How to expand boundaries of evaluator: A Reference Based LLM-as-Evaluator Method

Anonymous ACL submission

Abstract

Evaluating the performance of large language models (LLMs) in diverse domains has been a significant challenge due to the limitations of traditional evaluation metrics and the high cost of manual annotation. This paper introduces the Reference-based LLM-as-Evaluator (Ref-Eval) framework, which leverages the strengths of LLMs in text comprehension and instruction-following to assess model responses. The Ref-Eval framework employs a multi-round dialogic evaluation process, condensing extensive external references into distinct knowledge units, clustering them for efficient evaluation, and iteratively refining questions based on model responses. Experimental results on multiple domain-specific text datasets demonstrate that Ref-Eval achieves a high consistency with human evaluation, saving computational resources and enhancing evaluation accuracy. This approach not only addresses the limitations of existing LLM evaluation methods but also provides a scalable and efficient way to assess model performance in knowledge-intensive tasks.

1 Introduction

Assessing the quality of text generated by language models has been a challenging task for a long time (Chang et al., 2024). Traditional approaches rely on a large amount of manually annotated ground truth benchmarks (Hendrycks et al., 2020; Huang et al., 2024; Gu et al., 2024) as well as n-gram automatic evaluation metrics (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) used to align the ground truth and generated text. However, these metrics, primarily word co-occurrence-based, have been shown to neglect semantics and are insufficient in identifying factual inaccuracies. The demanding cost of annotation and the limitations of n-gram automatic evaluation metrics make it difficult to evaluate open-ended results in various knowledge-intensive domains.

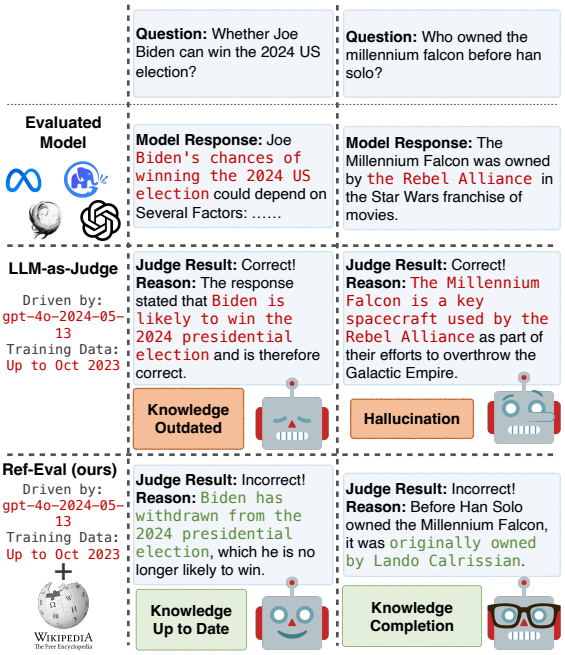


Figure 1: The effect of external reference.

With impressive text comprehension and instruction-following capabilities, "Large Language Model (LLM) as Evaluator" has emerged as a solution to address the previously mentioned evaluation challenges to some extent (Bai et al., 2024; Zheng et al., 2024; Chan et al., 2023). Existing research (Bai et al., 2024; Li et al., 2024) enables LLMs to generate domain-specific questions by starting from predefined terms or keywords and iteratively refining the questions based on model responses. However, these approaches overlook a crucial issue: **the knowledge within Evaluator LLM is limited, incomplete, and often outdated.**

To address the limitations of knowledge within LLMs, it is intuitive to consider incorporating external references (van Schaik and Pugh, 2024; Xie et al., 2023). As shown in Fig. 1, the reference text can correct the model's original misjudgments, providing a more accurate basis for evaluation. How-

ever, integrating external references into the LLM's evaluation process introduces several challenges. (1) The length of external references often exceeds the LLM's processing capacity, making it difficult for the model to extract relevant information efficiently. (2) Analyzing extensive reference materials demands significant computational resources, which raises concerns about scalability and efficiency in practical applications. (3) Due to the LLM's sensitivity to prompts, ensuring consistent and accurate evaluations across references with varying expression styles is challenging. These challenges must be addressed to unlock the full potential of using external references in LLM evaluations.

To overcome these challenges, we propose the **Ref**erence-based LLM-as-**Eval**uator framework (**Ref-Eval**). Ref-Eval is designed to effectively manage and utilize large volumes of external reference by decomposing them into discrete, manageable knowledge units. These units are organized into clusters based on their thematic content, and the evaluation process involves generating targeted questions for each cluster to assess the performance of the model being evaluated. The evaluation focuses on whether the responses adequately cover all relevant knowledge units within each cluster, thereby ensuring the comprehensiveness and precision of the assessment. Any knowledge units that remain unaddressed, indicating that the evaluated model's accuracy cannot be determined, are regrouped for further questioning in subsequent rounds. By adopting this iterative clustering and questioning approach, Ref-Eval not only mitigates the limitations of the LLM's internal knowledge but also enhances the robustness of the evaluation process by efficiently leveraging the rich information contained in external references, thereby reducing the overall evaluation overhead.

