arXiv:2403.19305v2 [cs.CL] 15 Apr 2024

MATEval: A Multi-Agent Discussion Framework for Advancing Open-Ended Text Evaluation

Yu Li[1,⋆], Shenyu Zhang[1,⋆], Rui Wu[2], Xiutian Huang[2], Yongrui Chen[1],
Wenhao Xu[2(🖂)], Guilin Qi[1 (🖂)], and Dehai Min[1]

[1] Southeast University, Nanjing, China
{yuli_11, shenyuzhang, yrchen, gqi, zhishanq}@seu.edu.cn
[2] Ant Group, Hangzhou, China
{guli.wr, xiutian.hxt, hao.xuwh}@antgroup.com

Abstract. Recent advancements in generative Large Language Models (LLMs) have been remarkable, however, the quality of the text generated by these models often reveals persistent issues. Evaluating the quality of text generated by these models, especially in open-ended text, has consistently presented a significant challenge. Addressing this, recent work has explored the possibility of using LLMs as evaluators. While using a single LLM as an evaluation agent shows potential, it is filled with significant uncertainty and instability. To address these issues, we propose the **MATEval**: A "**M**ulti-**A**gent **T**ext **Eval**uation framework" where all agents are played by LLMs like GPT-4. The MATEval framework emulates human collaborative discussion methods, integrating multiple agents' interactions to evaluate open-ended text. Our framework incorporates self-reflection and Chain-of-Thought (CoT) strategies, along with feedback mechanisms, enhancing the depth and breadth of the evaluation process and guiding discussions towards consensus, while the framework generates comprehensive evaluation reports, including error localization, error types and scoring. Experimental results show that our framework outperforms existing open-ended text evaluation methods and achieves the highest correlation with human evaluation, which confirms the effectiveness and advancement of our framework in addressing the uncertainties and instabilities in evaluating LLMs-generated text. Furthermore, our framework significantly improves the efficiency of text evaluation and model iteration in industrial scenarios.

Keywords: Multi-Agent · Large Language Models · Text Evaluation

1 Introduction

Evaluating the text generated by large language models (LLMs) has long been a challenging task, Traditional manual evaluation methods are not only time-consuming and laborious but also expensive [2]. Although methods like BLEU [14], Rouge [10], and METEOR [1] have achieved success in scenarios such as

[⋆] Equal Contributors.

machine translation, these automated evaluation methods are limited in the context of open-ended text generation [11]. Recently, LLMs have been used as evaluators such as G-Eval [4], but these methods exhibit unstable and uncertain evaluation effects [15] [16]. Even certain collaboration frameworks could alleviate this problem by employing multi-agent discussion, *e.g.* ChatEval [3], however, the current methods of multi-agent collaboration remain limited to simple interactions, without fully harnessing the potential for agents' *thinking* and *planning*. Additionally, reaching a consensus within multi-agent discussion frameworks continues to be a challenging issue. Furthermore, traditional text evaluation models typically provide only a score without explaining it, making it difficult for reviewers to trust the reliability of these scores. They still need to manually identify errors, obviously slowing down the collection of bad cases and, consequently, affecting the pace of model iteration in industrial scenarios.

To address the above challenges, this paper introduces a Multi-Agent Text Evaluation Framework (MATEval). In this framework, we simulate the human collaborative process in evaluating texts generated by LLMs and propose a novel multi-agent discussion strategy. This strategy integrates self-reflection [13] and Chain-of-Thought (CoT) [17] concepts, as self-reflection focuses on understanding the depth of issues but may lead to rigid thinking. Meanwhile, strategies based on the CoT emphasize the refinement of problems but may lack in-depth analysis of specific issues. Therefore, we combine the two approaches by guiding agents through prompts to decompose evaluation questions and focus on only one sub-question in each discussion round. During each round of the discussion, agents engage in self-reflection, considering peer inputs to enrich issue comprehension. This approach strengthens agents' self-assessment and critical thinking, broadening their evaluation scope for open-ended text and aligning results more closely with human evaluations.

Furthermore, our framework introduces a feedback mechanism at the end of each discussion round to evaluate the quality and efficiency of the discussions, encouraging agents to reach a consensus. The comprehensive evaluation report generated by our framework details error types, specific locations, in-depth explanations, and corresponding scores.

For practical applications in industry, we provide two report formats: a question-and-answer format for strategy analysis and a text report designed to help business personnel quickly identify errors and facilitate iterative improvement of LLMs. Our framework has achieved significant results in the story text evaluation task at Alipay, markedly enhancing the efficiency of model iteration.

To summarize, our main contributions in this paper are:

1. We propose a Multi-Agent Evaluation Framework called MATEval1, which enhances the reliability of scoring by providing accurate diagnostic reports for text generated by LLMs. This framework not only facilitates model iteration in industrial scenarios but also significantly boosts audit efficiency.
2. We propose a novel method to integrate self-reflection and CoT in our multi-agent framework. Additionally, we creatively introduce a feedback mech-

anism at the end of each discussion round to resolve disagreements and facilitate the achievement of consensus.

