: The results from the formal run on the reserved set. Baselines 1, 2, 3, and 4 correspond to direct prompting of ChatGLM3-6B, Baichuan2-13B, ChatGLM-Pro, and GPT-4o, respectively. The best result is highlighted in bold.

Team	Dialogue Generation			Text Expansion			Summary Generation			Non-Factoid QA			Overall		
	acc	τ	ρ	*acc*	τ	ρ	*acc*	τ	ρ	*acc*	τ	ρ	*acc*	τ	ρ
Baseline1	0.5583	0.3228	0.3495	0.5029	0.1236	0.1293	0.5976	0.2589	0.2712	0.6445	0.3717	0.3948	0.5759	0.2692	0.2862
Baseline2	0.5518	0.1647	0.1750	0.5021	0.0698	0.0725	0.6112	0.2693	0.2813	0.6044	0.2664	0.2791	0.5674	0.1926	0.2020
Baseline3	0.5924	0.2921	0.3135	0.5495	0.2831	0.2985	0.7079	0.4390	0.4575	0.6975	0.4682	0.4990	0.6368	0.3706	0.3921
Baseline4	0.6595	0.4423	0.4797	0.5543	**0.3963**	**0.4248**	0.7029	0.3886	0.4180	0.7441	**0.4896**	**0.5281**	0.6652	0.4292	0.4627
KNUIR	0.6778	0.4404	0.4717	0.5512	0.3141	0.3430	0.7375	0.4524	0.4914	0.6951	0.4102	0.4297	0.6654	0.4043	0.4340
ISLab	/	/	/	0.5241	0.3609	0.4035	**0.7658**	0.5117	0.5632	/	/	/	/	/	/
UCLWI	**0.7756**	**0.5798**	**0.6426**	0.5266	0.3482	0.3815	0.7273	**0.5432**	**0.5763**	0.6853	0.4105	0.4291	0.6787	**0.4704**	**0.5074**
PanguIR	0.7444	0.5611	0.6091	**0.5581**	0.3432	0.3775	0.7479	0.5097	0.5520	**0.7528**	0.4175	0.4534	**0.7008**	0.4579	0.4980

❖ Analysis

Comparing different methods

- overall, PanguIR achieves the best performance in terms of accuracy (acc), while UCLWI excels in Kendall's Tau (τ) and Spearman's Rank correlation coefficients (ρ)
- For each subtask, UCLWI excels in all three metrics for Dialogue Generation and in τ and ρ for Story Generation. PanguIR outperforms others in acc for Text Expansion and Non-Factual QA, and ISLab leads in acc for Summary Generation.

Comparing different evaluation metrics

- the results of τ and ρ are almost consistent
- acc sometimes differs from the results of these two coefficients

Comparing different subtasks

- the Text Expansion dataset is the most challenging
- Dialogue Generation is the easiest of the four tasks

❖ Conclusions

- AEOLLM this year received a total of 48 runs from 4 different teams, showcasing a variety of approaches to evaluating LLMs across four distinct subtasks: dialogue generation, text expansion, summary generation, and non-factoid question answering.
- (1) Comparing different methods, overall, PanguIR achieved the best performance in terms of accuracy (acc), while UCLWI excelled in Kendall's Tau (τ) and Spearman's Rank correlation coefficients (ρ).
- (2) Considering multiple metrics is necessary to provide a more comprehensive assessment of the performance of a method.
- (3) The Text Expansion dataset is the most challenging, with the highest acc being only 0.5581. This presents a challenging scenario for future method optimization.
- Looking ahead, we plan to further extend the AEOLLM task to better and more comprehensively evaluate LLMs.

Email: chenjj826@gmail.com