ACL 2023

LLM-BLENDER:

Ensembling LLMs with Pairwise Ranking & Generative Fusion



Dongfu Jiang[♡] **Xiang Ren**[♣♠] **Bill Yuchen Lin**[♠]

dongfu@zju.edu.cn, xiangren@usc.edu, yuchenl@allenai.org

[♠]Allen Institute for Artificial Intelligence

[♣]University of Southern California [♡]Zhejiang University

Allen Institute for AI

USC University of Southern California

No single LLM is the best for all!

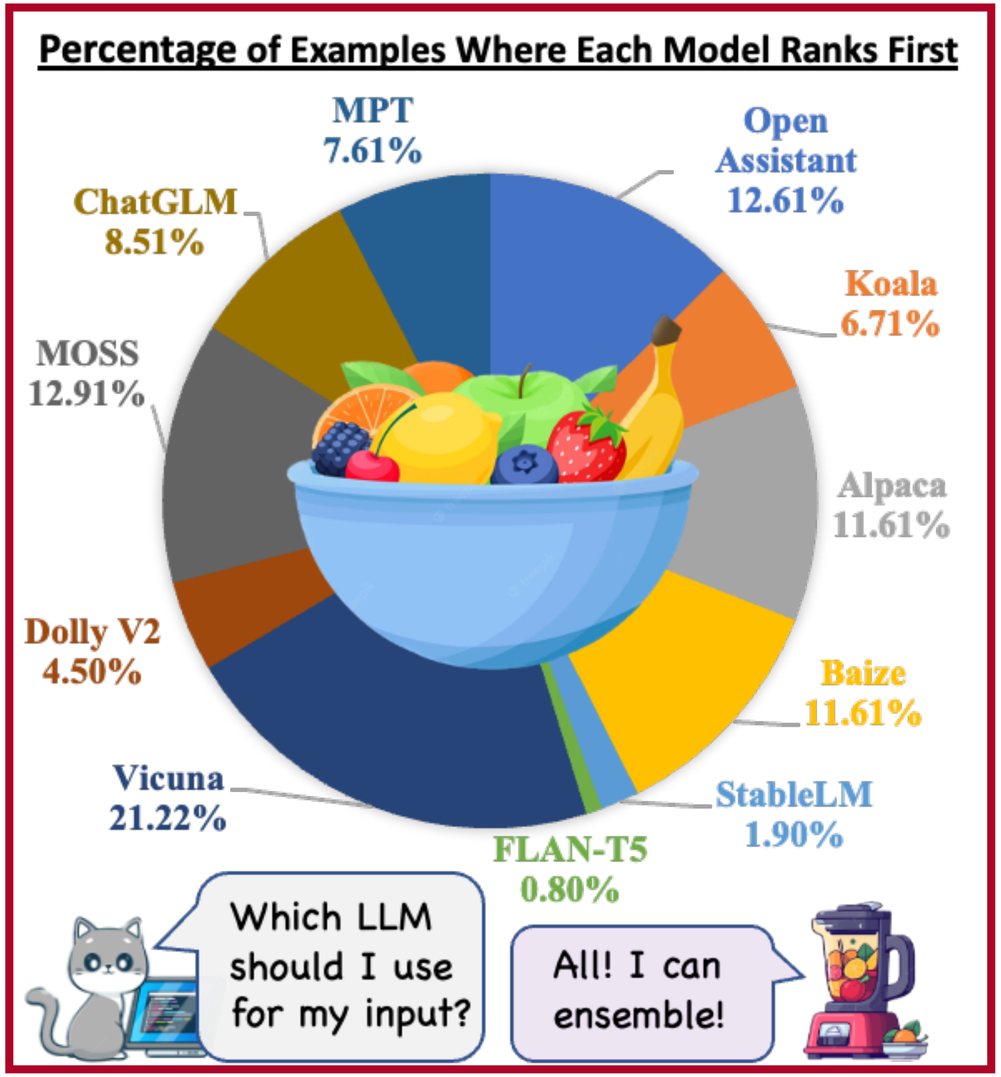
AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Evaluator: GPT-4 Claude Filter: Community Verified Minimal

