

# Paper Review

25 April 2025


# TRUST OR ESCALATE: LLM JUDGES WITH PROVABLE GUARANTEES FOR HUMAN AGREEMENT

Jaehun Jung<sup>1</sup>   Faeze Brahman<sup>1 2</sup>   Yejin Choi<sup>1 2</sup>

<sup>1</sup>University of Washington

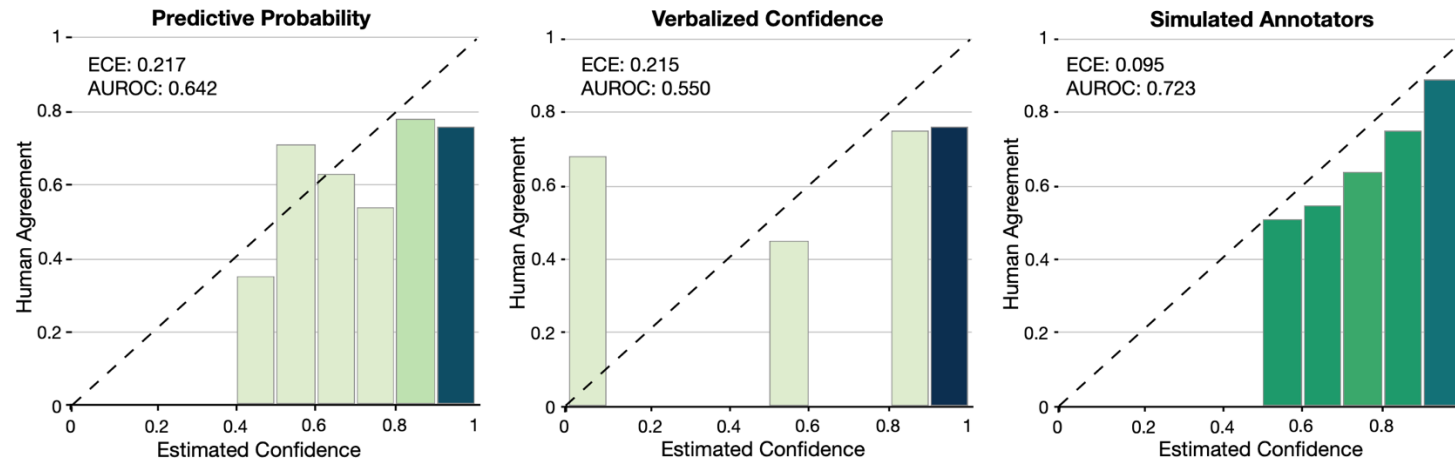
<sup>2</sup>Allen Institute for Artificial Intelligence

## ABSTRACT

We present a principled approach to provide LLM-based evaluation with a rigorous guarantee of human agreement. We first propose that a reliable evaluation method should not uncritically rely on model preferences for pairwise evaluation, but rather assess the confidence of judge models and selectively decide when to trust its judgement. We then show that under this *selective evaluation* framework, human agreement can be provably guaranteed—such that the model evaluation aligns with that of humans to a user-specified agreement level. As part of our framework, we also introduce *Simulated Annotators*, a novel confidence estimation method that significantly improves judge calibration and thus enables high coverage of evaluated instances. Finally, we propose *Cascaded Selective Evaluation*, where we use cheaper models as initial judges and escalate to stronger models only when necessary—again, while still providing a provable guarantee of human agreement. Experimental results show that *Cascaded Selective Evaluation* guarantees strong alignment with humans, far beyond what LLM judges could achieve without selective evaluation. For example, on a subset of Chatbot Arena where GPT-4 almost never achieves 80% human agreement, our method, even while employing substantially cost-effective models such as Mistral-7B, *guarantees* over 80% human agreement with almost 80% test coverage. 

# Motivation

- Prior LLM evaluators either assume perfect calibration or lack statistical control over disagreement risk
- LLM-as-a-judge exhibit systematic biases and miscalibrations
- There's a **cost vs trust** trade-off



# Aims

Problem	Proposed Solution
Lack of tight statistical control over disagreement risk.	Selective evaluation framework (1) with threshold calibration (2)
Poor calibration of confidence.	Simulated annotators (3)
Cost vs trust trade-off	Cascaded selective evaluation (4)

# Background

- We are dealing with pairwise prompt here
- i.e. [Generation A] vs [Generation B], which is better?

