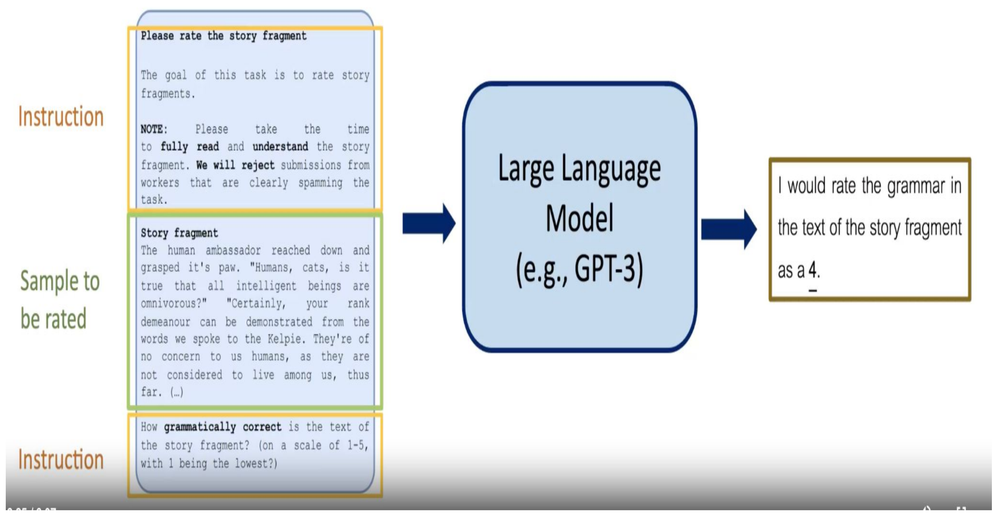
OFLMEval: An Open Source Language Model with Better Finegrained Human Alignment

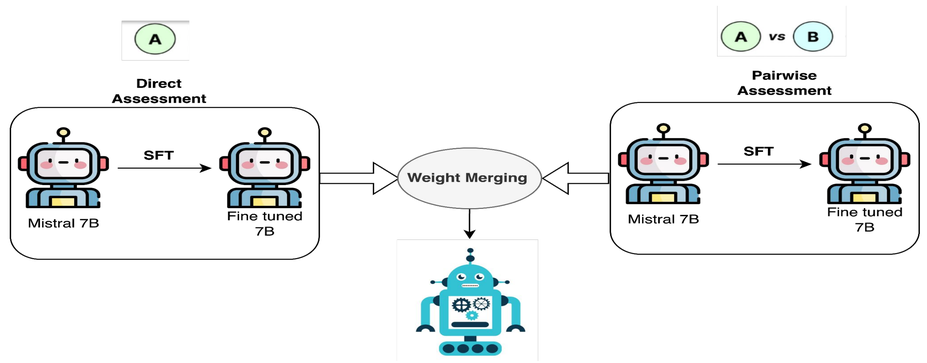
RFEST 2024

Problem Definition and Motivation

- Human evaluation is costly, time consuming and is hard to reproduce.
- Existing works (GPTEval, GPTScore, etc) use GPT-4 as an evaluators but challenges:-
  - **Data Exposure**
  - **Transparency**
  - **Unpredictable API changes**
  - **controllability, and affordability**
- ***Aim***: To build a strong Open source Language model for fine-grained evaluation of language models