Model Name	Win Rate
GPT-4	95.28%
Claude	88.39%
ChatGPT	86.09%
WizardLM 13B	75.31%
airoboros 65B	73.91%
airoboros 33B	73.29%
Guanaco 65B	71.80%
Vicuna 13B	70.43%
LLaMA 33B OASST RLHF	66.52%
Minotaur 13B	66.02%
Guanaco 33B	65.96%
Nous Hermes 13B	65.47%
Vicuna 7B	64.41%

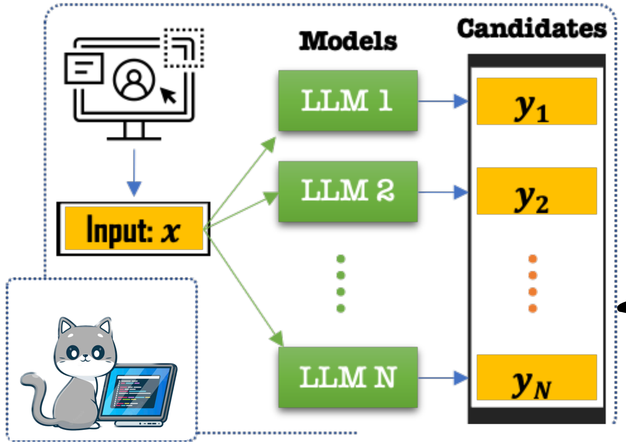
Average Overall Performance



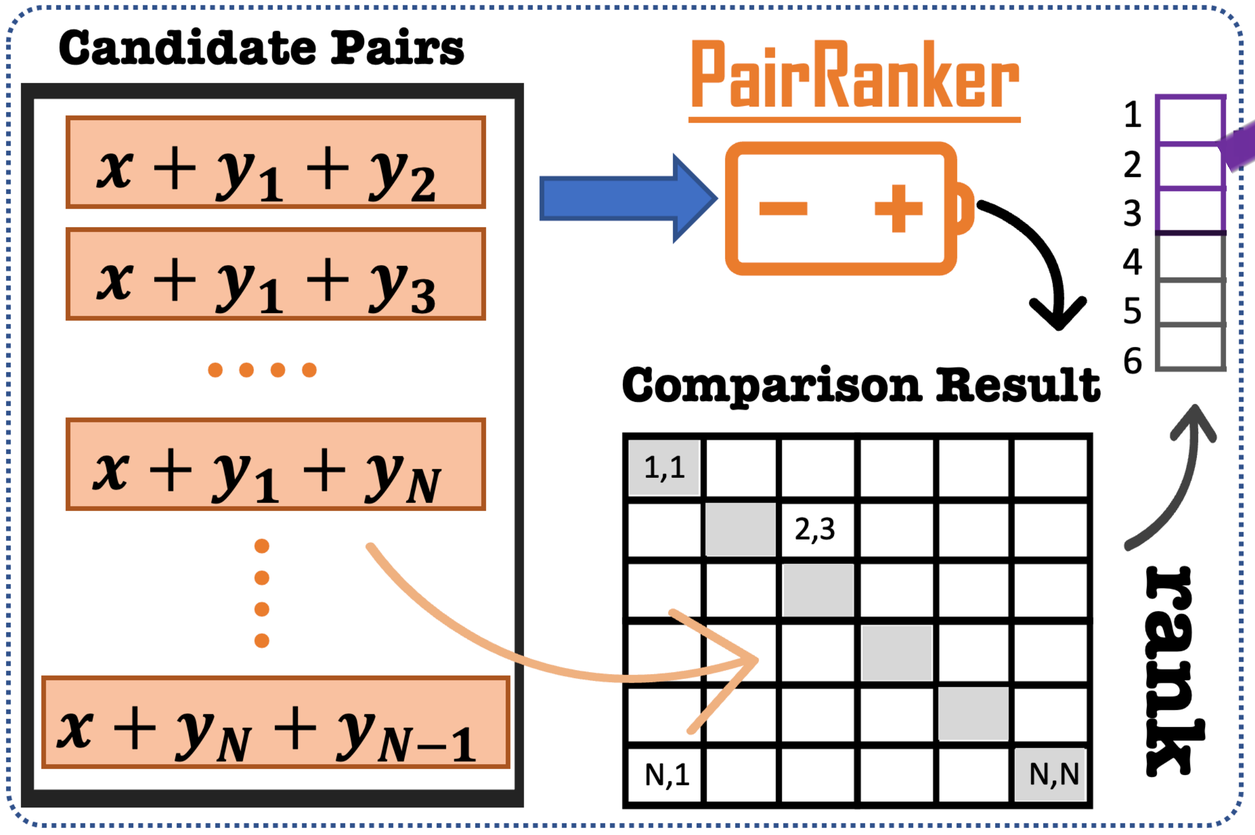
The optimal LLMs for different examples can significantly vary!