3. We conduct comprehensive experiments on two English and two Chinese story text datasets, including one constructed based on Alipay's business story text dataset. Our experimental results showcase the effectiveness of our framework and its high correlation with human evaluations. [1]

2 Related Work

Traditional NLG Evaluation: For a significant period, open-ended text evaluation primarily depended on human annotations, which incurs substantial human and financial costs. Subsequent automated NLG evaluations employ computational models to evaluate the quality of generated texts, such as BLEU [14], ROUGE [10], and METEOR [1]. Embedding-based metrics refer to the evaluation of generated texts by measuring the semantic similarity between generated texts and reference texts based on word or sentence embeddings. BERTScore [18] calculates the similarity between generated text and reference text based on BERT's contextual embedding. $\text{RUBER}_{\text{BERT}}$ [5] is also based on BERT embeddings to measure the similarity of texts with and without references through processes such as pooling and MLP operations.

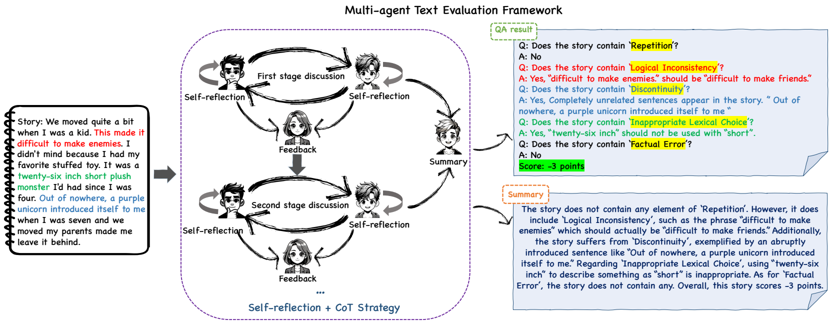
LLM-based Evaluators: GPTScore [4] utilizes models such as GPT-3 to evaluate text quality, predicated on the assumption that generative pre-trained models assign higher probabilities to high-quality generated texts by given instructions and context. Recent studies also explore the potential of using ChatGPT as an NLG evaluator [15]. G-Eval [12] demonstrates the evaluation of NLG outputs using prompts in LLMs like ChatGPT through Chain-of-Thought (CoT) methods.

Communicative Agents: Recently, the concept of using agents for communication and collaboration to accomplish specific tasks gains widespread application. CAMEL [9] introduces a cooperative agent framework called *role-playing*, enabling individual agents to collaboratively solve complex tasks autonomously. ChatEval [3] applies the multi-agent approach to text evaluation, constructing a multi-agent jury to explore the impact of different communication strategies in evaluating open-ended questions and traditional NLG tasks.

3 Methodology

In this section, we will provide a detailed exposition of the design of the evaluation framework within MATEval 1, the utilization of various strategies, and the functional specifications of different roles.

[1] We have made the datasets and results used in our experiments publicly available at `https://github.com/kse-ElEvEn/MATEval`. Due to the user privacy of Alipay, we cannot make the "Ant" dataset public.



Story: We moved quite a bit when I was a kid. This made it difficult to make enemies. I didn't mind because I had my favorite stuffed toy. It was a twenty-six inch short plush monster I'd had since I was four. Out of nowhere, a purple unicorn introduced itself to me when I was seven and we moved my parents made me leave it behind.

Q: Does the story contain 'Repetition'?
A: No
Q: Does the story contain 'Logical Inconsistency'?
A: Yes, "difficult to make enemies." should be "difficult to make friends."
Q: Does the story contain 'Discontinuity'?
A: Yes, Completely unrelated sentences appear in the story. " Out of nowhere, a purple unicorn introduced itself to me "
Q: Does the story contain 'Inappropriate Lexical Choice'?
A: Yes, "twenty-six inch" should not be used with "short".
Q: Does the story contain 'Factual Error'?
A: No
Score: -3 points

The story does not contain any element of 'Repetition'. However, it does include 'Logical Inconsistency', such as the phrase "difficult to make enemies" which should actually be "difficult to make friends." Additionally, the story suffers from 'Discontinuity', exemplified by an abruptly introduced sentence like "Out of nowhere, a purple unicorn introduced itself to me." Regarding 'Inappropriate Lexical Choice', using "twenty-six inch" to describe something as "short" is inappropriate. As for 'Factual Error', the story does not contain any. Overall, this story scores -3 points.

Fig. 1: The overall process diagram of the MATEval Framework. The input to the framework is a text with quality questions, which after going through a multi-agent discussion that combines self-reflection and CoT strategies, outputs a detailed evaluation report.