We conduct experiments on multiple domain-specific text datasets. The experimental results indicate that Ref-eval achieves a consistency of more than 96% with human results. Compared to the theoretical best method, our method saves 55.9% in money expense and saves 10k dollars in our dataset of 18M tokens. Ref-eval also provides instructive performance characteristics of models on real-world knowledge, for example, knowledge with brief or complex content and models with different capacities of in-context learning (Min et al., 2022).

2 Related work

Open-ended question answering benchmarks for models include MMLU (Hendrycks et al., 2020), C-eval (Huang et al., 2024), Xiezhi (Gu et al., 2024), etc. These benchmarks contain a series of questions that are described in natural language and require the model to give an open-ended answer. However, these benchmarks rely on a large number of manual annotations and cannot be updated with the latest knowledge. At present, some methods propose to use LLM to automatically build updatable benchmarks, such as LM-as-an-Examiner (Bai et al., 2024) and TreeEval (Li et al., 2024). In these methods, in the scenario without a benchmark, LLM raises questions according to its knowledge. However, the knowledge of LLM itself cannot be complete, and the language model is biased, which will lead to incomplete questions and deviation of questions to a certain extent.

Traditional open evaluation metrics are based on n-gram to measure semantic similarity between texts, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), etc. These methods can automate the evaluation of natural language texts. However, these methods lack expressive power and can not distinguish the key information, such as some negative words, semantically. To solve these problems, BERTScore (Zhang et al., 2019), GPTScore (Fu et al., 2023), and other methods use language models to derive similarities between candidate answers and reference answers. In the latest evaluation work on LLM, LLM as Judge (a paradigm that uses LLM to evaluate open results) (Zheng et al., 2024; Chan et al., 2023) is proposed. This method has strong interpretability and scalability, so it has been widely considered. These methods still rely either on artificially constructed reference answers or on knowledge of the LLM itself, and the problem of the limitations of knowledge remains unsolved.

3 Reference-based LLM-as-Evaluator Framework

As shown in Fig. 2, Ref-Eval consists of five main components, and the description of these components is as follows:

Step 1 Preparing Knowledge Unit Candidates: The Ref-Eval framework begins by preparing knowledge units that will be utilized throughout the evaluation process. These units are sourced primarily from two places: external references or

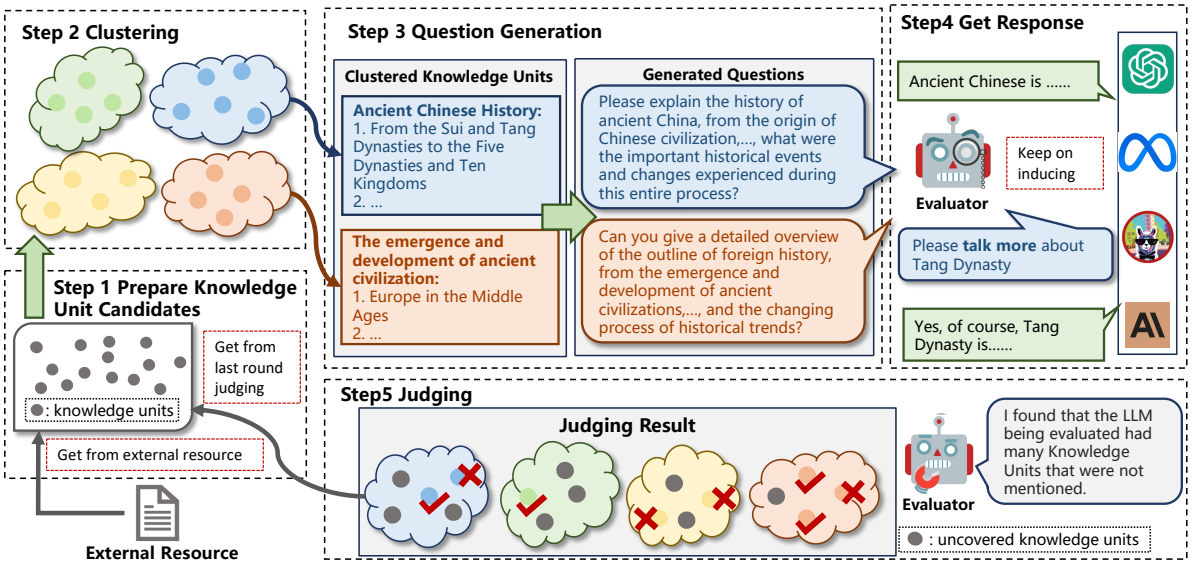


Figure 2: Ref-Eval consists of five main components that form a cycle of assessment.

knowledge units that were not covered in previous evaluation rounds.

