# CourtEval: A Courtroom-Based Multi-Agent Evaluation Framework

**Sandeep Kumar***
Indian Institute of Technology Patna
sandeep_2121cs29@iitp.ac.in

**Abhijit A Nargund**
Samsung R&D Bangalore, India
abhijit.an@samsung.com

**Vivek Sridhar**
Samsung R&D Bangalore, India
v.sridhar@samsung.com

## Abstract

Automated evaluation is crucial for assessing the quality of natural language text, especially in open-ended generation tasks, given the costly and time-consuming nature of human evaluation. Existing automatic evaluation metrics like ROUGE and BLEU often show low correlation with human judgments. As large language models (LLMs) continue to evolve, researchers have explored their use as alternatives to human evaluators. Although single-agent approaches have shown potential, results indicate that further progress is required to close the gap between their performance and the quality of human assessments. Acknowledging that human evaluations involve multiple annotators, the multi-agent approach allows LLMs to collaborate, enhancing efficiency and effectiveness in handling complex tasks. In this paper, we present CourtEval, a novel Multi-Agent Evaluation Framework modeled after courtroom dynamics. Each agent takes on a distinct role: the Grader, similar to a judge, assigns an initial score; the Critic, like a prosecutor, challenges this score; and the Defender, akin to a defense attorney, defends it. Based on the input from both the Critic and Defender, the Grader re-evaluates the score, leading to a more balanced and fair final decision through this adversarial process. CourtEval substantially outperforms the previous state-of-the-art methods in two meta-evaluation benchmarks in NLG evaluation, SummEval and TopicalChat.

## 1 Introduction

Evaluating the quality of text, whether generated by language models or written by humans, has long posed a significant challenge, consistently attracting considerable attention from researchers and practitioners alike (Celikyilmaz et al., 2020). Conventional approaches mainly depend on human-annotated texts (Callison-Burch, 2009), a method often regarded as excessively time-consuming and expensive. Automatic evaluation metrics for Natural Language Generation (NLG) tasks help minimize the reliance on human evaluations, which are often costly and time-consuming to gather (Zhu and Bhat, 2020). To tackle this issue, evaluation metrics based on n-grams, such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005), have been developed as alternatives to costly human evaluations.

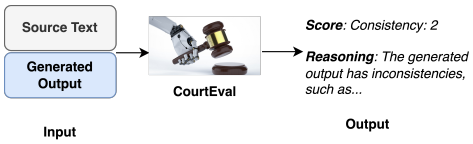


Figure 1: Illustration of the Input and Output

However, these approaches have demonstrated a relatively low correlation with human evaluations, especially in tasks involving open-ended generation or those requiring specialized domain knowledge (Novikova et al., 2017; Kumar et al., 2023a). While embedding-based similarity from pre-trained models can alleviate some of these shortcomings, (Zhang et al., 2020; Belz and Gatt, 2008) highlighted that such similarity, termed "human-likeness", and pointed out that generating human-like text may not always reflect the quality of the output.

In view of the impressive text understanding and instruction-following capabilities of recent LLMs, a body of literature (Zheng et al., 2023; Liu et al., 2023a; Kumar et al., 2024; Chiang and Lee, 2023; Gao et al., 2023; Shen et al., 2023) has adopted LLMs as evaluators to assess the quality of responses to open-ended questions or traditional NLG tasks, including dialogue response generation and summarization. Research findings show that LLMs can mimic human behavior and provide evaluations aligned with human judgments, offering a

*This work was done during an internship at Samsung R&D Institute India, Bangalore.



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 25875–25887

scalable and transparent alternative to expensive, labor-intensive human assessments.

In the human evaluation processes, relying on a single perspective can introduce bias and instability in the results (Karpinska et al., 2021; Kumar et al., 2023b). While a single powerful LLM can already tackle various missions, emerging studies suggest that multiple LLMs can further improve one another through debate and cooperation (Li et al., 2023; Chen et al., 2024). By incorporating multiple LLMs into an integrated group and designing specific interaction mechanisms, different LLMs can engage in proposing and deliberating unique responses and thought processes across several rounds. This approach leads to enhanced factuality of generated responses (Du et al., 2024) and improvement in the completion of arduous tasks (Li et al., 2024). Furthermore, the multi-agent group also addresses and mitigates the Degeneration-of-Thought (DOT) problem (Du et al., 2024).

Our proposed Multi-Agent Evaluation Framework is inspired by the structure of a courtroom, where different roles provide opposing perspectives to arrive at a fair judgment. In this framework, the Grader acts as the judge, initially evaluating the text and assigning a score. The Critic functions like a prosecutor, rigorously questioning and critiquing the Grader's decision. In response, the Defender, similar to a defense attorney, counters the Critic's arguments by defending the Grader's judgment. The Grader then reconsiders the feedback from both the Critic and Defender to form a more balanced final decision. This setup mirrors the adversarial nature of courtroom debates, leveraging multiple perspectives to ensure a robust and well-justified evaluation process.

We summarize our contribution below:-

- We introduce CourtEval, a novel multi-agent framework inspired by the dynamics of a courtroom, where various roles engage in debate from opposing perspectives to ensure a balanced and fair judgment.
- Our proposed CourtEval substantially outperforms the previous state-of-the-art methods in two meta-evaluation benchmarks in NLG evaluation, SummEval and TopicalChat.
- We conduct a comprehensive qualitative analysis demonstrating how the interactions between the Critic and Defender enhance evaluation accuracy, offering unique insights into how adversarial debate refines scoring decisions.