# Methodology

## 1. Provide a **Selective Evaluation Framework**

$$(f_{LM}, c_{LM})(x) = \begin{cases} f_{LM}(x) & \text{if } c_{LM}(x) \geq \lambda, \\ \emptyset & \text{otherwise.} \end{cases}$$

Goal: guarantee  $P(f_{LM}(x) = y_{human} | c_{LM}(x) \geq \lambda) \geq 1 - \alpha$

Important terminology:

- $\alpha$  is the ‘risk tolerance’ => maximum fraction of LLM judgement you are willing to see disagree with the human majority
- $\lambda$  is the ‘confidence threshold’ => model will only accept judgements where the  $c_{LM}(x) \geq \lambda$

# Methodology

2. Find the optimal  $\lambda$ , which is  $\hat{\lambda}$ 
  - Use a small human-annotated calibration set
  - Do a grid search from  $\lambda_1 = 0.999$  downwards:
    - Compute the empirical error  $\hat{R}(\lambda)$
    - Find the upper confidence bound of this error  $\hat{R}^+(\lambda)$
    - If value of  $\hat{R}^+(\lambda)$  is below target risk  $\alpha$ , continue searching
    - Else, stop and take the last value of  $\lambda$  as  $\hat{\lambda}$

$$\hat{R}(\lambda) = \frac{1}{n(\lambda)} \sum_{(x, y_{human}) \in D_{cal}} \mathbb{1}\{f_{LM}(x) \neq y_{human} \wedge c_{LM}(x) \geq \lambda\},$$

# Methodology

## 3. Simulated annotators

- For each input  $x$ , repeat the pairwise prompt under  $N$  different few-shot contexts (each with  $K$ -labelled examples).
- Obtain the per-context probability and take average

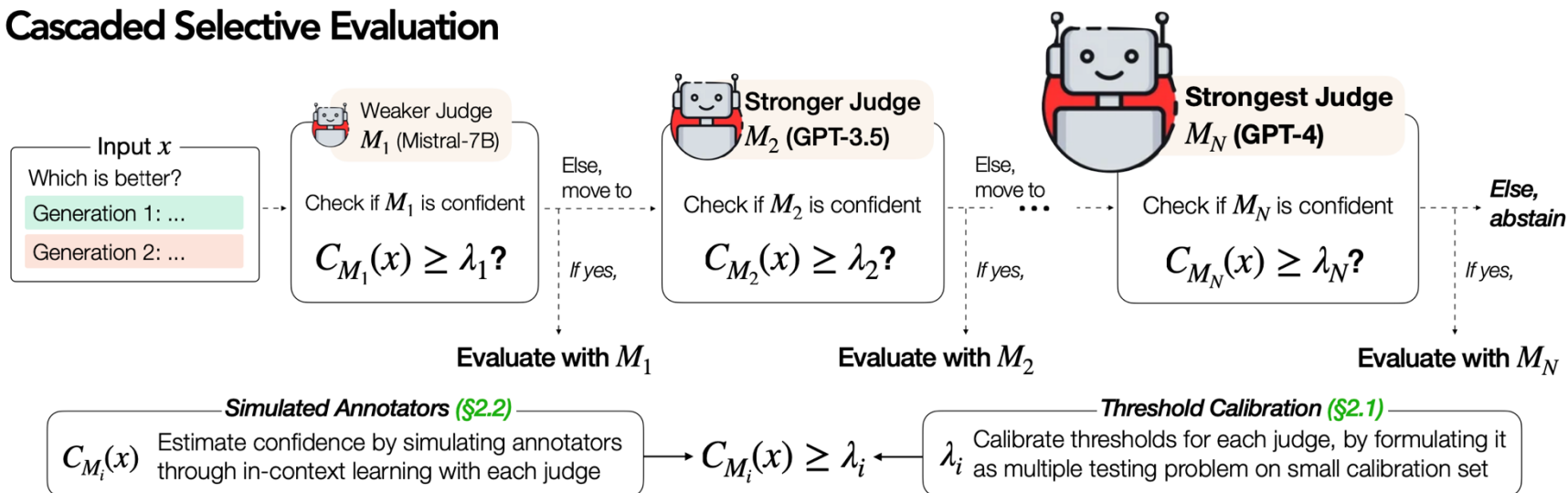
$$c_{LM}(x) = \max_y \frac{1}{N} \sum_{j=1}^N p_{LM}(y|x; (x_{1,j}, y_{1,j}), \dots, (x_{K,j}, y_{K,j})),$$