Step 2 Clustering: Once the initial knowledge units are collected, the next step is to cluster these units. Clustering is performed to effectively organize knowledge units into coherent groups to facilitate the evaluation process. This organization is achieved by analyzing the semantic relationships between different knowledge units. Each knowledge unit, which may consist of facts, concepts, and information relevant to the related field, is transformed into an embedding representation. These units are then classified based on the proximity of their embeddings in semantic space, effectively creating subsets of knowledge that are thematically related.

Step 3 Question Generation: Ref-Eval leverages an LLM to generate questions for each cluster. This step utilizes the capabilities of the language model to pose questions that are not only relevant to the theme of the cluster but also aimed at probing the depth and accuracy of the presented knowledge. The questions are slightly adjusted during the evaluation based on the responses provided by the language model being evaluated to uncover any inconsistencies or gaps in the model's understanding of the given references. The questions are slightly adjusted during the evaluation due to the response made by the evaluated language model, which aims to uncover any inconsistencies or gaps in the language model's understanding of the given reference.

Step 4 Get Response: In this phase, the evaluator gets the response from the other evaluated targets. These responses are crucial as they represent the model's current understanding and knowledge. The Ref-Eval analyzes these responses to determine the model's proficiency and the accuracy of the information it provides. If responses are found missing the point of the question, Ref-Eval will keep on inducing to cover most of the knowledge units used to generate the question.

Step 5 Judging: The final step involves instructing the evaluation LLM to assess the extent to which the evaluated model's responses cover the specified knowledge units. In this phase, Ref-Eval labels the covered knowledge units as either correct or incorrect. For knowledge units that remain uncovered by the response of the evaluated model, Ref-Eval identifies and retains them for reorganization and re-evaluation, starting again from the initial step in subsequent assessment rounds.

4 Experiment Setup

4.1 Dataset and models

We experimented using text data sourced from various domains, including **Wiki** (Vrandečić and Krötzsch, 2014), **CodeGPT** (Xiaoxuan et al., 2023), **Legalbench** (Guha et al., 2024), and **PubMedQA** (Jin et al., 2019). We use the textual context of these datasets as a textual reference in evaluation. Details of the datasets are presented in Tab.1.

The models under evaluation include the widely used **GPT4-turbo** (GPT4), **GPT3.5-turbo** (GPT3.5), and **LLaMA2_chat_13b**

Dataset	Source	#Words	#Para.
Comp	CodeGPT[1]	2,371k	5,511
Legal.term	Legalbench[2]	233k	695
Wikitext	Wiki[3]	11,834k	4,396
Med.rand	PubMedQA[4]	877k	2,000
Med.sim	PubMedQA	351k	2,000

Table 1: Datasets information. **Comp** - Code-Text task of CodeGPT with **computer science** QA pairs. **Legal.term** - Definitions from Legalbench for **law interpretation**. **Wikitext** - High-quality Wikipedia articles, including various **world knowledge**. **Med.rand** and **Med.sim** - Randomized and similarity-based (details in A.2) subsets from PubMedQA, including knowledge of **biomedical**.

(LLaMA2). Additionally, we incorporated the **PMC_LLaMA_13b** (PMC), fine-tuned on the LLaMA2 architecture, specifically tailored for the biomedical field, to assess its performance within the related domain.

4.2 Baseline

In our evaluation, we compare Ref-eval against several baseline methods:

- **BLEU**, a probabilistic measure based on n-gram matching by comparing the overlap of n-grams between two sentences.
- **GPT Score**, an embedding-based indicator utilizing pre-trained language model embeddings to compare the semantic similarity between generated and reference texts.
- **LLM-as-Judge**, adopting the LLM-as-judge paradigm and relying on the evaluator (LLM) to judge response correctness.
- **Top-baseline**, assessing response correctness by aligning the answer provided by the model and the related textual context in the datasets **one-by-one**, which is superior to Ref-eval with more detail.

Furthermore, for comparative purposes, we introduce a **Human** baseline, and details are shown in A.3. Three annotators manually formulated questions and annotated model responses across datasets with the original text visible. Annotations were scored 0 for incorrect and 1 for correct answers, and average scores are detailed in the A.3.