LLM-BLENDER

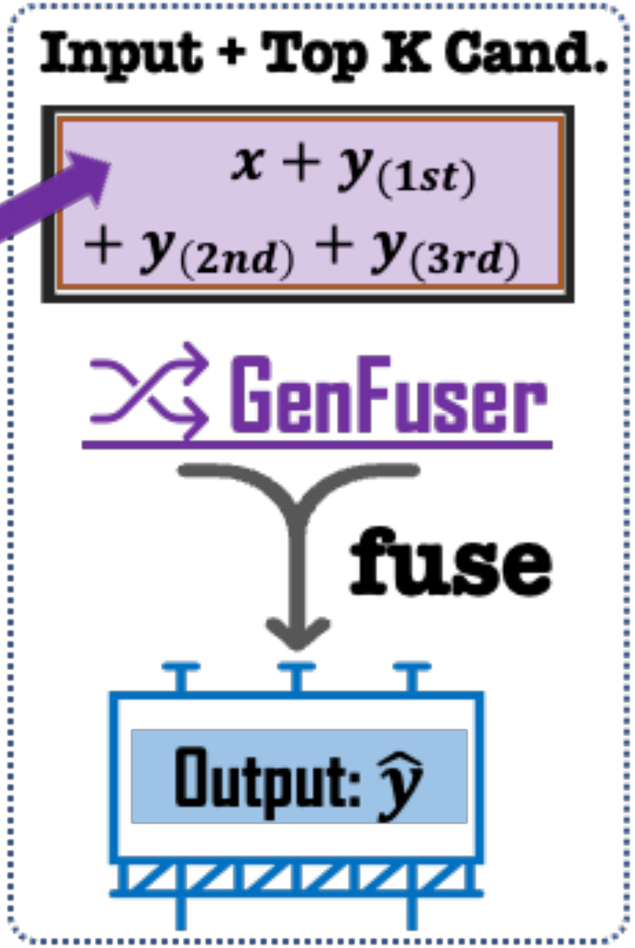
A simple LLM ensemble learning framework for LLMs