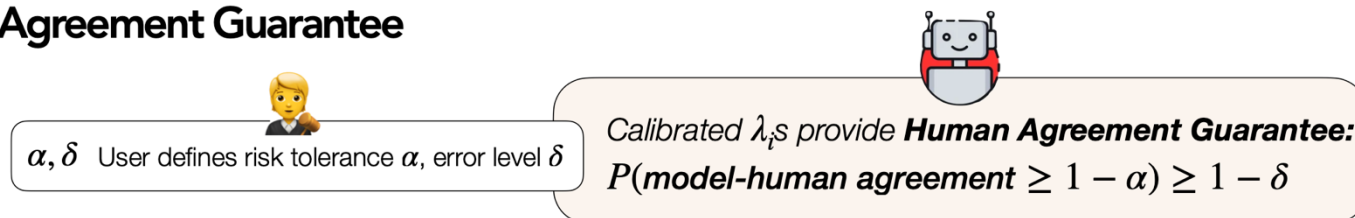
# Methodology

## 4. Cascaded selective evaluation

### Cascaded Selective Evaluation



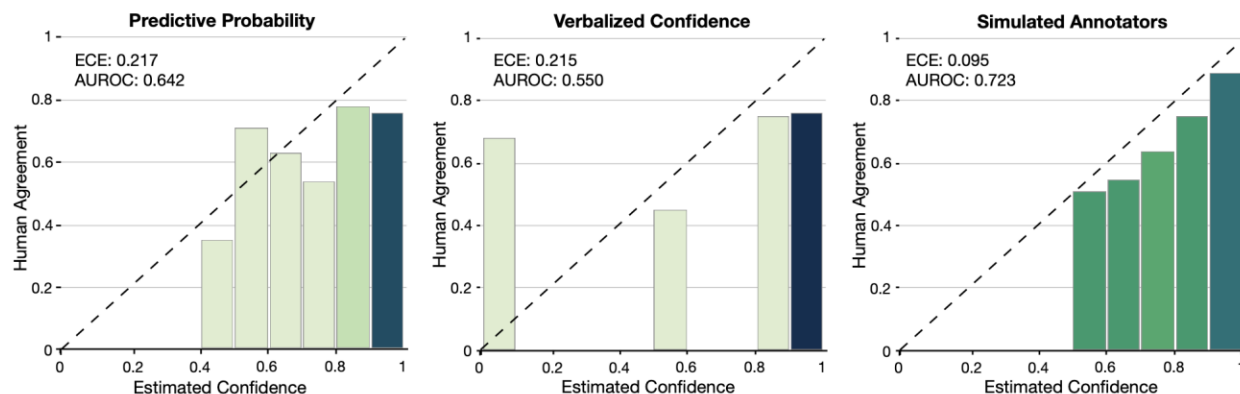
### Human Agreement Guarantee



# Results

Table 1: Performance of confidence measures across judge models. **Simulated Annotators consistently outperforms baselines both in calibration and failure prediction, especially improving the reliability of weaker judge models (*GPT-3.5-turbo* and *Mistral-7B*).**

Dataset		AlpacaEval				TL;DR			
Method		Acc.	ECE ↓	AUROC	AUPRC	Acc.	ECE ↓	AUROC	AUPRC
<i>GPT-4-turbo</i>	Predictive Probability	0.724	0.217	0.642	0.852	0.760	0.196	0.731	0.890
	Verbalized Confidence	0.724	0.215	0.550	0.774	0.760	0.194	0.548	0.792
	Randomized Annotators	0.720	0.113	0.705	0.866	0.779	0.079	0.734	0.905
	Simulated Annotators (Maj.)	0.730	0.106	0.718	0.873	0.783	0.062	<b>0.755</b>	<b>0.921</b>
	Simulated Annotators (Ind.)	<b>0.734</b>	<b>0.095</b>	<b>0.723</b>	<b>0.877</b>	<b>0.788</b>	<b>0.039</b>	<b>0.755</b>	<b>0.921</b>
<i>GPT-3.5-turbo</i>	Predictive Probability	0.644	0.293	0.581	0.691	0.667	0.228	0.653	0.786
	Verbalized Confidence	0.644	0.306	0.505	0.595	0.667	0.211	0.607	0.716
	Simulated Annotators (Ind.)	<b>0.694</b>	<b>0.058</b>	<b>0.632</b>	<b>0.793</b>	<b>0.725</b>	<b>0.043</b>	<b>0.704</b>	<b>0.842</b>
<i>Mistral-7B-it</i>	Predictive Probability	0.618	0.374	0.457	0.579	0.661	0.306	0.613	0.735
	Verbalized Confidence	0.618	0.414	0.490	0.627	0.661	0.335	0.578	0.680
	Simulated Annotators (Ind.)	<b>0.684</b>	<b>0.075</b>	<b>0.632</b>	<b>0.772</b>	<b>0.696</b>	<b>0.103</b>	<b>0.654</b>	<b>0.807</b>

