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

Evaluation Types

- **Pairwise comparison**: An LLM judge is presented with a question and two answers, and tasked to determine which one is better or declare a tie.
- **Single answer grading**: Alternatively, an LLM judge is asked to directly assign a score to a single answer.

Dataset

- SummEval CNN/DailyMail)[1] 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.

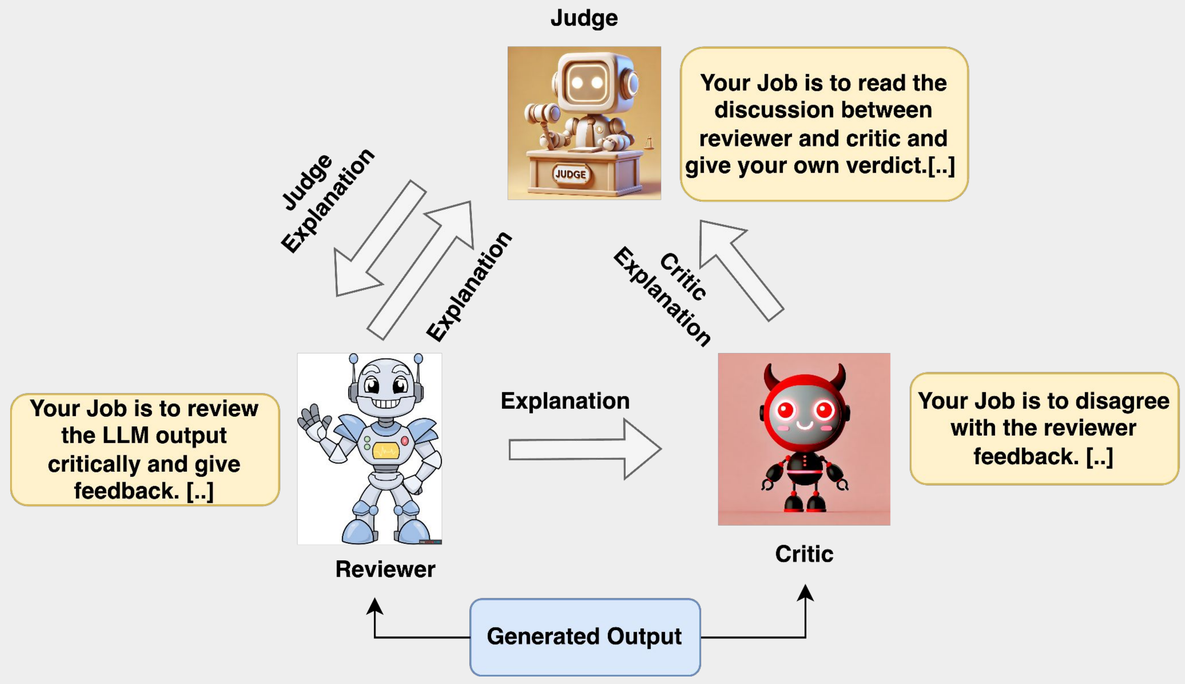
Results and Discussion (Weight Merging)

- In method-2, We LoRA Instruction finetuned Mistral 7b separately with direct assessment and pairwise dataset and merged the weights.
- Final Weight = α*θ_LoRA_pairwise + (1-α)*θ_LoRA_direct
- Pairwise Results:-

	COH	FLU	REL	CON
Llama3-8b	0.5319	0.2723	0.5545	0.2159
Llama3-13b	0.5613	0.3121	0.5745	0.2541
Mistral 7b	0.5733	0.3341	0.5782	0.3284
Mistral 7b (IFT)	0.5913	0.3921	0.5943	0.3891
Mistral 7b (IFT +WM))	**0.6311**	**0.4213**	**0.61328**	**0.4074**

- We found direct assessment and pairwise both helps each other evidenced by increase in their scores.

Proposed Methodology-1 (Debate Agent)