## 2 Related Works

### 2.1 Reference-based text Evaluation

Reference-based text evaluation Previously, model-free scores that evaluate machine-generated text based on a golden candidate reference such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002). BERTScore (Zhang et al., 2020) calculates a quality score based on how similar the reference and candidate texts' BERT (Devlin et al., 2019) embeddings. AEval is a summarization metric based on question-answering, where wh-questions are generated from the reference summary, and the candidate summary is scored according to the percentage of questions it answers correctly (Deutsch et al., 2021). However, collecting reference texts can be costly or even impossible when real-time text quality estimation is required. As a result, there is growing interest in creating automatic evaluation metrics that do not rely on reference texts, often known as reference-free metrics (Deutsch et al., 2022). In our work we performed reference-free text evaluation ie: we didn't included the reference text for evaluation.

### 2.2 LLM based Evaluation

Recently, employing language models as a judge has gained attention as a promising paradigm to mimic the depth and granularity that human evaluation offers (Zheng et al., 2023; Liu et al., 2023b). Many studies propose a simple method for evaluating text quality by providing task-specific instructions. For example, GPTScore (Fu et al., 2024) suggests that better instructions and context are associated with higher probabilities according to GPT-3. G-Eval (Liu et al., 2023a) adopts an automatic chain-of-thought approach.

### 2.3 LLM based Multi-Agent Frameworks

Multi-agent has been investigated for various tasks including autonomous research (Schmidgall et al., 2025), long context (Zhang et al., 2024b), Peer review (Xu et al., 2023), competitive debate with agents performing specific roles such as searcher and analyzer Agent4Debate (Zhang et al., 2024a), mathematical reasoning and logical problem-solving, translation (Liu et al., 2024b). ChatEval is a multiagent framework (Chan et al., 2023), it treats LLM agent independently generates

scores through simultaneous discussions, and the final score is an average of these scores. In contrast another multiagent for NLG task DEBATE (Kim et al., 2024) employs a single scoring agent (Scorer) with a Devil's Advocate providing critical feedback. Another Multi-Agent Debate framework (MAD) (Liang et al., 2024), in which multiple agents express their arguments in the state of "tit for tat" and a judge manages the debate process to obtain a final solution. MAD encourages divergent thinking but lacks the role diversity needed to comprehensively address biases or errors in this task. However our proposed CourtEval utilizes specialized roles (Grader, Critic, Defender, and Controller) within a modular and systematic workflow. While MAD relies on two agents (affirmative and negative debaters) engaging in competitive interactions overseen by a judge, this binary setup limits the complexity and depth of analysis, as it lacks mechanisms for balancing harsh criticism or refining scores iteratively. The Multiagent Debate framework (Du et al., 2023) proposed a framework for improving mathematical and strategic reasoning using complementary approach to improve language responses where multiple language model instances propose and debate their individual responses and reasoning processes over multiple rounds to arrive at a common final answer. However it faces significant challenges due to its reliance on different LLMs as agents, making it difficult to apply uniformly across tasks. CHATEVAL (Chan et al., 2023) is one of the multi-agent evaluation frameworks based on group discussion by Simultaneous-Talk.

Our approach differs from the existing methods, which often rely on either single-agent prompting or multi-agent frameworks that lack deep interaction. Instead, we draw inspiration from courtroom dynamics, using a diverse multi-agent structure where a judge evaluates contrasting perspectives from the agents. This debate-driven approach allows the judge to generate more well-reasoned and balanced scores, grounded in thorough discussion.

## 3 Methodology

In this section, we elaborate on the principal components in CourtEval debater agents, diverse role specification, communication strategy, and provide a detailed overview of each component's role and functionality. Debater agents are one of the most significant components in our framework. Our proposed debater agents Controller, Grader, Critic and Defender is designed to act as an AI assistant, using LLMs but not requiring human input or code execution. We treat each individual LLM as an agent and ask them to generate their response from the given prompt.

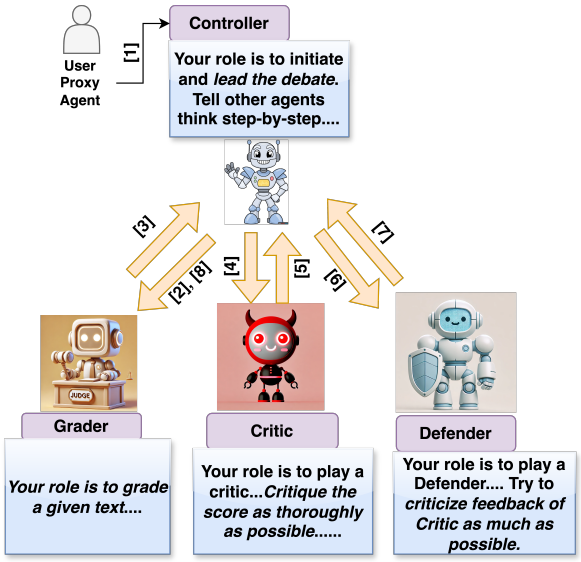


Figure 2: Overview of our proposed CourtEval Multi-Agent framework. The numbers by the arrows indicate the steps outlined in Section 3.6

CourtEval is a framework that simulates a mock conversation between an LLM agents and a user proxy agent. Here a user proxy agent is an agent playing the user's role in conversations with the LLM agent. The workflow of this framework is presented in Figure 2. The user proxy agent takes a LLM evaluation problem to be solved as input and would initiate a conversation with the Controller agent. The initial message from the user proxy agent consists of an initial prompt and the problem to be solved. Another distinct benefit of this framework is that it enables multi-turn dialogues, which can be particularly useful in addressing complex evaluation that require multi-step reasoning.

### 3.1 Prompt

We utilized the prompts presented in (Liu et al., 2023a) as the foundation for our task descriptions and aspect definitions. Given that CourtEval is structured to promote logical reasoning among agents through uniform debate standards, we applied a zero-shot chain-of-thought approach (Wei et al., 2022).