3.1 Design of the Framework

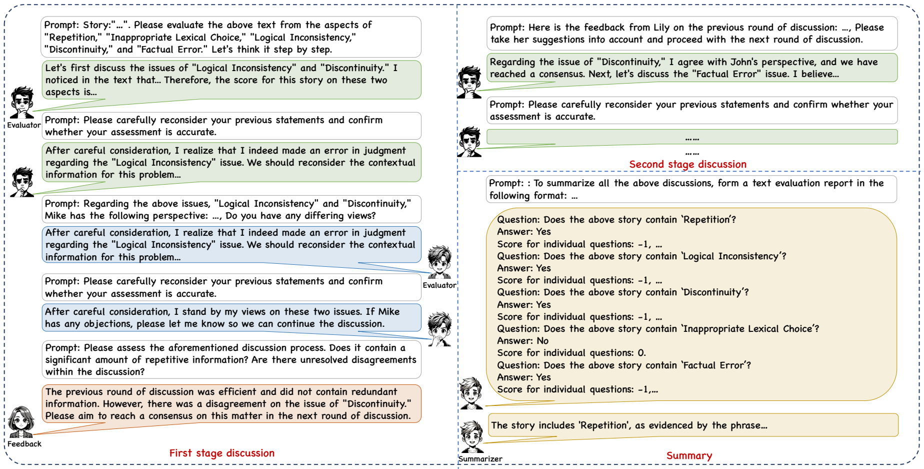
Our framework primarily consists of agents with different roles combined with discussion strategies. The roles of agents we utilize include *Evaluator Agent*, *Feedback Agent*, and *Summarizer Agent*, who collaborate to complete text evaluation tasks. The *Evaluator Agent* is the main entity in the evaluation task, the *Feedback Agent* plays a crucial role in improving discussion quality and promoting consensus, and the *Summarizer Agent* is indispensable for consolidating discussion information, summarizing, and forming evaluation reports. In our framework, we employ a discussion strategy that integrates self-reflection and CoT.

The framework accepts text as input, which may contain various quality issues. The output is a detailed evaluation report outlining error type, location, explanation, and score. We present the results in two formats: one is a Q&A format conducive to evaluating correlation, allowing easy extraction of correlation scores for similarity calculations. The other is a report format that is conducive to iteration by relevant business personnel. This enables business personnel to quickly identify text issues and refine models using the analysis reports, enhancing efficiency. This is shown in the right part of Figure 1.

3.2 Application of Different Roles

In this section, we introduce several key roles within our framework and their respective functions.

Evaluator Agent: The core element in the framework is the evaluator, for which we use GPT-4 to conduct multi-round evaluations and responses that are guided through carefully designed prompts. The evaluator stores and processes statements from other agents, using this as a reference for dialogue history. Each agent not only receives responses from others but also provides their own statements, with the entire process requiring minimal human intervention.



Prompt: Story:"...". Please evaluate the above text from the aspects of "Repetition," "Inappropriate Lexical Choice," "Logical Inconsistency," "Discontinuity," and "Factual Error." Let's think it step by step.

Let's first discuss the issues of "Logical Inconsistency" and "Discontinuity." I noticed in the text that... Therefore, the score for this story on these two aspects is...

Prompt: Please carefully reconsider your previous statements and confirm whether your assessment is accurate.

After careful consideration, I realize that I indeed made an error in judgment regarding the "Logical Inconsistency" issue. We should reconsider the contextual information for this problem...

Prompt: Regarding the above issues, "Logical Inconsistency" and "Discontinuity," Mike has the following perspective: ..., Do you have any differing views?

After careful consideration, I realize that I indeed made an error in judgment regarding the "Logical Inconsistency" issue. We should reconsider the contextual information for this problem...

Prompt: Please carefully reconsider your previous statements and confirm whether your assessment is accurate.

After careful consideration, I stand by my views on these two issues. If Mike has any objections, please let me know so we can continue the discussion.

Prompt: Please assess the aforementioned discussion process. Does it contain a significant amount of repetitive information? Are there unresolved disagreements within the discussion?

The previous round of discussion was efficient and did not contain redundant information. However, there was a disagreement on the issue of "Discontinuity." Please aim to reach a consensus on this matter in the next round of discussion.

Prompt: Here is the feedback from Lily on the previous round of discussion: ..., Please take her suggestions into account and proceed with the next round of discussion.

Regarding the issue of "Discontinuity," I agree with John's perspective, and we have reached a consensus. Next, let's discuss the "Factual Error" issue. I believe...

Prompt: Please carefully reconsider your previous statements and confirm whether your assessment is accurate.

......

......

Prompt: : To summarize all the above discussions, form a text evaluation report in the following format: ...

