Enhancing LLM-as-a-Judge via Multi-Agent Collaboration

Yiyue Qian[1], Shinan Zhang[1], Yun Zhou[2], Haibo Ding[2], Diego Socolinsky [1], Yi Zhang [2]

1. Amazon AWS Generative AI Innovation Center, USA
2. Amazon AWS Bedrock, USA

Yiyue Qian: https://yiyueqian.github.io/

Introduction

Background and Motivation:

- The rapid advancement of LLMs has revolutionized AI-generated content evaluation, making the LLM-as-a-Judge paradigm increasingly popular.
- Recent studies have demonstrated the potential of using single LLMs as evaluators. These approaches have shown promising results in automating evaluation processes across various dimensions including coherence, relevance, and fluency.

Major Challenges:

- Single-LLM evaluations lack robustness due to inherent biases from their pre-training data and knowledge.
- While recent works have developed agent-based frameworks to address these limitations, these approaches often lack the flexibility and efficiency needed for diverse evaluation scenarios. These challenges underscore the need for a more robust and adaptable evaluation framework.

Research Goal:

- we aim to propose a novel multi-agent evaluation framework that implements a structured (i.e., three-phrase) collaborative assessment process to assess the generation from LLMs.

CollabEval

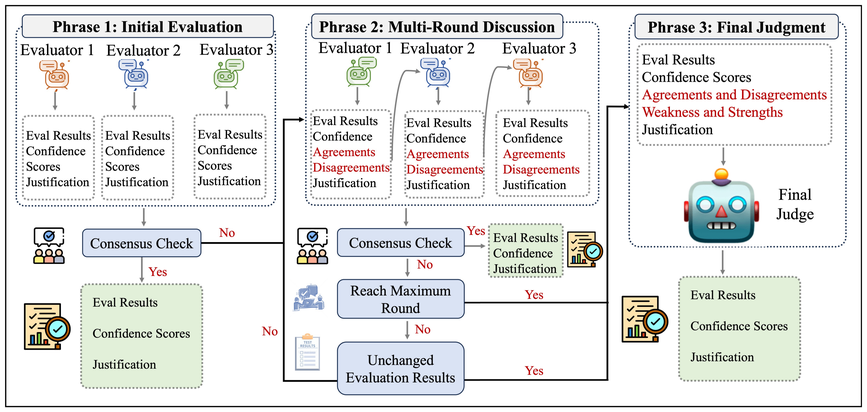
We have proposed a novel multi-agent evaluation framework that implements a three-phase collaborative evaluation process: initial evaluation, multi-round discussion, and final judgment.

- **Initial evaluation**, where different agents independently assess the content;
- **Multi-round collaborative discussion**, where agents share and refine their evaluations through structured dialogue, including confidence scores, agreements, disagreements, and reasoning.
- **Final judgment**, where ultimate evaluation decisions are made based on prior discussions.

Major Contributions

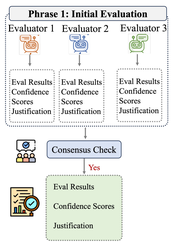
- We introduce a three-stage evaluation framework that uniquely combines independent assessment with collaborative refinement among agents.
- CollabEval supports both criteria-based and pairwise comparisons across multiple dimensions, demonstrating superior performance over single-LLM evaluations via extensive experimental validation.
- Our framework maintains strong performance even when individual LLMs show weaknesses, while ensuring efficiency through strategic consensus checking and early termination.

Framework of CollabEval



Proposed Method

Initial Evaluation

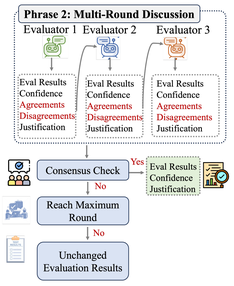


Independent Assessment: CollabEval employs multiple independent evaluators to conduct initial assessments including evaluation results, confidence scores, and detailed justifications for their assessments.

Consensus Check: CollabEval performs a consensus check to determine whether the evaluators have reached agreement in their judgments.

Evaluation Return: If consensus is achieved, the system returns the final evaluation results, demonstrating efficient early termination. However, if evaluators show significant disagreement, the process advances to Phase 2, where evaluators engage in multi-round discussions to resolve differences and refine their assessments.

Multi-Round Discussion

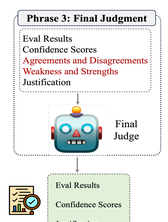


Agents Collaboration: Evaluators share their initial evaluations, confidence scores, and justifications with each other.