### 3.2 Controller

The job of controller is to communicate with the Grader, Critic and Defender agents.

The define role of the Controller below:

> Your role is to initiate and lead the debate. Tell other agents think step-by-step.

Controller leads the debate. It sends prompt instruction to Grader, Critic and Defender along with the prompt instruction received from user proxy agent. Controller receives responses from the others agents and sends instructions to other agents. It should be noted that the whole process does not require human intervention. Grader, Critic and Defender doesn't interact directly with each other but through the controller.

### 3.3 Grader

The role of grader is receive evaluation steps from controller and respond with a grade score. The define role of the Grader below:

> You are required to assign a score to a provided text. Carefully review both the instructions and the text, then assess the text in a logical manner.

### 3.4 Critic

Critic has become an important mechanism for enhancing the reasoning performance of LLMs. (Zheng et al., 2024; Kalyanpur et al., 2024). Inspired by (Kim et al., 2024) we used critic as an agent in your framework. We defined role of the critic below:-

> Your role is to play/act as a critic. Use step-by-step reasoning to evaluate the given score. Critically assess whether the score is appropriate based on the content. If you believe the score is not justified, provide a detailed critique. Challenge/Critique the score as thoroughly as possible.

### 3.5 Defender

The Defender's role in this multi-agent model is crucial for ensuring a balanced and fair evaluation process. By carefully assessing the critic's feedback, the Defender aims to identify potential biases, overstatements, or inaccuracies in the criticism itself. The Defender serves as a counterbalance to the critic, ensuring that the feedback isn't disproportionately harsh or misguided. This approach not only helps in refining the quality of feedback but also ensures that the score and overall assessment are a true reflection of the text's merit, based on a thorough and adversarial process of scrutiny and defense. The Defender thus plays a vital role in preventing undue penalization and fostering a more accurate and justified evaluation. The define role of the Defender below:

> Your role is to play a Defender. Your logic has to be step-by-step. Read the paper and its summary and Critically review the feedback by the critic and assess whether the feedback is accurate. Try to criticize feedback as much as possible.

### 3.6 Implementation Steps

This section details the operational principles of CourtEval, as outlined in Algorithm 1 and Figure 2, through eight distinct steps. CourtEval features four interactive LLM agents—Controller, Grader, Critic, and Defender each assigned a specific role. These agents communicate and exchange information with one another to facilitate the evaluation process. The eight steps implemented in the CourtEval process are as follows:

1. The process begins when a user agent submits a natural language generation (NLG) evaluation task, specifying the aspects to be assessed. These inputs are sent as prompts to the Controller.

2. The Controller initiates a debate by forwarding the task and the evaluation aspects to the Grader, requesting an initial score and accompanying rationale.

3. The Grader evaluates the task according to the given instructions and returns a score along with a justification to the Controller.

4. The Controller passes this feedback to the Critic for further analysis and evaluation.

5. The Critic examines the Grader's response and attempts to challenge or dispute the score as rigorously as possible.

6. The Defender reviews the Critic's arguments and responds by defending the original score, offering counterarguments to the critique.

7. The Controller considers both the Critic's and Defender's perspectives and instructs the Scorer—acting as a judge—to reassess the score, updating it if necessary based on the strength of the arguments.

**Algorithm 1:** Multi-Agent Scoring Framework with Defender

**Input :** NLG task $T$, aspects $A$, and number of iterations $n$
**Output :** Final score for the task

Define agents: Commander ($C$), Scorer ($S$), Critic ($Cr$), Defender ($D$);
$P \leftarrow C(T, A)$ `// Generate prompts based on task and aspects` $Score \leftarrow S(P)$ `// Compute the initial score`
**for** $i = 1$ *to* $n$ **do**
    $C$ sends $(P, Score)$ to $Cr$; `// Forward prompt and score to Critic`
    $C$ sends $(P, Score)$ to $D$; `// Forward prompt and score to Defender`
    $FeedbackCr \leftarrow Cr(Score)$; `// Receive feedback from Critic`
    $FeedbackD \leftarrow D(Score)$; `// Receive feedback from Defender`
    **if** *'NO ISSUE' in* $FeedbackCr$ *and* $FeedbackD$ **then**
        **break** `// Terminate if both agents find no issues`
    **else**
        $C$ sends $FeedbackCr$ and $FeedbackD$ to $S$; `// Send feedback to Scorer`
        $Score \leftarrow S(FeedbackCr, FeedbackD)$; `// Update score based on feedback from both agents`

**Result :** Final score for the task

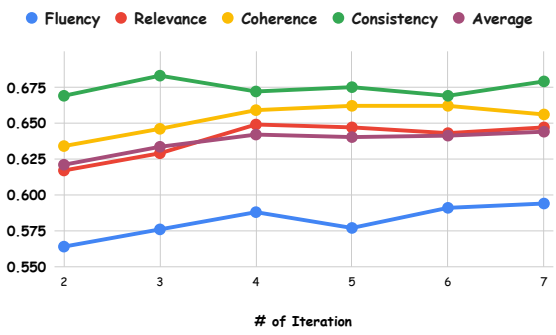


Figure 3: Impact of number of iteration on CourtEval; 'n' refers to the nubmer of debate iterations among multi-agents

8. The revised score is then re-evaluated by the Critic. If new feedback is introduced, the process loops back to step 3 and continues until either the Critic or the Defender issues a `STOP` signal to end the cycle.