Question: Does the above story contain 'Repetition'?
Answer: Yes
Score for individual questions: -1, ...
Question: Does the above story contain 'Logical Inconsistency'?
Answer: Yes
Score for individual questions: -1, ...
Question: Does the above story contain 'Discontinuity'?
Answer: Yes
Score for individual questions: -1, ...
Question: Does the above story contain 'Inappropriate Lexical Choice'?
Answer: No
Score for individual questions: 0.
Question: Does the above story contain 'Factual Error'?
Answer: Yes
Score for individual questions: -1,...

The story includes 'Repetition', as evidenced by the phrase...

Fig. 2: The diagram includes prompts and dialogue that incorporates a process of discussion with self-reflection, CoT, feedback mechanisms and final summary.

Feedback Agent: The feedback agent evaluates the content and quality of each discussion round. It focuses on identifying inefficient dialogues and disagreements. If issues are detected, it suggests improvements for subsequent rounds to enhance efficiency and consensus through prompts.

Summarizer Agent: After all discussions are concluded, the summarizer compiles the entire process and outcomes. It provides a Q&A format evaluation report, detailing the identification, analysis, and scoring of various issues. Additionally, we provide a comprehensive text-based format evaluation report that includes detailed problem descriptions and is easy to read to help improve model performance in industrial production.

3.3 Feedback Mechanism

The feedback mechanism is a well-designed component in our framework. At the end of each discussion round, we use a prompt to guide a *feedback agent* to summarize and evaluate the discussion. Its role is to steer subsequent discussions towards less repetition, enhance the efficiency of the discussion, and importantly, guide the participants towards reaching a consensus. All of these are achieved by conveying the feedback provider's remarks to the agents involved in the discussion.

3.4 Combined Self-reflection and CoT

Self-reflection Strategy: After each agent's statements, they engage in a process of self-reflection. Guided by the prompt, agents adjust their statements by integrating the viewpoints of other agents. The final statements of each agent are stored and used as historical information for subsequent discussions. In a new round of discussion, the statements from the previous round are stored as historical information.

CoT Strategy: We guide agents through prompts to autonomously decompose problems and address only one sub-problem in each round of discussion. Meanwhile, each agent's statements are stored and used as historical information for subsequent discussions.

Combined Self-reflection and CoT: As shown in Algorithm 1. Combining self-reflection and the CoT is an important strategy employed in our framework. Agents autonomously decompose the question according to the prompt, focusing on one sub-question in each round of multiple discussions: $\mathcal{D}(\mathcal{Q}) = \{\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_{\mathrm{n}}\}$, $\mathcal{P}_{\mathrm{i}} = \mathcal{I}(\mathcal{Q}_{\mathrm{i}}, \mathcal{H})$, where $\mathcal{I}$ represents the formation of preliminary ideas based on the prompt, $\mathcal{Q}$ is the evaluation task, $\mathcal{Q}_{\mathrm{n}}$ is the sub-question and $\mathcal{H}$ is the history information. They then optimize their statements through self-reflection: $\mathcal{R}_{\mathrm{i}} = \mathcal{S}(\mathcal{P}_{\mathrm{i}}, \mathcal{H})$. Next, update the history: $\mathcal{H} = \mathcal{H} \cup \{\mathcal{R}_i\}$. After each round of discussion, a feedback provider evaluates the discussion to reduce repetition and disagreement: $\mathcal{E} = \mathrm{E}(\mathcal{H})$. Finally, a summarizer compiles all statements to produce the evaluation report: $\mathcal{R}_{\mathrm{final}} = \mathrm{Summary}(\mathcal{H})$. The overall prompt and the flow of the discussion are illustrated in Figure 2.

Algorithm 1 Self-Reflection and Chain-of-Thought Discussion with Feedback

```
Input: Given text, set of sub-questions $\mathcal{Q}$ decomposed by LLMs-based agents, number
    of agents $\mathcal{N}$
Output: Final evaluation report $R_{\mathrm{final}}$
$\mathcal{H} \leftarrow \{\}$, Agents $\leftarrow \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_N\}$, SubQIndex $\leftarrow 0$
while SubQIndex $<$ length($\mathcal{Q}$) do
    SubQCurrent $\leftarrow \mathcal{Q}$[SubQIndex]          ▷ select the current sub-question
    SubQIndex $\leftarrow$ SubQIndex $+ 1$                 ▷ increment the sub-question index
    for each $\mathcal{A}_i$ in Agents do
        $\mathcal{P}_i \leftarrow$ Formulate_Idea($\mathcal{Q}_{\mathrm{i}}, \mathcal{H}$)          ▷ generate preliminary ideas
        $\mathcal{R}_i \leftarrow$ Self_Reflection($\mathcal{P}_i, \mathcal{H}$)          ▷ reflect on their statements
        $\mathcal{H} \leftarrow \mathcal{H} \cup \{\mathcal{R}_i\}$     ▷ update historical record with reflected statement
    end for
    $\mathcal{E} \leftarrow$ Evaluate($\mathcal{H}$)               ▷ evaluator assesses this round's discussion
    $\mathcal{F}_{\mathrm{round}} \leftarrow$ Feedback($\mathcal{E}$)             ▷ provide feedback $F_{\mathrm{round}}$ to all agents
end while
$\mathcal{R}_{\mathrm{final}} \leftarrow$ Summary($\mathcal{H}$)              ▷ summarizer generates error analysis report
return $\mathcal{R}_{\mathrm{final}}$
```