Iterative Process: The discussion proceeds iteratively, with evaluators using all available data from both initial evaluations and ongoing discussions to refine their assessments.

Consensus Check: First, the system examines whether all evaluators have reached consensus on their evaluations at the current-round discussion. If consensus is achieved, the system returns the final results. Otherwise, CollabEval then proceeds to verify two additional conditions: whether the maximum number of discussion rounds has been reached, and whether the evaluation results remain unchanged from the previous round.

Final-Judge Evaluation



Final Judge: When the multi-round discussion fails to reach consensus or evaluations remain unchanged, CollabEval employs a strong model as the final judge. The final judge makes the ultimate evaluation decision by analyzing all evaluation results from previous rounds, confidence scores and justifications, areas of agreement and disagreement among evaluators, and the progression of evaluations through discussion rounds.

Experimental Analysis

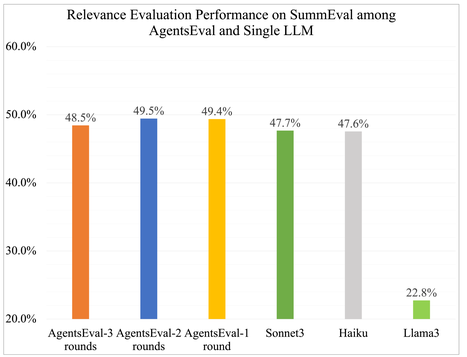
Comparisons with Baseline Models on SummEval Data

Model Setting	Accuracy (%)	Avg Rounds	Gap 1 Ratio (%)	Gap 2 Ratio (%)	Gap 3 Ratio (%)	Gap 4 Ratio (%)	Over-eval Ratio (%)	Under-eval Ratio (%)
Relevance								
CollabEval	**49.5**	2.073	87.8	12.0	0.5	0	31.9	68.1
Single-LLM Sonnet	47.7	1	85.5	13.7	1.6	0	29.7	70.3
Single-LLM Haiku	47.6	1	84.9	14.7	1.1	0	30.2	69.8
Single-LLM Llama3	22.8	1	76.7	23.3	0.0	0	100.0	0.0
Coherence								
CollabEval	**40.4**	2.343	77.8	20.8	1.5	0	63.3	36.7
Single-LLM Sonnet	37.4	1	71.4	23.9	4.9	0	66.4	33.6
Single-LLM Haiku	38.9	1	76.9	22.4	0.8	0	63.4	36.6
Single-LLM Llama3	29.5	1	77.0	22.0	2.2	0	25.4	74.6
Fluency								
CollabEval	**46.9**	2.103	77.8	18.0	4.5	0	21.9	78.1
Single-LLM Sonnet	46.8	1	65.9	24.0	21.4	5	29.7	70.3
Single-LLM Haiku	13.8	1	75.9	22.3	6.2	0	30.2	69.8
Single-LLM Mistral	45.8	1	86.7	13.3	0.0	0	25.0	75.0
Consistency								
CollabEval	**48.2**	2.181	79.6	18.2	7.0	0	10.2	89.8
Single-LLM Sonnet	46.9	1	65.8	25.2	19.8	0	10.4	89.6
Single-LLM Haiku	12.6	1	77.7	20.9	5.3	0	4.7	95.3
Single-LLM Mistral	55.9	1	93.8	5.4	2.8	0	24.4	75.6

Comparisons with Baseline Models on Arena Chatbot and Arena Human Preference Datasets

Model Setting	Accuracy (%)	Average Rounds	GT_Win _Pred_Tie Ratio(%)	GT_Tie_Pred_Win Ratio (%)
Chatbot Arena Data				
CollabEval (Ours)	**60.2**	1.542	50.00	2.63
Round-Table Agents Eval	57.7	1.214	15.84	43.97
Single-LLM Mistral Large	58.2	1	45.54	4.22
Single-LLM Haiku	57.2	1	46.30	3.38
Single-LLM Llama3 70b	59.7	1	53.85	0.00
Arena Human Preference Data				
CollabEval (Ours)	**51.5**	1.517	53.20	9.07
Round-Table Agents Eval	48.7	1.258	12.70	47.37
Single-LLM Sonnet	48.4	1	48.06	13.95
Single-LLM Mistral Large	50.5	1	54.95	5.25
Single-LLM Llama3 70b	48.8	1	55.47	0.39

Discussion Rounds Discussion



Evaluation Patterns Discussion