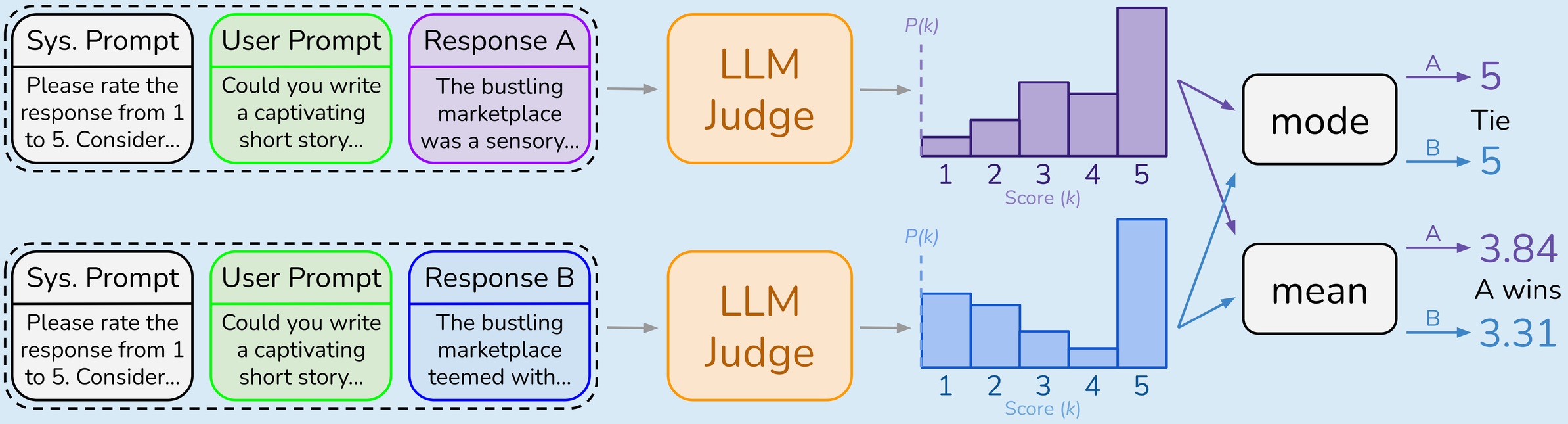
Improving LLM-as-a-Judge Inference with the Judgment Distribution

Victor Wang[1], Michael J.Q. Zhang[2], Eunsol Choi[2]