4 Experiments

4.1 Implementation Details

In the MATEval framework, we select OpenAI's GPT-4 as our LLMs due to its outstanding performance and API accessibility. We set the temperature parameter to 0 for result reproducibility. GPT-4's easy access facilitated effective and coherent multi-agent interactions in our experiments.

4.2 Dataset

We mainly apply our framework to the evaluation of story texts generated by LLMs in Alipay business scenarios. So in the experiment, we mainly focus on two open-ended story datasets: ROCStories (**ROC**) [6] and WritingPrompts (**WP**) [6]. Considering GPT-4's context length limitations and the need for storing multi-round discussion contexts, we truncate WP stories to the first 200 words, ensuring textual integrity at the truncation point. To test model generalizability, we conduct similar experiments on two Chinese datasets. These include the *Chinese **LO**ng **T**ext understanding and generation* **(LOT)** [7] dataset, comprising human-written stories averaging 106 words, and a dataset of Chinese fairy tales constructed using prompts from Alipay's business data with GPT-3.5 named **Ant**, mainly involving fables and fairy tales with an average story length of 125 words. Considering GPT-4's request rate limits and high usage costs, we select the first 200 stories from each dataset for multi-agent discussion experiments.

Using the GPT-4 interface, we introduce five basic types of errors into 200 story texts across different datasets to simulate possible problems in stories. They are ***Repetition*****(REP), *Logical Inconsistency* (LINC), *Discontinuity* (DCONT), *Inappropriate Lexical Choice* (ILC), and *Factual Error* (FER)**. Repetition includes redundant sentences or excessive word use; Logical Inconsistency encompasses antonym substitution and polarity shifts in sentences; Discontinuity involves sequencing errors or unrelated content; Inappropriate Lexical Choice refers to misused quantifiers or pronouns; and Factual Error denotes contradictions with established knowledge. We hire five annotators to assess these datasets, ensuring the data aligns with human preferences. Both manual and multi-agent scoring follow the same criteria: starting from zero, each error deducts one point, with scores tallied for each error type and the total for each text.

4.3 Compared Methods

Referenced Metrics: The ***BLEU*** [14] score is used to evaluate lexical similarity between candidate and reference texts. ***ROUGE-L*** [10] focuses on the longest common subsequence to assess the fluency and coherence of texts. And ***RUBER-BERT*** **[5],** an enhancement of the original RUBER model with BERT's contextual embeddings, includes both referenced and unreferenced versions. The referenced version, ***RUBER-BERTr***, measures the similarity between candidate responses and reference texts using BERT word embeddings.

Unreferenced Metrics: The RUBER's unreferenced version ***RUBER-BERTu*** [5] predicts relevance between responses and queries using BERT word embeddings followed by operations such as pooling and MLP. The BERT-based ***UNION*** [8] model distinguishes human-written stories from automatically generated negative samples and corrects the interference of negative samples. ***ChatEval*** [3] evaluates open-ended Q&A quality through multi-agent framework, and we select its most effective *One-by-One* approach for comparison.

Our Methods: In our experience, we compare various strategies:

- **Single-Agent(SA)**: LLMs directly evaluate stories without multi-agent.
- **One-by-One** [3]**(O_b_O)**: Agents sequentially evaluate stories in multi-round discussions, without optimization strategies.
- **Self-Reflection (SR)**: Agents conduct self-reflection, considering their and others' previous statements during discussions.
- **Chain-of-Thought (CoT)**: Agents break down the assessment problem through prompts, solving one sub-problem in each discussion round.
- **Self-Reflection + CoT (SR+CoT)**: By combining CoT and self-reflection strategies, agents first decompose questions for discussion, then engage in self-reflection each round.

In all these strategies, we employ feedback mechanisms at the end of each discussion round, as well as a final summary.

4.4 Experimental Results

Our experiments on the ROC and WP datasets are presented in Table 1, using MATEval framework strategies to evaluate narrative texts. We calculate Spearman (ρ) and Kendall (τ) correlation coefficients to compare models' evaluations with human judgments.

Analyzing the experimental results on the ROC and WP datasets, we draw the following conclusions:

LLMs-based methods show better Spearman (ρ) and Kendall (τ) correlations compared to traditional n-gram and Bert-based methods, proving their effectiveness in text evaluation.

Multi-agent discussions generally surpass single-agent evaluations in performance, suggesting they significantly improve text evaluation quality.