## 4 Experiments

### 4.1 Dataset

We used the *SummEval* (Fabbri et al., 2020) and Topical-Chat (Mehri and Eskenazi, 2020) dataset for our experiments. SummEval is a benchmark developed by the Yale LILY Lab and Salesforce Research for assessing summarization models on the English CNN/DailyMail dataset (Hermann et al., 2015) which is a widely used benchmark for evaluating summarization models. It includes 1,600 samples, comprised of 100 unique source texts, each paired with 16 different summary versions. Each summary in the dataset is annotated with human ratings on a scale from 1 to 5, providing a fine-grained assessment of its quality. It provides human ratings on summarization tasks for four key metrics: *fluency*, *coherence*, *consistency*, and *relevance*. We calculate the Spearman and Kendall tau correlation scores for each source text, and the results are averaged at the summary level.

Topical-Chat is a knowledge-grounded human to-human conversation dataset to evaluate four dimensions: *naturalness*, *coherence*, *engagingness*, and *groundedness*. It comprises a total of 360 samples, which are 60 source texts, each with 6 facts and responses. We compute Pearson and Spearman correlation scores for each source text and then take the average at the text level.

### 4.2 Implementation Details

We choose to utilize models from OpenAI's GPT family as our LLMs in CourtEval, including GPT-4o and ChatGPT (GPT-3.5-turbo) and Gemini Pro and Llama-3.1-70B Instruct and set the temperature

| | | SummEval | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Average | | Coherence | | Consistency | | Fluency | | Relevance | |
| | | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| **Others** | ROUGE-L† | 0.165 | 0.128 | 0.128 | 0.099 | 0.115 | 0.092 | 0.105 | 0.084 | 0.311 | 0.237 |
| | BERTScore† | 0.225 | 0.175 | 0.284 | 0.211 | 0.110 | 0.090 | 0.193 | 0.158 | 0.312 | 0.243 |
| | UniEval† | 0.474 | 0.377 | 0.575 | 0.442 | 0.446 | 0.371 | 0.449 | 0.426 | 0.325 | 0.325 |
| | MOVERSCore† | 0.191 | 0.148 | 0.159 | 0.118 | 0.157 | 0.127 | 0.129 | 0.105 | 0.318 | 0.244 |
| | BARTScore† | 0.385 | 0.305 | 0.448 | 0.342 | 0.382 | 0.315 | 0.356 | 0.356 | 0.273 | 0.273 |
| **Gemini** | G-Eval | 0.290 | 0.250 | 0.354 | 0.278 | 0.399 | 0.370 | 0.160 | 0.147 | 0.249 | 0.208 |
| | DEBATE | 0.424 | 0.358 | 0.472 | 0.363 | 0.518 | 0.506 | 0.314 | 0.382 | 0.389 | 0.259 |
| | CourtEval | 0.446 | 0.382 | 0.496 | 0.373 | 0.528 | 0.511 | 0.340 | 0.376 | 0.420 | 0.269 |
| **Llama 70B** | GPTScore | 0.353 | 0.278 | 0.481 | 0.359 | 0.358 | 0.275 | 0.248 | 0.239 | 0.326 | 0.242 |
| | G-Eval | 0.342 | 0.333 | 0.245 | 0.196 | 0.456 | 0.444 | **0.389** | 0.346 | 0.279 | 0.224 |
| | DEBATE | 0.421 | 0.403 | 0.506 | 0.488 | 0.468 | 0.454 | 0.341 | 0.336 | 0.369 | 0.353 |
| | CourtEval | **0.452** | **0.425** | **0.551** | **0.505** | **0.496** | **0.471** | 0.371 | **0.362** | **0.393** | **0.372** |
| **GPT-3.5** | GPTScore | 0.381 | 0.311 | 0.514 | 0.394 | 0.390 | 0.321 | 0.273 | 0.259 | 0.348 | 0.269 |
| | G-Eval | 0.394 | 0.344 | 0.284 | 0.242 | 0.501 | 0.480 | 0.415 | 0.390 | 0.306 | 0.265 |
| | DEBATE | 0.463 | 0.438 | 0.554 | 0.511 | 0.506 | 0.482 | 0.389 | 0.378 | 0.402 | 0.381 |
| | CourtEval | **0.492** | **0.466** | **0.585** | **0.540** | **0.535** | **0.515** | **0.418** | **0.406** | **0.430** | **0.401** |
| **GPT-4o** | G-Eval | 0.517 | 0.437 | 0.506 | 0.407 | 0.582 | 0.536 | 0.480 | 0.404 | 0.501 | 0.400 |
| | ChatEval | 0.524 | 0.454 | 0.473 | 0.409 | 0.590 | 0.512 | 0.497 | 0.429 | 0.535 | 0.468 |
| | MAD | 0.525 | 0.492 | 0.463 | 0.510 | 0.580 | 0.502 | 0.530 | 0.419 | 0.525 | 0.536 |
| | DEBATE | 0.592 | 0.570 | 0.605 | 0.583 | 0.638 | 0.609 | 0.538 | 0.511 | 0.588 | 0.577 |
| | CourtEval | **0.621** | **0.598** | **0.634** | **0.612** | **0.669** | **0.639** | **0.564** | **0.536** | **0.617** | **0.605** |

