AEOLLM2: Automatic Evaluation of LLMs 2

- ***Challenge:*** Through AEOLLM, our findings revealed that current automatic evaluation methods of LLMs still remain limited in effectiveness when applied to long-form generation tasks, such as **text expansion**.
- ***Growing trend:*** Use LLMs to generate **Deep Research Reports:**

in-depth research	exceed 1500 tokens	Factual accuracy	structured organization	thematic coherence

- ***Existing Benchmarks:*** Remain focused primarily on short to medium-length texts (typically less than 500 tokens), leaving long-form scenarios largely underexplored.

Table 1: Statistics of benchmark datasets.

Dataset	Total	Average Length (tokens)			Label Count			Type	Type Count
		Prompt	Answer A	Answer B	First	Second	Tie		
RewardBench [11]	2985	439	165	163	1490	1495	0	Chat Chat_Hard Safety Reasoning	358 456 739 1432
MTBench [12]	2396	885	339	335	929	929	538	Turn1 Turn2	1204 1192
PreferenceBench [13]	1998	495	177	177	980	1018	0	-	-

- To address this gap, we propose a new subtask in **AEOLLM-2: Deep Research Evaluation.**
- This subtask focuses on the automated evaluation of long-form deep research reports generated by LLMs.
- Participants will be asked to develop evaluation methods that **automatically** score the quality of the generated reports.