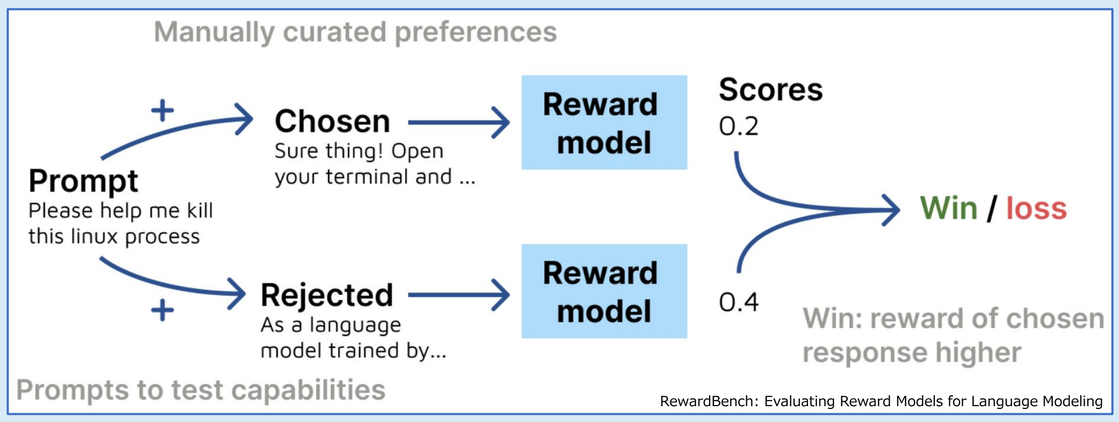
[1] The University of Texas at Austin

[2] NYU

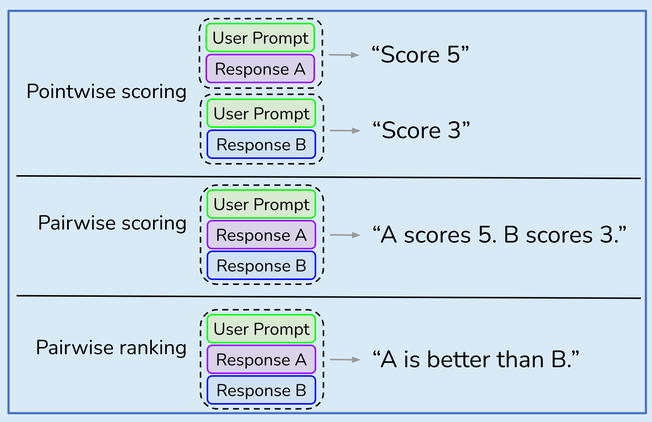
What is the best way to leverage the **LLM judgment distribution** to improve agreement with human preferences?



Evaluating LLM-as-a-Judge



LLM-as-a-Judge Settings



Methods

[mode / mean]
×
[CoT / no-CoT]

Main Results

CoT and no-CoT are comma-separated

Model	Setting	Method	Reward Bench	MT-Bench
GPT-4o	point score	mode	85.1, 84.0	81.9, 80.5
		mean	87.4, **88.0**	**83.6**, 83.2
	pair score	mode	86.7, 87.4	86.2, 86.5
		mean	87.1, **87.6**	86.3, **86.8**
	pair rank	mode	88.4, 89.7	86.3, 85.6
		mean	88.6, **90.5**	**87.3**, 85.9
Llama -3.1-8B	point score	mode	69.6, 72.2	74.9, 71.9
		mean	72.7, **79.3**	78.7, **81.5**
	pair score	mode	71.7, 75.2	**82.6**, 82.4
		mean	72.1, **76.8**	82.3, 81.2
	pair rank	mode	68.9, 58.9	76.2, 63.0
		mean	**74.2**, 68.6	**80.0**, 76.5
Mistral-7B	point score	mode	60.4, 62.7	59.5, 66.2
		mean	63.8, **72.1**	62.6, **74.0**
	pair score	mode	67.3, 68.9	79.3, 79.8
		mean	68.1, **71.0**	80.0, **80.4**
	pair rank	mode	56.3, 53.8	51.5, 51.5
		mean	**63.9**, 59.1	**73.5**, 65.5
Prometheus -2-7B	point score	mode	64.3, 66.0	72.5, 73.5
		mean	64.6, **75.2**	72.1, **81.6**
	pair score	mode	**71.0**, 68.7	78.4, 80.8
		mean	70.5, 70.8	78.3, **80.9**
	pair rank	mode	59.6, 48.2	51.5, 43.0
		mean	**69.7**, 48.8	**75.4**, 33.4

The mean outperforms the mode in 42 out of 48 cases.

No-CoT outperforms CoT in 14 out of 16 cases when using the mean for pointwise or pairwise scoring.

We attribute the harmful effect of CoT to **sharpening**: the initial entropy in the judgment is lost as the model commits to one instantiation of a reasoning trace.

Removing CoT benefits the mean more than the mode, revealing the **synergy between eliciting** (using no-CoT) **and utilizing** (using mean) **distributional judgment**.

Sharpening

Model	Setting	RewardBench	MT-Bench
GPT-4o	point score	.039, .103	.041, .116
	pair score	.042, .066	.038, .064
	pair rank	.002, .065	.012, .114
Llama -3.1-8B	point score	.060, .101	.068, .093
	pair score	.054, .106	.047, .092
	pair rank	.215, .318	.186, .331

Table 2: Average standard deviation of judgment distribution, with judgment options rescaled to $[0, 1]$. Comma-separated values in each cell are with and without CoT. No-CoT always has a greater standard deviation.

We verify the phenomenon of sharpening, showing that no-CoT leads to higher standard deviations than CoT.

See our paper for further analysis on novel methods in the pointwise, pairwise, and listwise settings (evaluating one, two, and many responses at a time).