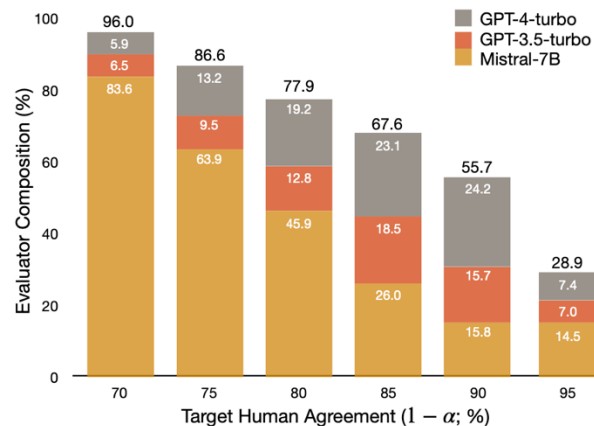
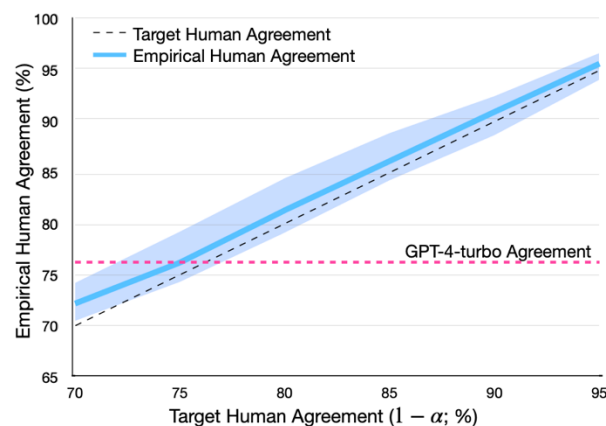


- Simulated Annotators reduces ECE by 50% and improving AUROC by 13% for GPT-4.
- Simulated Annotators improve the reliability of the weaker judge models.
- Ablation studies shows Simulated Annotators gains actually come from simulating diverse human preference.

# Results

- Tested this framework on three datasets:
  1. TL;DR dataset (evaluating summaries)

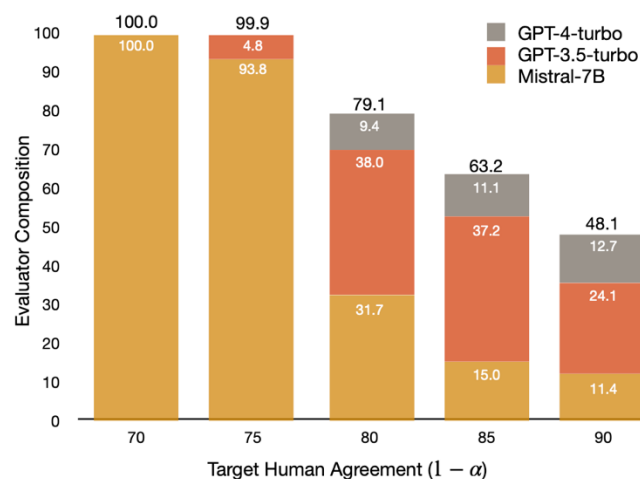
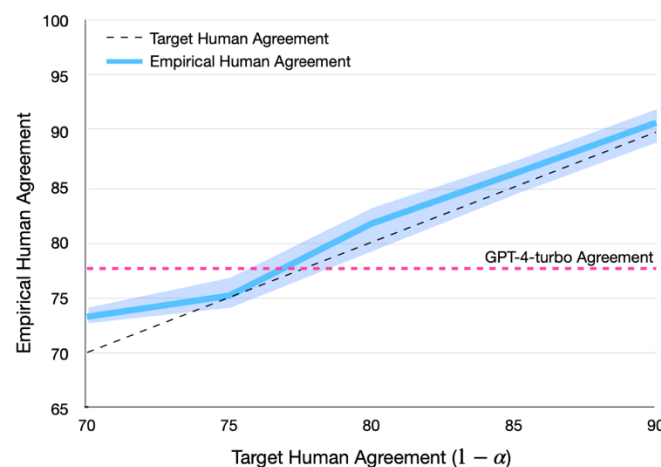
Method	Evaluator Composition (%)			Coverage (%)	Guarantee Success Rate (%)
	Mistral-7B	GPT-3.5-turbo	GPT-4-turbo		
No Selection	0.0	0.0	100.0	100.0	0.0
Heuristic Selection	0.0	0.0	100.0	89.6	42.0
Cascaded Heuristic Selection	59.6	15.0	25.5	64.6	0.0
Point-Estimate Calibration	100.0	0.0	0.0	5.6	54.7
	0.0	100.0	0.0	9.4	79.0
	0.0	0.0	100.0	57.7	47.5
<b>Cascaded Selective Evaluation</b>	<b>28.3</b>	<b>28.2</b>	<b>43.5</b>	<b>55.7</b>	<b>90.8</b>