In analyzing the strategies used within the MATEval framework, we found that employing self-reflection or the CoT independently produces unstable results across different error types. In some cases, these methods even underperformed compared to single-agent evaluations. This might be due to inherent flaws when applying these strategies separately. For example, self-reflection can lead to rigid thinking in multiple discussion rounds, where agents often repeat earlier content without adding new insights. On the other hand, using CoT alone often results in superficial and divergent perspectives, offering only a basic analysis of each error type without delving deeper.

The combination of self-reflection and CoT achieved the best overall correlation, particularly excelling over other methods in evaluating *Logical Inconsistency*, *Discontinuity*, and *Inappropriate Lexical Choice*. It significantly improves the evaluation of *Discontinuity* compared to the single-agent method, demonstrating the framework's high sensitivity to textual coherence. However, its effectiveness was lower for *Repetition* and *Factual Error*. Agents often misidentified emotionally similar sentences as repetitive, despite clear definitions. Furthermore, the framework's evaluation of *Factual Errors* was limited by LLMs constraints, specifically the absence of external knowledge affecting common sense

Table 1: Correlation of evaluation results with human judgment using different models and different strategies of MATEval on the ROC/WP dataset, where SA stands for Single-Agent, SR denotes Self-Reflection, and CoT represents Chain-of-Thought. The symbols ρ/τ respectively indicate the Spearman/Kendall correlation. The highest correlation values are highlighted in bold.

Strategy		REP		LINC		DCONT		ILC		FER	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROC	BLEU	0.318	0.260	0.193	0.153	0.156	0.128	0.037	0.031	-0.010	-0.008
	ROUGE-$_\text{L}$	-0.017	-0.014	0.129	0.102	0.202	0.165	0.056	0.045	0.104	0.084
	RUBER_r	0.036	0.035	0.054	0.049	0.315	0.297	-0.018	-0.017	-0.176	-0.166
	RUBER_u	-0.111	-0.091	0.038	0.031	0.131	0.107	0.134	0.110	0.180	0.146
	UNION	-0.093	-0.076	0.091	0.071	-0.018	-0.015	0.057	0.046	0.072	0.059
	SA	0.699	0.694	0.268	0.253	0.318	0.312	0.240	0.236	0.545	0.538
	O_b_O	0.698	0.692	0.170	0.160	0.356	0.349	0.259	0.248	0.484	0.473
	SR	0.691	0.680	0.169	0.154	0.354	0.339	0.144	0.138	0.498	0.478
	CoT	0.743	**0.737**	0.189	0.180	0.288	0.282	0.213	0.205	0.502	0.491
	SR+CoT	**0.735**	0.728	**0.281**	**0.264**	**0.391**	**0.382**	**0.263**	**0.256**	**0.575**	**0.561**
WP	BLEU	0.087	0.071	0.096	0.073	0.039	0.033	-0.114	-0.091	0.009	0.007
	ROUGE-$_\text{L}$	0.092	0.074	0.127	0.096	0.083	0.068	-0.046	-0.037	0.049	0.040
	RUBER_r	0.038	0.036	-0.020	-0.018	-0.081	-0.076	0.035	0.033	0.076	0.071
	RUBER_u	-0.102	-0.084	0.054	0.041	-0.006	-0.005	-0.006	-0.007	0.111	0.089
	UNION	0.048	0.039	0.010	0.008	-0.110	-0.090	-0.038	-0.031	0.052	0.042
	SA	0.258	0.246	0.107	0.095	0.111	0.105	0.192	0.1802	0.176	0.171
	O_b_O	0.386	0.380	0.183	0.166	0.081	0.075	0.089	0.082	**0.299**	**0.286**
	SR	**0.491**	**0.483**	0.120	0.107	0.224	0.209	0.057	0.051	0.214	0.208
	CoT	0.132	0.129	0.159	0.139	0.203	0.191	0.002	0.001	0.218	0.211
	SR+CoT	0.430	0.417	**0.215**	**0.188**	**0.265**	**0.248**	**0.290**	**0.266**	0.299	0.286

error detection, highlighting the need to integrate external knowledge for future multi-agent framework enhancements.

4.5 Ablation Study

To assess the effectiveness of different modules in MATEval, we conducted ablation experiments on the ROC dataset, involving the removal of the feedback mechanism, omitting Q&A format explanations, and not using a multi-agent approach. Table 3 demonstrates that the complete MATEval framework surpasses its ablated versions, confirming the significance of both feedback mechanisms, explanations and multi-agent methods. Specifically, this establishes the importance of our feedback mechanisms in promoting discussion consensus and enhancing relevance. It also proves that providing scoring explanations by LLMs significantly enhances evaluation effectiveness.

Table 2: Correlation of evaluation results with human judgment using different models and different strategies of MATEval on the LOT/Ant dataset. The highest correlation values are highlighted in bold.