1 Run *N* LLMs for the given input in parallel.

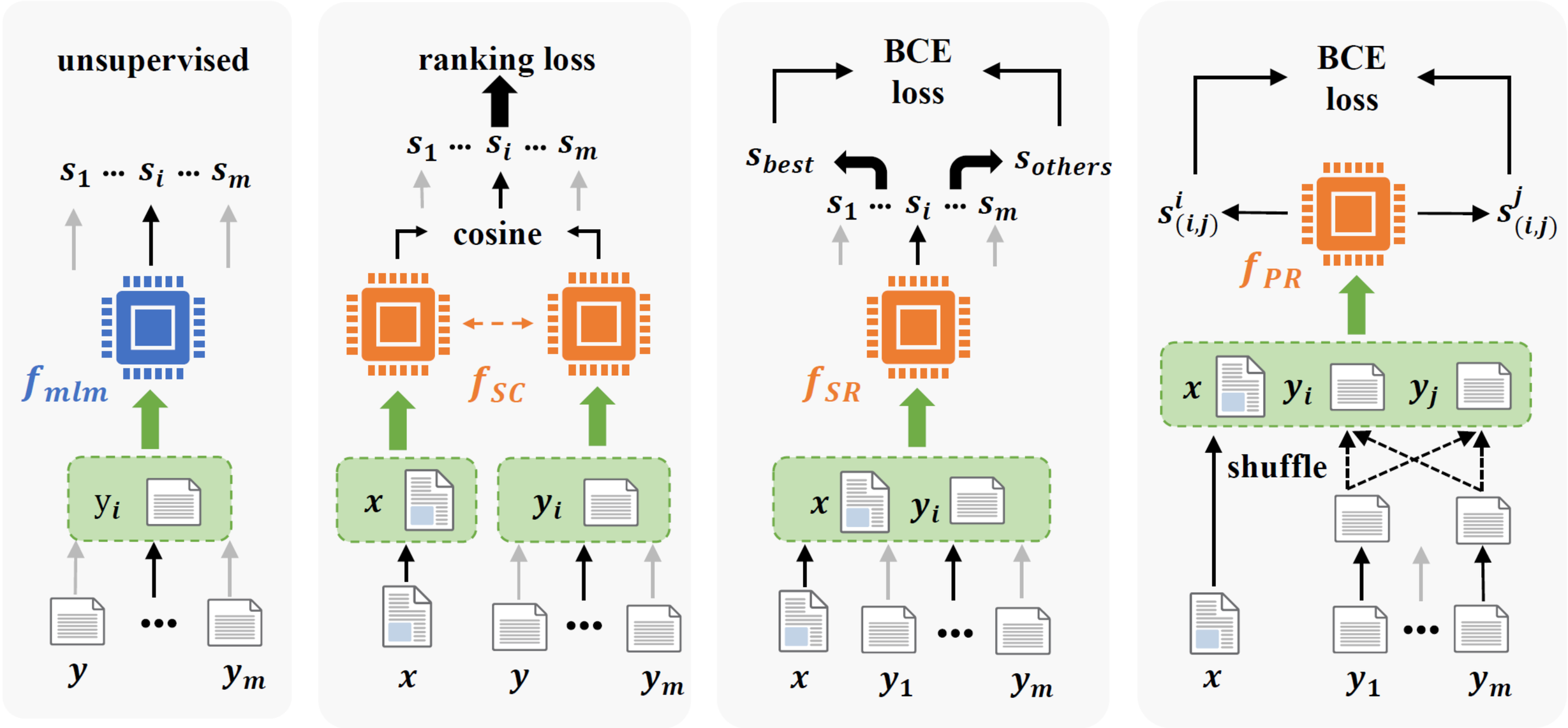


2 Compare each **pair** of candidates, and **rank** them!



3 Generate final output by **fusing** the top 3 candidates.

Ranking Candidates: Baseline & Ours



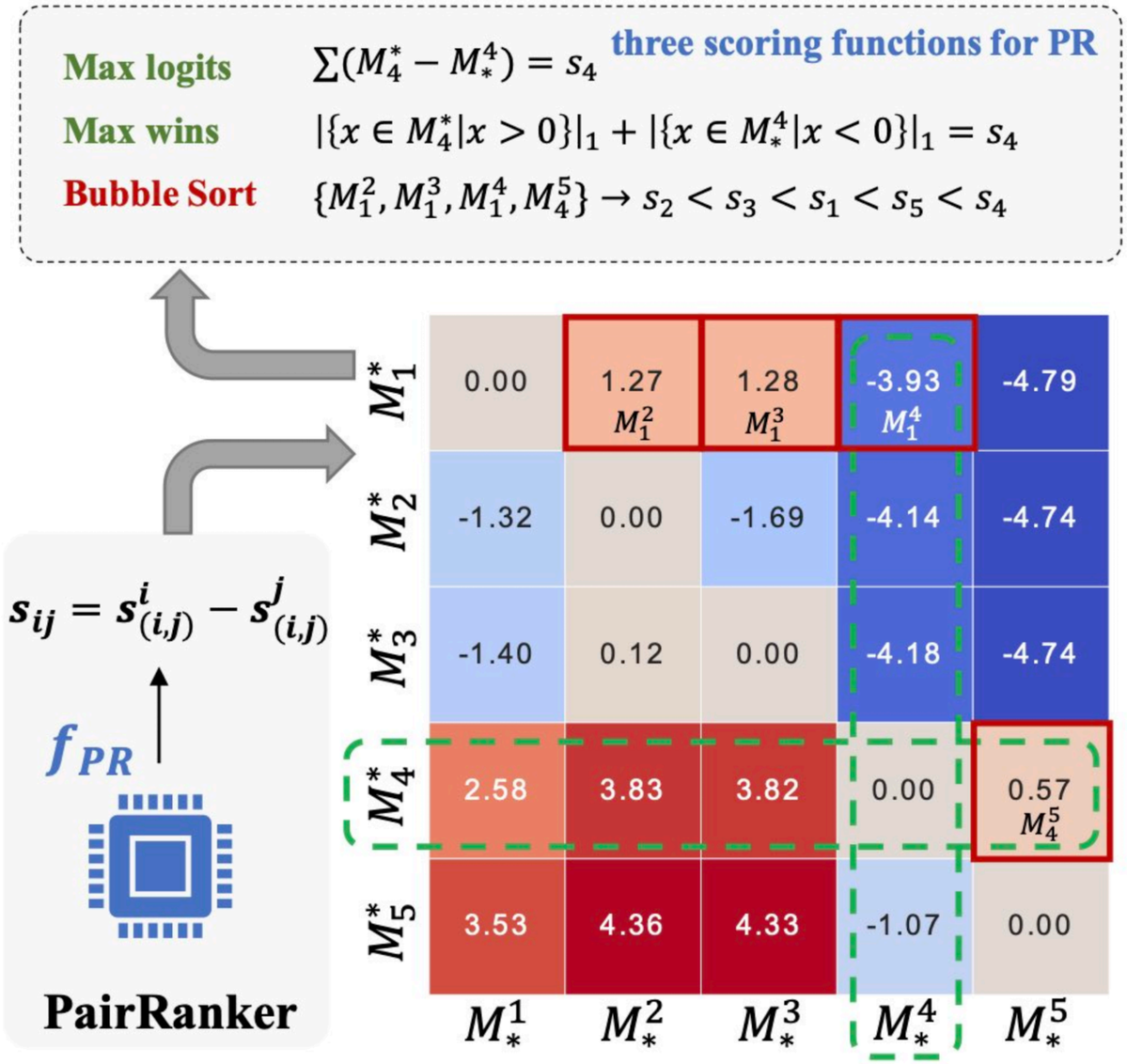
MLM-Scoring

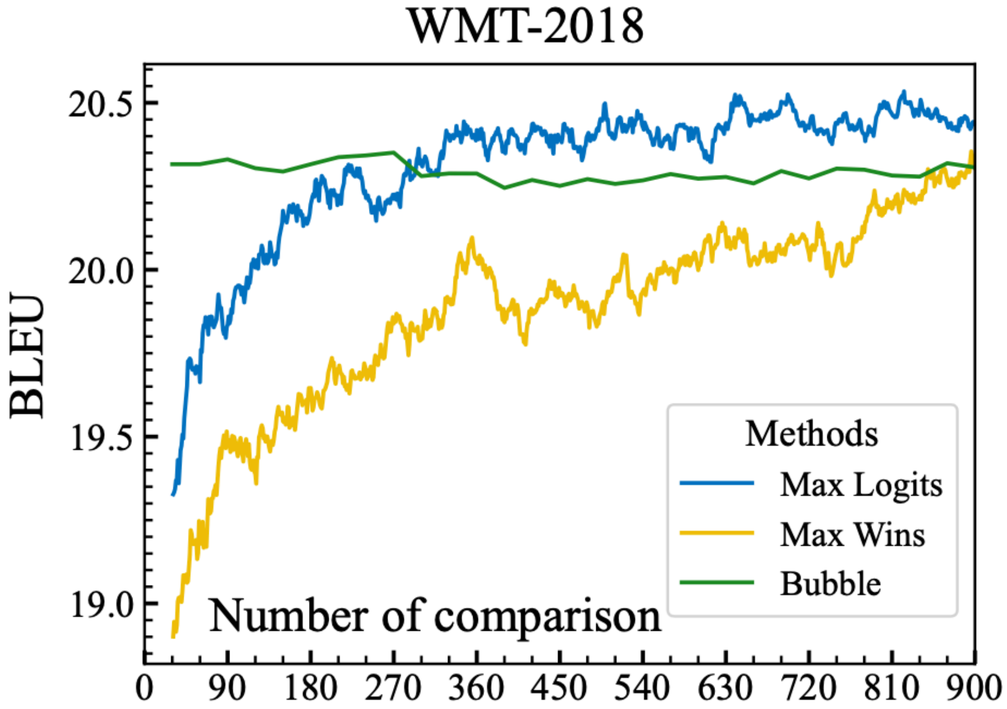
SimCLS

SummaReranker

PairRanker

Ranking by pairwise comparisons





Ranking Methods	Pearson Correlation ↑	Spearman's Correlation ↑	Spearman's Footrule ↓
Random	0.00	0.00	48.27
MLM-Scoring	-0.02	-0.01	47.16
SimCLS	39.89	38.13	29.32
SummaReranker	41.13	39.10	29.69
PairRanker	**46.98**	**44.98**	**27.52**