# Results

## 2. Chat(bot) Arena

Method	Evaluator Composition (%)			Coverage (%)	Guarantee Success Rate (%)
	Mistral-7B	GPT-3.5-turbo	GPT-4-turbo		
No Selection	0.0	0.0	100.0	100.0	0.0
Heuristic Selection	0.0	0.0	100.0	95.2	0.1
Cascaded Heuristic Selection	57.1	15.2	27.7	79.7	0.3
Point-Estimate Calibration	100.0	0.0	0.0	0.0	0.0
	0.0	100.0	0.0	40.5	57.2
	0.0	0.0	100.0	60.9	54.4
<b>Cascaded Selective Evaluation</b>	<b>23.7</b>	<b>58.8</b>	<b>17.5</b>	<b>63.2</b>	<b>91.0</b>



# Results

## Impact of number of annotators, $N$ .

- Good guarantee success rate even at small  $N$ , but improve on coverage.

Table 6: Impact of number of simulated annotators  $N$  on ChatArena, with  $1 - \alpha = 0.85$ . Larger number of simulations generally leads to better coverage, while human agreement is guaranteed even with a small  $N$ . **For all values of  $N$ , Cascaded Selective Evaluation guarantees high agreement with humans while reducing the API cost by 40 % compared to GPT-4 without abstention.**

Method	Empirical Human Agreement (%)	Coverage (%)	Guarantee Success Rate (%)	Relative API Cost
GPT-4 ( $N = 1$ )	77.8	100.0	0.0	1.000
Cascaded Selective Evaluation ( $N = 1$ )	85.2	60.9	90.3	0.655
GPT-4 ( $N = 2$ )	78.2	100.0	0.0	2.000
Cascaded Selective Evaluation ( $N = 2$ )	85.7	61.5	90.8	1.288
GPT-4 ( $N = 3$ )	78.1	100.0	0.0	3.000
Cascaded Selective Evaluation ( $N = 3$ )	85.5	62.1	90.3	1.920
GPT-4 ( $N = 5$ )	78.5	100.0	0.0	5.000
Cascaded Selective Evaluation ( $N = 5$ )	85.8	63.2	91.0	2.849

# Results

Impact of model composition with Chat(bot) Arena.

- Weaker cascade uses Mistral-7B, Mixtral-8x7B and GPT-3.5
- Stronger cascade uses GPT-4 instead of Mixtral-8x7B

Guaranteed human agreement even with the weaker cascade.

Method	Empirical Human Agreement (%)	Coverage (%)	Guarantee Success Rate (%)	Relative API Cost
GPT-4	77.8	100.0	13.9	1.000
Cascaded Selective Evaluation ( <i>stronger</i> )	80.2	77.6	90.5	0.215
Cascaded Selective Evaluation ( <i>weaker</i> )	80.3	68.3	90.8	0.126
Cascaded Selective Evaluation ( <i>weaker</i> + <i>GPT-4</i> )	80.4	78.2	90.6	0.192

# Strengths

- Rigorous human-agreement guarantee, that helps automatic evaluation.
- Cost-effective cascading.
- Model-agnostic.

# Limitations

- Binary, pairwise scope.
- Still requires (some) human-labelled calibration.
- May still be expensive to ensure high coverage.