Strategy		REP		LINC		DCONT		ILC		FER	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
LOT	SA	**0.829**	**0.817**	0.120	0.110	0.336	0.324	0.179	0.175	0.284	0.279
	O_b_O	0.770	0.764	0.142	0.131	0.249	0.239	0.069	0.066	**0.362**	**0.349**
	SR	0.751	0.735	0.054	0.048	0.282	0.267	0.118	0.112	0.296	0.284
	CoT	0.636	0.628	0.026	0.024	0.215	0.206	0.051	0.049	0.155	0.151
	SR+CoT	0.811	0.798	**0.197**	**0.185**	**0.354**	**0.341**	**0.182**	**0.175**	0.341	0.333
Ant	SA	0.522	0.517	0.281	0.275	0.231	0.231	0.318	0.316	0.495	0.489
	O_b_O	0.545	0.538	0.145	0.141	-0.010	-0.010	0.011	0.018	0.347	0.343
	SR	0.563	0.557	0.185	0.175	0.069	0.069	0.187	0.184	0.368	0.360
	CoT	0.572	0.562	0.034	0.039	0.024	0.024	0.040	0.042	0.358	0.352
	SR+CoT	**0.694**	**0.676**	**0.403**	**0.377**	**0.287**	**0.282**	**0.424**	**0.417**	**0.552**	**0.544**

Table 3: Correlation of evaluation results with human judgment of ablation experiments on the ROC dataset.

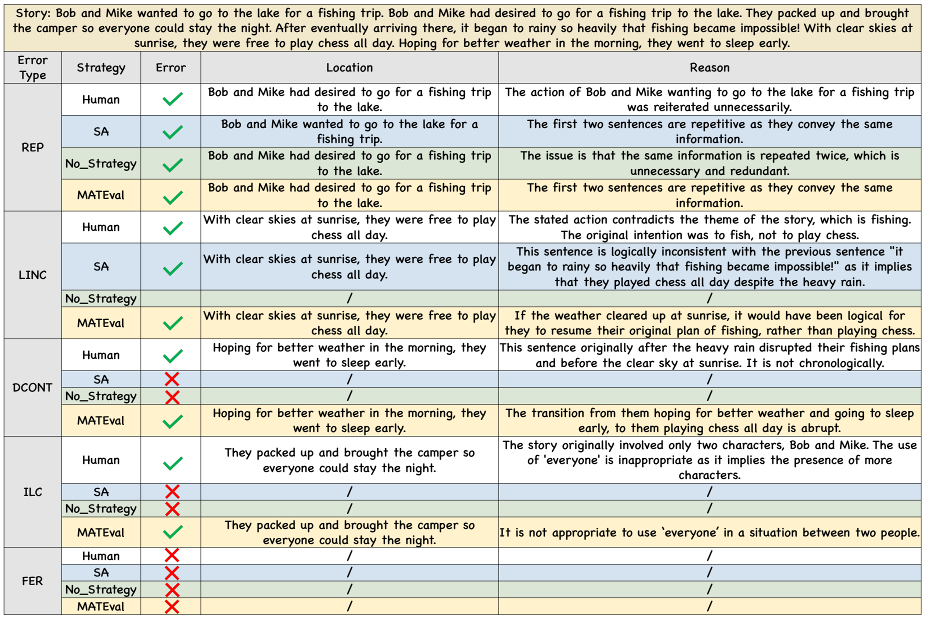
Strategy	REP		LINC		DCONT		ILC		FER	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
MATEval-FB	0.567	0.567	0.039	0.038	0.259	0.255	0.266	0.260	0.477	0.473
MATEval-QA	0.612	0.597	0.071	0.068	0.011	0.011	0.132	0.127	0.283	0.281
MATEval-multi	0.699	0.694	0.268	0.253	0.318	0.312	0.240	0.236	0.545	0.538
MATEval	**0.735**	**0.728**	**0.281**	**0.264**	**0.391**	**0.382**	**0.263**	**0.256**	**0.575**	**0.561**

4.6 Generalization Experiments

To verify the generalizability of our framework across different languages and in the industrial field, we experimented with two Chinese datasets: *LOT* and a story text dataset *Ant*, derived from Alipay's business data. The findings in Table 2 were similar to those from English datasets. Interestingly, single-agent evaluations often perform better in Chinese, possibly due to its unique language and sentence structure. This suggests that optimizing agents for different languages may require tailored adjustments for optimal performance.

4.7 Case Study

To showcase the MATEval framework's effectiveness in real-world industrial settings, we illustrated its application using a story text example shown in Table 3. We compared our framework against manual evaluation, single-agent evaluation, and the strategy-less One_by_One evaluation method. Experimental results



Story: Bob and Mike wanted to go to the lake for a fishing trip. Bob and Mike had desired to go for a fishing trip to the lake. They packed up and brought the camper so everyone could stay the night. After eventually arriving there, it began to rainy so heavily that fishing became impossible! With clear skies at sunrise, they were free to play chess all day. Hoping for better weather in the morning, they went to sleep early.