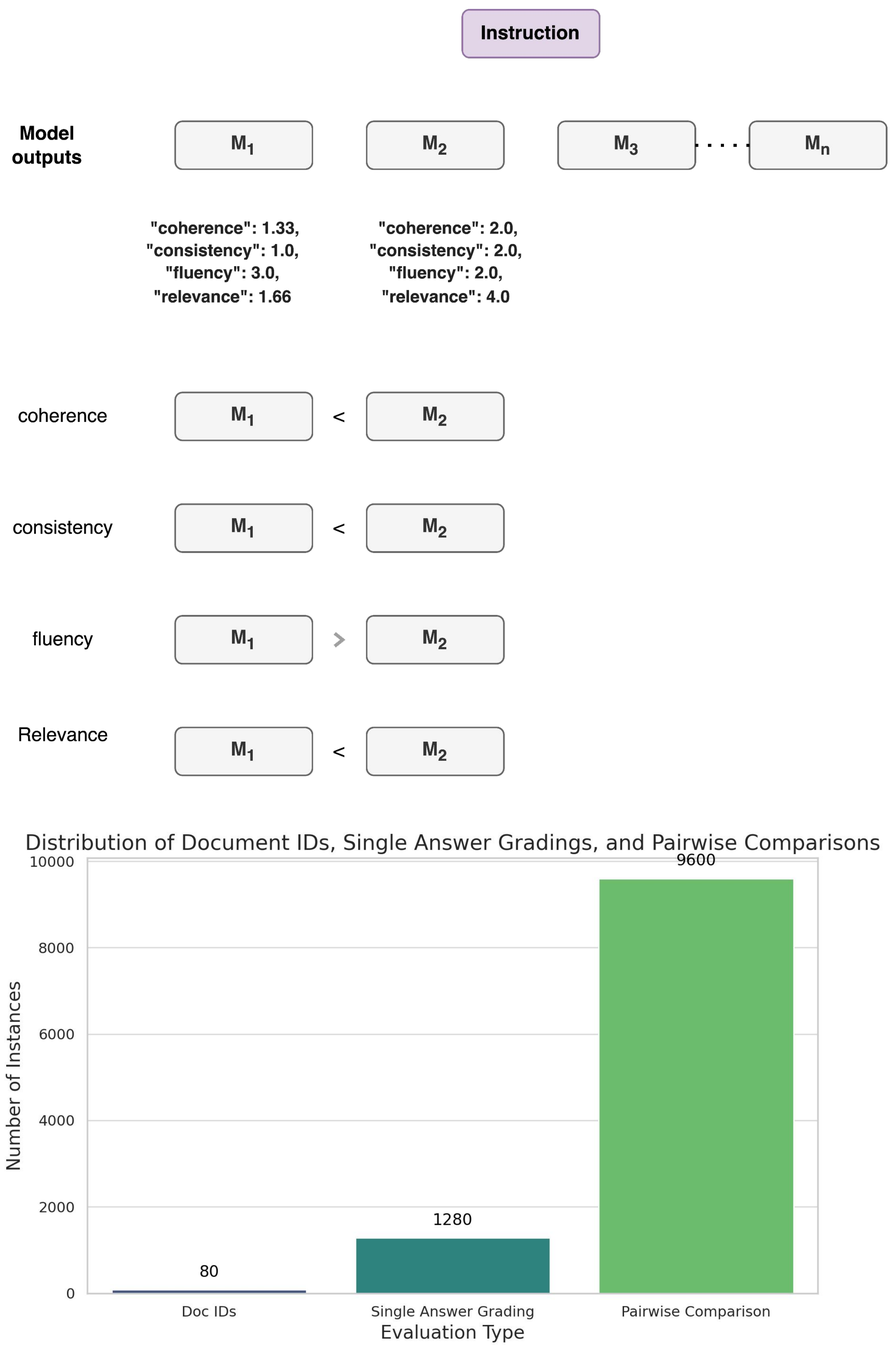


Methodology



Dataset

- SummEval is a benchmark that compares different evaluation methods or summarization.
- It gives human ratings for four aspects of each summary: fluency, coherence, consistency and relevance.
- It is built on the CNN/DailyMail dataset (Hermann et al., 2015)



Experimental Results

Fluency	Pearson	Spearman	Kendall
GPT-4	0.5924	0.5058	0.4554
Llama 70B	0.5812	0.4823	0.4346
Llama3-8b	0.2581	0.2217	0.2052
Llama3-8b (CoT)	0.2474	0.2294	0.2123
Mistral 7b	0.4067	0.3614	0.341
Mistral 7b (IFT)	**0.4239**	**0.3712**	**0.3761**
Mistral 7b (IFT) + WM	**0.4451**	**0.3872**	**0.3892**
Relevance			
GPT-4	0.5882	0.5636	0.4529
Llama 70B	0.509	0.4736	0.4168
Llama3-8b	0.3638	0.3441	0.3003
Mistral 7b	0.3935	0.3647	0.3192
Mistral 7b (IFT)	0.4123	0.3711	0.3201
Mistral 7b (IFT) + WM	**0.4311**	**0.3812**	**0.333**
Consistency			
GPT-4	0.5906	0.5007	0.4199
Llama 70B	0.6806	0.5949	0.5693
Llama3-8b	0.5147	0.4361	0.4167
Llama3-8b (CoT)	0.5061	0.4229	0.4019
Mistral 7b	0.4306	0.4207	0.4046
Mistral 7b (IFT)	0.4531	0.4421	0.4547
Mistral 7b (IFT) + WM	**0.4628**	**0.4512**	**04581**
Coherence			
GPT-4	0.5851	0.5711	0.4626
Llama 70B	0.5777	0.5723	0.4985
Llama3-8b	0.3518	0.3435	0.296
Llama3-8b (CoT)	0.3089	0.3065	0.2647
Mistral 7b	0.367	0.3665	0.3131
Mistral 7b (IFT)	**0.388**	0.3711	0.3201
Mistral 7b IFT + WM	0.376	**0.3861**	**0.3411**

Contact information: Sandeep Kumar | Team: Natural Language Intelligence