MixInstruct: A benchmark for LLM Ensambles

110k Instruction-Following Examples

Sources	#Examples	Source	I/O Tokens
Alpaca-GPT4	22,862	GPT-4	22 / 48
Dolly-15K	7,584	Human	24 / 53
GPT4All-LAION	76,552	ChatGPT	18 / 72
ShareGPT	3,002	ChatGPT	36 / 63
Total	110K	Mix	20 / 66

Table 1: **Statistics of `MixInstruct`.** It contains 110K examples and we randomly split the dataset into train/dev/test in 100K/5K/5K sizes.

11 Open-Source LLMs

Open Assistant (LAION-AI, 2023)
Vicuna (Chiang et al., 2023)
Alpaca (Taori et al., 2023)
Baize (Xu et al., 2023)
MOSS (Sun and Qiu, 2023)
ChatGLM (Du et al., 2022)
Koala (Geng et al., 2023)
Dolly V2 (Conover et al., 2023)
Mosaic MPT (MosaicML, 2023)
StableLM (Stability-AI, 2023)
Flan-T5 (Chung et al., 2022)

Three Auto Metrics: BERTScore, BLUERT, BARTScore
GPT-based Pairwise Comparison Results

Evaluation

Category	Methods	BERTScore↑	BARTScore↑	BLEURT↑	GPT-Rank↓	Beat Vic(%)↑	Beat OA(%)↑
LLMs	Open Assistant (LAION-AI, 2023)	**74.68**	**-3.49**	**-0.41**	**3.90**	63.09	N/A
	Vicuna (Chiang et al., 2023)	69.60	-3.57	-0.66	4.13	N/A	64.64
	Alpaca (Taori et al., 2023)	71.46	-3.63	-0.57	4.62	56.80	61.43
	Baize (Xu et al., 2023)	65.57	-3.62	-0.70	4.86	52.84	56.36
	MOSS (Sun and Qiu, 2023)	64.85	-3.74	-0.75	5.09	51.77	51.83
	ChatGLM (Du et al., 2022)	70.38	-3.60	-0.63	5.63	44.14	45.78
	Koala (Geng et al., 2023)	63.96	-3.96	-0.94	6.76	40.01	39.03
	Dolly V2 (Conover et al., 2023)	62.26	-3.90	-0.91	6.90	33.39	31.57
	Mosaic MPT (MosaicML, 2023)	63.21	-3.81	-0.86	7.19	30.89	30.16
	StableLM (Stability-AI, 2023)	62.47	-4.20	-1.02	8.71	21.63	19.97
	Flan-T5 (Chung et al., 2022)	64.92	-4.58	-1.24	8.81	23.92	19.97
Analysis	Oracle (ChatGPT as Ranker)	69.82	-3.37	-0.54	**1.00**	**100**	**100**
Rankers	Random	66.36	-3.84	-0.80	6.14	38.75	39.05
	MLM-Scoring	64.77	-4.11	-0.94	7.00	34.02	30.50
	SimCLS	75.84	-3.37	-0.35	3.82	57.28	33.12
	SummaReranker	75.37	-3.40	-0.37	3.90	**58.94**	25.05
	PAIRRANKER	**76.23**	**-3.31**	**-0.32**	**3.53**	58.21	**37.04**
LLM-BLENDER	PR ($K = 3$) + GF	**79.25**	**-3.09**	**-0.17**	**3.11**	**68.62**	**76.73**

Conclusion

- **LLM-BLENDER** is a simple ensemble learning framework for LLMs.
  - Two sub-modules: PairRanker & GenFuser
  - Largely improve the overall performance of existing LLMs.

- **MixInstruct**: a dataset for evaluating ensemble learning of LLMs
  - 100k/5k/5k examples of instruction-following datapoints

- A unified codebase for evaluation and future development:
  - https://yuchenlin.xyz/LLM-Blender