Error Type	Strategy	Error	Location	Reason
REP	Human	✓	Bob and Mike had desired to go for a fishing trip to the lake.	The action of Bob and Mike wanting to go to the lake for a fishing trip was reiterated unnecessarily.
	SA	✓	Bob and Mike wanted to go to the lake for a fishing trip.	The first two sentences are repetitive as they convey the same information.
	No_Strategy	✓	Bob and Mike had desired to go for a fishing trip to the lake.	The issue is that the same information is repeated twice, which is unnecessary and redundant.
	MATEval	✓	Bob and Mike had desired to go for a fishing trip to the lake.	The first two sentences are repetitive as they convey the same information.
LINC	Human	✓	With clear skies at sunrise, they were free to play chess all day.	The stated action contradicts the theme of the story, which is fishing. The original intention was to fish, not to play chess.
	SA	✓	With clear skies at sunrise, they were free to play chess all day.	This sentence is logically inconsistent with the previous sentence "it began to rainy so heavily that fishing became impossible!" as it implies that they played chess all day despite the heavy rain.
	No_Strategy		/	/
	MATEval	✓	With clear skies at sunrise, they were free to play chess all day.	If the weather cleared up at sunrise, it would have been logical for they to resume their original plan of fishing, rather than playing chess.
DCONT	Human	✓	Hoping for better weather in the morning, they went to sleep early.	This sentence originally after the heavy rain disrupted their fishing plans and before the clear sky at sunrise. It is not chronologically.
	SA	✗	/	/
	No_Strategy	✗	/	/
	MATEval	✓	Hoping for better weather in the morning, they went to sleep early.	The transition from them hoping for better weather and going to sleep early, to them playing chess all day is abrupt.
ILC	Human	✓	They packed up and brought the camper so everyone could stay the night.	The story originally involved only two characters, Bob and Mike. The use of 'everyone' is inappropriate as it implies the presence of more characters.
	SA	✗	/	/
	No_Strategy	✗	/	/
	MATEval	✓	They packed up and brought the camper so everyone could stay the night.	It is not appropriate to use ‘everyone’ in a situation between two people.
FER	Human	✗	/	/
	SA	✗	/	/
	No_Strategy	✗	/	/
	MATEval	✗	/	/

Fig. 3: The schematic diagram illustrating the comparison of results generated by the MATEval framework and other methods.

demonstrate that our method is basically consistent with human evaluations, unlike other methods, which show some gaps, thereby confirming the high correlation between our approach and human evaluation.

5 Conclusion

In this paper, we proposed the MATEval framework, which enhances the evaluation performance of open-ended story text generated by LLMs in the industrial field. Extensive experiments show that MATEval's evaluation results on two classic story datasets are more aligned with human preferences than those of existing methods. In the Alipay industrial scenario, our framework significantly improves review efficiency and evaluation accuracy, serving as an effective aid.[2][3]

In future work, we can fine-tune LLMs as agents in the industrial field to complete specific domain tasks. When solving a complex domain task, we can enable these domain-specific agents to collaborate with each other, thereby enhancing their capability and efficiency in addressing challenges.

[2] This work was supported by Ant Group

[3] This work was supported by the Natural Science Foundation of China (Grant No. U21A20488). We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

1. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL. pp. 65–72 (2005)
2. Callison-Burch, C.: 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, EMNLP 2009. pp. 286–295
3. Chan, C.M., et al: Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023)
4. Fu, J., et al: Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166 (2023)
5. Ghazarian, S., et al: Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. arXiv preprint arXiv:1904.10635 (2019)
6. Guan, J., et al: Openmeva: A benchmark for evaluating open-ended story generation metrics. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers). pp. 6394–6407 (2021)
7. Guan, J., et al: LOT: A story-centric benchmark for evaluating chinese long text understanding and generation **10**, 434–451 (2022)
8. Guan, J., Huang, M.: UNION: an unreferenced metric for evaluating open-ended story generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP. pp. 9157–9166 (2020)
9. Li, G., et al: Camel: Communicative agents for" mind" exploration of large scale language model society. arXiv preprint arXiv:2303.17760 (2023)
10. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (Jul 2004)
11. Liu, C., et al: How to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP. pp. 2122–2132 (2016)
12. Liu, Y., et al: 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. pp. 2511–2522 (2023)
13. Madaan, A., et al: Self-refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651 (2023)
14. Papineni, K., et al: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
15. Wang, J., et al: Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048 (2023)
16. Wang, P., et al: Large language models are not fair evaluators. arXiv preprint arXiv:2305.17926 (2023)
17. Wei, J., et al: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
18. Zhang, T., Kishore, V., Wu, F., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR (2020)