| | | Topical-Chat | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Average | | Naturalness | | Coherence | | Engagingness | | Groundedness | |
| | | r | $\rho$ | r | $\rho$ | r | $\rho$ | r | $\rho$ | r | $\rho$ |
| **Others** | ROUGE-L† | 0.243 | 0.244 | 0.176 | 0.146 | 0.193 | 0.203 | 0.295 | 0.300 | 0.310 | 0.327 |
| | BERTScore† | 0.262 | 0.273 | 0.226 | 0.209 | 0.214 | 0.233 | 0.317 | 0.335 | 0.291 | 0.317 |
| | UniEval† | 0.552 | 0.417 | 0.455 | 0.330 | 0.602 | 0.455 | 0.573 | 0.430 | 0.577 | 0.453 |
| | MOVERSCore† | 0.222 | 0.238 | 0.169 | 0.170 | 0.247 | 0.259 | 0.275 | 0.269 | 0.198 | 0.147 |
| | BARTScore† | 0.293 | 0.276 | 0.287 | 0.266 | 0.251 | 0.225 | 0.411 | 0.406 | 0.226 | 0.205 |
| **Gemini** | G-Eval | 0.345 | 0.322 | 0.264 | 0.237 | 0.006 | -0.003 | 0.476 | 0.442 | 0.621 | 0.611 |
| | DEBATE | 0.469 | 0.444 | 0.423 | 0.440 | 0.205 | 0.121 | 0.570 | 0.552 | 0.675 | 0.664 |
| | CourtEval | 0.480 | 0.454 | 0.433 | 0.450 | 0.215 | 0.131 | 0.580 | 0.562 | 0.685 | 0.674 |
| **Llama 70B** | GPTScore | 0.336 | 0.312 | 0.286 | 0.264 | 0.252 | 0.185 | 0.429 | 0.434 | 0.379 | 0.368 |
| | G-Eval | 0.378 | 0.380 | 0.318 | 0.350 | **0.493** | 0.490 | 0.347 | 0.356 | 0.355 | 0.326 |
| | DEBATE | 0.423 | 0.469 | 0.447 | 0.465 | 0.396 | 0.543 | 0.435 | 0.463 | 0.417 | 0.405 |
| | CourtEval | **0.457** | **0.507** | **0.490** | **0.496** | 0.438 | **0.585** | **0.460** | **0.503** | **0.440** | **0.444** |
| **GPT-3.5** | GPTScore | 0.374 | 0.345 | 0.330 | 0.291 | 0.281 | 0.221 | 0.468 | 0.455 | 0.420 | 0.414 |
| | G-Eval | 0.419 | 0.414 | 0.365 | 0.380 | 0.536 | 0.525 | 0.373 | 0.379 | 0.404 | 0.371 |
| | DEBATE | 0.464 | 0.497 | 0.478 | 0.495 | 0.439 | 0.567 | 0.459 | 0.490 | 0.437 | 0.437 |
| | CourtEval | **0.497** | **0.532** | **0.512** | **0.530** | **0.470** | **0.607** | **0.492** | **0.525** | **0.468** | **0.468** |
| **GPT-4o** | G-Eval | 0.617 | 0.622 | 0.621 | 0.618 | 0.627 | 0.614 | 0.543 | 0.584 | 0.675 | 0.679 |
| | ChatEval | 0.667 | 0.652 | 0.680 | 0.648 | 0.662 | 0.636 | 0.682 | 0.681 | 0.698 | 0.686 |
| | MAD | 0.682 | 0.670 | 0.685 | 0.660 | 0.648 | 0.645 | 0.685 | 0.674 | 0.698 | 0.693 |
| | DEBATE | 0.696 | 0.685 | 0.690 | 0.673 | 0.646 | 0.641 | 0.690 | 0.671 | 0.703 | 0.699 |
| | CourtEval | **0.748** | **0.736** | **0.741** | **0.724** | **0.742** | **0.722** | **0.756** | **0.752** | **0.807** | **0.794** |

Table 1: Spearman ($\rho$) and Kendall-Tau ($\tau$) correlations for SummEval, alongside Pearson (r) and Spearman ($\rho$) correlations for Topical-Chat, between machine and human evaluations. †: results from (Liu et al., 2023a); all other results are reproduced by the authors.

to 0 to ensure reproducibility. The rationale behind this selection is the exceptional performance these models offer, being among the most advanced and powerful in the world. When this study was conducted, the cost for processing input tokens with gpt-4o was 2.50 per 1M input tokens while generating output tokens was priced at 10.00 per 1M output tokens.

### 4.3 Baselines

Our focus of this paper is to solve the issues faced by LLM based text evaluation by proposing a specialized multi-agent framework. So, we compared our method with the available baselines on LLM text evaluation. We extensively evaluate the performance of CourtEval with baselines, including a traditional evaluator, ROUGE-L (Lin, 2004), which measures lexical overlap between generated and reference texts. We also compare against pretrained

language model-based evaluators: BERTScore (Zhang et al., 2020), which computes token similarity in embedding space; MoverScore (Zhao et al., 2019), which extends BERTScore with Earth Mover's Distance for improved semantic alignment; BARTScore (Yuan et al., 2021), which leverages a pretrained BART model to assess generation likelihood; and UniEval (Zhong et al., 2022), which evaluates text quality across multiple dimensions using LLM-based scoring. Additionally, we include comparisons with recent LLM-based evaluators, GPTScore (Fu et al., 2024), which utilizes GPT models for reference-based and reference-free evaluation, and G-Eval (Liu et al., 2023a), which incorporates Chain-of-Thought (CoT) reasoning and probability-weighted summation for enhanced evaluation reliability. MAD (Liang et al., 2024) employs multiple agents engaging in a "tit-for-tat" argumentation, with a judge overseeing the debate to reach a final decision. DEBATE (Kim et al., 2024) utilizes a single scoring agent (Scorer) with a Devil's Advocate offering critical feedback for NLG tasks. CHATEVAL (Chan et al., 2023) is a multi-agent evaluation framework using group discussion via Simultaneous-Talk.