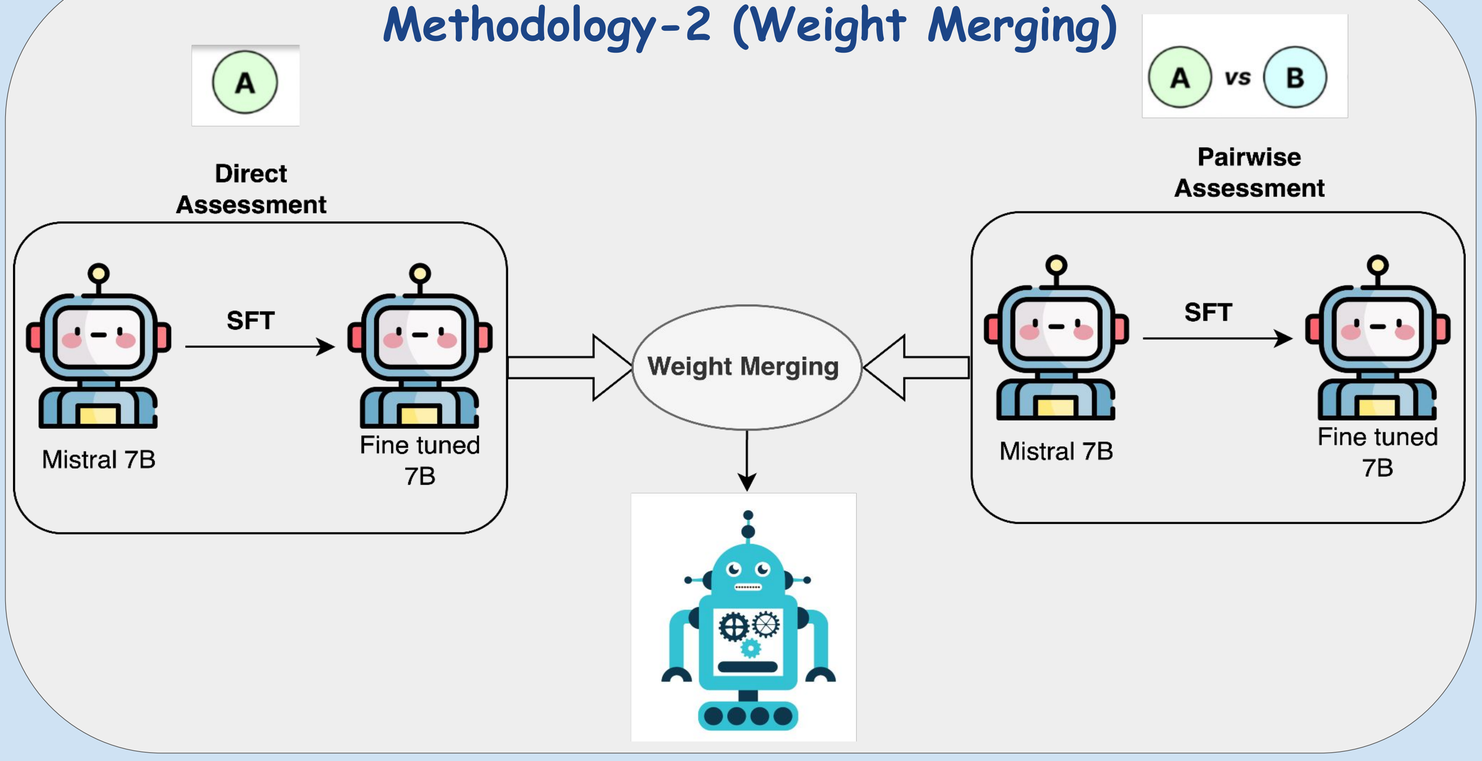
MultiAgent Result

- LLama 70B agentic framework beats GPT-4 direct scoring method by approx 1.5 points on Spearman(ρ), Pearson coefficient (r) (-1 to 1 scale) on each aspects.

	FLU		REL		CON		COH	
	r	ρ	r	ρ	r	ρ	r	ρ
GPT-4	0.5924	0.5058	0.5882	0.5636	0.5906	0.5007	0.5851	0.5711
Mistral 7b	0.4067	0.3614	0.3935	0.3647	0.4306	0.4207	0.367	0.3665
Llama 70B	0.5812	0.4823	0.509	0.4736	0.6806	0.5949	0.5777	0.5723
MultiAgent	**0.6212**	**0.5189**	**0.5982**	**0.5789**	**0.6987**	**0.6051**	**0.5982**	**0.5789**

- Drawbacks:
  - Requires much GPU
  - Slow inference
  - To solve this we introduce method-2

Methodology-2 (Weight Merging)



Direct Assessment Results (Weight Merging)

	FLU		REL		CON		COH	
	r	ρ	r	ρ	r	ρ	r	ρ
Mistral 7b	0.4067	0.3614	0.3935	0.3647	0.4306	0.4207	0.367	0.3665
Mistral 7b (IFT)	0.4239	0.3712	0.4123	0.3711	0.4531	0.4421	**0.388**	0.3711
Mistral 7b (IFT) + WM	**0.4451**	**0.3872**	**0.4311**	**0.3812**	**0.4628**	**0.4512**	0.376	**0.3861**

Contact information: Sandeep Kumar | Team: Natural Language Intelligence