CourtEval incorporates a feedback loop where the Grader revises its score based on input from both the Critic and Defender. Through multi-turn interactions, these agents collaboratively refine the evaluation by providing checks and balances that ensure balanced scoring and minimize the risk of over-penalization or unjustified adjustments. This iterative process enables fine-grained score adjustments and facilitates convergence toward more accurate evaluations. In contrast, MAD's process ends with the judge's decision, lacking mechanisms for iterative refinement, which limits its capacity to effectively address complex evaluation tasks.

### 4.4 Main results

We report the performance of our proposed framework CourtEval on the SummEval and Topical-Chat datasets on Table 1. As evident from the table, *CourtEval*, outperforms the traditional methods (ROUGE-L, BERTScore, MOVERScore, BARTScore) on each metrics on both the datasets. Result shows that GPT-4o outperforms GPT-3.5, Gemini-Pro and Llama-70B on every baselines. Our proposed method, CourtEval, outperforms the single-agent frameworks based on GPT-4o, achieving a 20.12% improvement over G-Eval and a 29.13% improvement over GPTScore on GPT-3.5 on the SummEval dataset. Similarly, CourtEval outperforms the single-agent frameworks on GPT-4o, achieving a 20.99% improvement over G-Eval and a 32.83% improvement over GPTScore on GPT-3.5 on the Topical-Chat dataset. With respect to existing multi-agent frameworks, our proposed method, CourtEval, achieves an 18.51% improvement over ChatEval and a 13.06% improvement over MAD and a 4.90% improvement over DEBATE with respect to average Spearman correlation on the SummEval dataset. We found that ChatEval suffers from the agreement problem: once the LLM has provided an incorrect score with convincing reason another agent gets to agree on it. However, in our proposed framework Critic Agent always criticises during the next iteration invoking more reasoning and distinct viewpoints for the judge Agent.

Also, CourtEval incorporates a feedback loop where the Grader revises its score based on input from both the Critic and Defender. Through multi-turn interactions, these agents collaboratively refine the evaluation by providing checks and balances that ensure balanced scoring and minimize the risk of over-penalization or unjustified adjustments. This iterative process enables fine-grained score adjustments and facilitates convergence toward more accurate evaluations. In contrast, MAD's process ends with the judge's decision, lacking mechanisms for iterative refinement, which limits its capacity to effectively address complex evaluation tasks. Additionally, CourtEval shows a 12.14% improvement over ChatEval and a 9.68% improvement over MAD and a 7.47% improvement over DEBATE with respect to average Spearman correlation on the Topical-Chat dataset.

### 4.5 Impact of Iteration Count on CourtEval Performance

To evaluate the impact of increasing the number of iterations on CourtEval's performance, we conducted an experiment utilizing GPT-4 on the SummEval dataset[1] illustrates how varying the number of iterations affects the evaluation metrics.

Our results indicate that all performance metrics improve as the number of iterations increases up to the fourth iteration. This suggests that iterative processing enhances the model's evaluation capa-

[1]Due to budget constraints, we conducted our experiments on a randomly selected 30% subset of the full dataset. This random sampling ensures that the subset is representative of the entire dataset, allowing us to maintain the validity and generalizability of our results despite the reduced sample size.

bilities during the initial stages. Beyond the fourth iteration, however, we observed fluctuations in the metrics, with performance sometimes increasing and sometimes decreasing. Despite these variations, the average values of the metrics remained relatively stable after the fourth iteration. This stability implies that additional iterations beyond the fourth do not significantly contribute to performance improvements and may introduce variability instead.

### 4.6 Qualitative Analysis

We performed qualitative analysis to understand how Critic and Defender's conversation helps Grader take better decision.

#### 4.6.1 How the Defender helps correct the Critic's feedback:

We observed several cases where the Defender played a crucial role in correcting the feedback provided by the Critic. In instances where the Critic's evaluation was overly harsh or imprecise, the Defender intervened with a more reasonable counter-argument. For example in Appendix Section A the initial score given by the Grader is 3, which is higher than the actual human score 2. This indicates that the grader awarded a higher score based on the text's coherence. Next, the Critic challenges the high score by pointing out mistakes in the previous assessment and suggests a score of 1. The Defender, however, argues that the critic's assessment is too harsh and recommends a score of 2, which matches the human score. Finally, the Grader considers the feedback from both the critic and the Defender and, based on their reasoning, finds the defender's argument stronger and more reasonable.

#### 4.6.2 How the Critic helps correct the Defender's feedback:

We observed serveral cases that the Critic plays an essential role when the Defender's feedback is not reasonable. For example in Appendix Section B, the Grader initially assigns a score of 3, which is higher than the actual human score of 2. This suggests that the Grader awarded a higher score based on the text's coherence. The Critic then challenges this score, highlighting mistakes in the previous assessment and recommending a score of 2. The Defender, however, argues that the Critic's assessment is overly harsh and advocates for retaining the score of 3, which aligns with the human score. Ultimately, the Grader reviews the feedback from both the Critic and the Defender and, after considering their reasoning, determines that the Critic's argument is more convincing and reasonable.

### 4.7 Cost and Feasibility of CourtEval Deployment

As CourtEval is deployed for real-world LLM evaluation tasks, its computational feasibility becomes a critical factor. CourtEval requires approximately three times the runtime per iteration compared to single-agent models (e.g., *G-Eval* or *GPT-Score*) due to its multi-agent structure. As shown in Section 4.5, performance improves up to four iterations, leading to a total runtime of $4 \times 3 = 12$ times that of single-agent models. However, the iteration count is adjustable based on computational resources and task requirements. Unlike many evaluation models requiring significant GPU resources, CourtEval operates via GPT-4o API calls, eliminating training and maintenance costs. This makes it accessible to small-scale industries and academic institutions without dedicated hardware.

To assess economic feasibility, we estimated the cost of evaluating a single LLM-generated output using GPT-4o. The average processing cost per instance for a single evaluation metric on SummEval is USD 0.0377, with 100 outputs costing approximately USD 3.77. While this cost accumulates over large-scale evaluations, it remains significantly lower than traditional human evaluation, which is both costly and time-intensive. Moreover, CourtEval's multi-agent setup ensures scalability despite multiple interacting agents, making it an efficient alternative to manual annotation. Additionally, CourtEval can serve as a pre-deployment evaluation tool for newly developed LLMs, reducing costs associated with extensive fine-tuning before release. Ongoing open-source LLM development efforts (Touvron et al., 2023; Liu et al., 2024a) may further reduce these costs in the near future. Further cost details and scalability considerations are provided in the Appendix C.

### 4.8 Error Analysis

In this section, we conducted a detailed human analysis to identify recurring error patterns in model predictions that deviated from expected outcomes.CourtEval's fluency evaluation shows three recurring errors: (1) Over-penalizing minor grammar issues, (2) favoring formality over context, and (3) harshly judging structural transitions despite

coherence. Detailed examples are provided in Appendix D.

## 5 Conclusion and Future Work

In conclusion, this work presents CourtEval, a novel multi-agent evaluation framework inspired by courtroom dynamics to enhance the quality, fairness, and reliability of automated text evaluations. The introduction of distinct roles—Grader, Critic, and Defender enables structured debate that mitigates bias and facilitates more accurate alignment with human judgments. Our experimental results show that CourtEval significantly outperforms traditional evaluation methods and state-of-the-art frameworks across natural language generation benchmarks, such as SummEval and TopicalChat, demonstrating a notable improvement.

For future work, we plan to expand the multi-agent structure with domain experts, optimize debate iterations with adaptive mechanisms, integrate real-world feedback, enhance multilingual and multimodal capabilities, and apply CourtEval to tasks like machine translation and conversational agents.

## Limitations

Although our proposed meta-evaluation method, CourtEval, outperforms single-agent approaches on benchmark datasets, it's crucial to acknowledge that implementing a multi-agent system inherently leads to higher costs. Thus, the processing costs must always be considered when evaluating the applicability of CourtEval. Our framework, like other LLM-as-a-judge approaches, is influenced by the choice of the underlying LLM judge, leading to some variation in scores. While our multi-agent debate mechanism helps mitigate inconsistencies, complete standardization across different judge models remains an open challenge.

## Ethics Statement

Our research presents a novel evaluation framework designed to assess the quality of generated text, demonstrating strong alignment with human judgments. However, we acknowledge the potential societal risks that could emerge from misuse of this technology. These risks include ethical concerns such as the automated creation of misleading or false information, the publication of machine-generated content that receives high ratings from our system, and the potential exploitation of the tool for deceptive purposes. This emphasizes the necessity for responsible usage and governance, underscoring the critical role of ethical guidelines in the advancement and deployment of natural language processing technologies.

## References


Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Short Papers*, pages 197–200. The Association for Computer Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. DUAL-REFLECT: enhancing large language models for reflective translation through dual learning feedback mechanisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Student Research Workshop, Bangkok, Thailand, August 11-16, 2024*, pages 693–704. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Trans. Assoc. Comput. Linguistics*, 9:774–789.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10960–10977. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *CoRR*, abs/2304.02554.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Aditya Kalyanpur, Kailash Saravanakumar, Victor Barres, Jennifer Chu-Carroll, David Melville, and David Ferrucci. 2024. Llm-arc: Enhancing llms with an automated reasoning critic. *arXiv preprint arXiv:2406.17663*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1265–1285. Association for Computational Linguistics.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. DEBATE: devil's advocate-based assessment and text evaluation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1885–1897. Association for Computational Linguistics.

Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023a. When reviewers lock horn: Finding disagreement in scientific peer reviews. *arXiv preprint arXiv:2310.18685*.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2024. Can large language models unlock novel scientific research ideas? *arXiv preprint arXiv:2409.06185*.

Sandeep Kumar, Guneet Singh Kohli, Tirthankar Ghosal, and Asif Ekbal. 2023b. Mup-scidocsum: Leveraging multi-perspective peer review summaries for scientific document summarization. In *International Conference on Asian Digital Libraries*, pages 250–267. Springer.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ruosen Li, Teerth Patel, and Xinya Du. 2024. PRD: peer rank and discussion improve large language model based evaluations. *Trans. Mach. Learn. Res.*, 2024.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024b. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2241–2252. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *CoRR*, abs/2305.13091.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023. Towards reasoning in large language models via multi-agent peer review collaboration. *arXiv preprint arXiv:2311.08152*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024a. Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate. *arXiv preprint arXiv:2408.04472*.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arik. 2024b. Chain of agents: Large language models collaborating on long-context tasks. *arXiv preprint arXiv:2406.02818*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 563–578. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xin Zheng, Jie Lou, Boxi Cao, Xueru Wen, Yuqiu Ji, Hongyu Lin, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. 2024. Critic-cot: Boosting the reasoning abilities of large language model via chain-of-thoughts critic. *arXiv preprint arXiv:2408.16326*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and

Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 94–108. Association for Computational Linguistics.

## A An example of how the Defender helps correct the Critic's harsh feedback.

**Aspect: Coherence**
**Human Score: 2**
**Grader Score: 3**
**Critic Score: 1**
**Defender Score: 2**
**Grader Revised Score: 2**
Grader Final Feedback: "The final revised score is **2**. Here's a breakdown of why: **The defender's argument is more balanced.** While the summary has significant flaws, it's not entirely incoherent. It does convey some basic information about Henderson's new contract and his performance. **The critic's score of 1 is too harsh.** The summary does have a basic structure, even if it's not perfect. **The defender's score of 2 is more accurate.** It acknowledges the summary's weaknesses while recognizing its minimal coherence.
The summary is flawed, but it's not completely devoid of coherence. It's important to consider the overall structure and flow of the summary, even if it contains errors and omissions."

In the above example, the initial score given by the Grader is 3, which is higher than the actual human score 2. This indicates that the grader awarded a higher score based on the text's coherence. Next, the Critic challenges the high score by pointing out mistakes in the previous assessment and suggests a score of 1. The Defender, however, argues that the critic's assessment is too harsh and recommends a score of 2, which matches the human score. Finally, the Grader considers the feedback from both the critic and the Defender and, based on their reasoning, finds the defender's argument stronger and more reasonable.

## B An example of how the Critic helps correct the Defender's feedback.

**Aspect: Coherence**
**Human Score: 2**
**Grader Score: 3**
**Critic Score: 2**
**Defender Score: 3**
**Grader Revised Score: 2**
Grader Final Feedback: "The final revised score is **2**. Here's a breakdown of why: The defender makes some valid points, but the critic's feedback is more accurate. While the summary does convey the news of the contract agreement, it lacks crucial details and logical flow, making it difficult to understand the full significance of the news.
Here's a breakdown of why a score of 2 is more appropriate:
* **Missing Key Information:** The summary omits the contract length (2020) and the significant salary increase (ŏ0a3100,000 per week). These are essential details that provide context and impact to the news.
* **Lack of Logical Flow:** The summary ends abruptly, leaving the reader with a sense of incompleteness. The final sentence about Henderson's advisors being in talks doesn't provide a clear conclusion or connection to the main point.
* **Redundancy:** While not a major issue, the repetition of Henderson's vice-captaincy and remaining contract time makes the summary feel clunky and less concise.
**Final Revised Score: 2**
The summary is not completely incoherent, but it lacks the necessary details and logical flow to be considered highly coherent. A score of 2 reflects its shortcomings while acknowledging that it does convey the basic news of the contract agreement."

In this example, the initial score given by the Grader is 3, which is higher than the actual human score of 2. This suggests that the Grader assigned a higher score based on the text's coherence. The Critic then challenges this score, pointing out mistakes in the previous assessment and proposing a score of 2. The Defender, however, argues that the Critic's assessment is too harsh and maintains that a score of 3 is more appropriate, aligning with the

human score. Ultimately, the Grader considers the feedback from both the Critic and the Defender and, after evaluating their reasoning, concludes that the Critic's argument is stronger and more reasonable.

## C Detailed Cost and Feasibility Analysis of CourtEval

### C.1 Computational Requirements

CourtEval involves three distinct roles—Judge (Grader), Prosecutor (Critic), and Defense Attorney (Defender)—resulting in approximately three times the runtime per iteration compared to single-agent models. As discussed in Section 4.5, performance improves up to four iterations, after which additional iterations yield diminishing returns. Running four iterations leads to a total runtime of $4 \times 3 = 12$ times that of single-agent models. However, CourtEval allows users to adjust the number of iterations based on available computational resources and task requirements.

### C.2 Cost Estimation

To evaluate the economic feasibility of CourtEval, we computed the cost of evaluating a single LLM-generated output using GPT-4o. For a single evaluation metric on the SummEval dataset (e.g., consistency), the average processing cost per instance is USD 0.0377 (Input: 19,287 tokens; Output: 2,712 tokens; input price: USD 1.25 per million tokens; output price: USD 5 per million tokens). This means evaluating 100 LLM outputs for a single metric costs approximately USD 3.77.

## D Error Analysis

We discuss the following key aspects based on our human error analysis:

- **Over-penalization for minor grammatical errors**: CourtEval tends to lower fluency scores significantly for small capitalization or punctuation mistakes, such as incorrect capitalization of names or misplaced spaces around punctuation marks. While these are valid errors, the human evaluators seem to weigh them less heavily when the overall coherence of the summary remains intact. This leads to cases where the LLM judge assigns a lower fluency score (e.g., a 2 or 3), while the human score remains higher (e.g., a 4 or 5).

- **Excessive focus on formality**: CourtEval tends to penalize colloquial or informal phrases (e.g., "put pen-to-paper"), whereas human evaluators often accept such expressions as appropriate within certain contexts, such as sports journalism. This difference in perceived fluency results in a score mismatch, with the LLM judge assigning lower scores for fluency due to informality, even though human evaluators may not consider it a significant issue.

- **Structural Transition Over-Penalization**: In particular, the LLM judge overly penalizes disjointed sentence structures, even when the overall summary still presents a coherent narrative from a human perspective. This suggests that the LLM model is heavily reliant on surface-level structural patterns, prioritizing logical flow over the comprehensiveness and clarity of the message being conveyed. Consequently, even summaries that adequately cover key points but lack smooth transitions receive lower scores. An example of this case has been explained in example 1.

**System Output:** Jordan henderson has provided liverpool with a lift after their fa cup heartache by agreeing a new long-term contract. The club's vice-captain had 14 months remaining on his current contract and his advisors had been in talks with liverpool since the beginning of this season. They have now reached a resolution and henderson is expected to put pen-to-paper on improved terms that are likely be worth in the region of £100,000.
Here, the LLM judge penalized the summary for failing to include broader context or transitions, particularly regarding the significance of the new contract or related developments. The human evaluators, however, found this summary more coherent, likely due to its clear focus on the main topic, even though transitions between the contract negotiations and financial terms could have been smoother. The LLM's strictness in judging the logical flow led to a much lower score than that of the human evaluators.
**LLM Judge Score (Coherence)**: 2
**Human Score (Coherence)**: 4.0