[1] https://github.com/zxx000728/CodeGPT/
[2] https://hazyresearch.stanford.edu/legalbench/
[3] https://huggingface.co/datasets/wikitext
[4] https://pubmedqa.github.io

5 Performance of Ref-Eval

Tab.2 displays the performance of all baselines measured by correlation with human baselines. Tab.3 displays the resource consumption of 2 strong baselines of them, Top-baseline and Ref-eval. The conclusions are as follows:

LLM-based methods exhibit significantly higher evaluation accuracy compared to both n-gram-based and embedding-based approaches. The correlation with human baselines has notably improved, ranging from approximately 0.3 for BLEU to 0.6 for GPT Score and approaching nearly 1 for Top-baseline, LLM-as-Judge, and Ref-eval. This substantial enhancement underscores the superior efficacy of LLM-based techniques in assessing model responses in alignment with human expectations. The inferior performance of n-gram-based and embedding-based methods can be attributed to their limited understanding of the text's intrinsic meaning. In contrast, LLMs offer robust capabilities in comprehending textual context, thereby enhancing the effectiveness of LLM-as-Judge methodologies.

External references enhance evaluation accuracy and stability by countering internal model biases. Firstly, references improve the accuracy of evaluation. Both the Top-baseline and Ref-eval baselines demonstrate a higher correlation, utilizing textual references to augment the LLM's precision in assessing answer correctness. External references effectively counteract the potential influence of erroneous internal model knowledge, as previously investigated in the literature (Xie et al., 2023). Secondly, adopting external references leads to superior stability and robustness against LLM only. Notably, the Top-baseline and Ref-eval methods, exhibit significantly lower variances of 1.9e-5 and 2.5e-4, respectively. In contrast, the LLM-as-Judge baseline demonstrates a higher variance of 0.04, indicating greater susceptibility to internal model biases.

Ref-eval reduces the cost of evaluation and maintains competitive correlation with Top-baseline. Tab. 3 demonstrates the efficiency advantage of Ref-eval. In contrast to Top-baseline, which necessitates more frequent API calls and processes greater data volumes, thereby escalating operational expenses, Ref-eval optimizes these aspects by concentrating on crucial tokens during the Judgment and Question generation phases. This strategy

	GPT4	GPT3.5	LLaMA2
N-Gram Based			
BLEU	28.1	38.4	9.4
Embedding Based			
GPT Score	60.6	62.3	41.0
LLM-as-Judge Based			
LLM(ref)	**99.8**	98.9	**99.2**
NoRef-eval	88.9	98.4	86.1
Ref-eval	97.6	**99.5**	96.3

Table 2: Correlations between different methods and human baseline, where the bold font indicates the highest correlation.

curtails superfluous overhead and diminishes API call frequencies, thereby achieving cost reduction and efficient evaluation. Deep analysis in Sec.6.2 also shows that the cost of Ref-eval is also related to the model capacity, and the cost of the high capacity model is less than Top-baseline.

6 Deep Analysis on Ref-Eval

6.1 Evaluation Results

Models with larger parameters have better performance, demonstrating superior knowledge depth and accuracy across diverse datasets. The results depicted in Tab.4 indicate that GPT4-turbo not only outperforms the competing models in each dataset but also exemplifies exceptional prowess in domains that demand specialized expertise, particularly in the biomedical field. Its top accuracy underscores its robust adaptability and the impact of large-scale model architectures in handling complex queries and knowledge-intensive tasks. In contrast, although GPT3.5-Turbo performs well, its accuracy on the Legal.term and Wiki datasets is significantly lower than that of GPT4-Turbo.

General world knowledge provides challenges to current models. GPT4-turbo's performance on the General Knowledge Wiki dataset was the highest among the models evaluated, with an accuracy of 82.6%. However, GPT4-turbo's performance on general knowledge was the worst, by a wide margin, compared to other specialized datasets. This suggests that despite its overall dominance, there is room to enhance its capabilities in processing and understanding broad, general world knowledge. While GPT3.5-Turbo and LLaMA2-Chat show much lower performance on the Wiki dataset.

Model fine-tuned in specific areas can significantly improve model performance. PMC_LLaMA is a model designed specifically for the biomedical field, and its architecture and parameters may be optimized for processing medical text and terminology. As can be seen from the table, PMC_LLaMA achieved relatively high accuracy on the Med.rand and Med.sim datasets. This shows that the model has significant advantages in processing medical-related tasks and may benefit from its specific training and fine-tuning in the biomedical field.

6.2 Factors in Ref-eval

The more knowledge units reduce in each chatting round, the better the LLM's ability to answer relevant knowledge questions. Fig.3 shows the performance of LLM across multiple iterations of Ref-eval. The slope of these curves, denoted as the difference between the count of knowledge units in the $(n+1)$th round and the nth round, represents the amount of knowledge units that are reduced in each round. More powerful models like GPT4-turbo tend to answer more questions in each chatting round, especially at the outset and particularly on challenging datasets. While LLaMA2-Chat consistently exhibits slower speeds compared to GPT4-turbo and GPT3.5-turbo, which indicates the low capacity of LLaMA2-Chat for question answering compared to the other models.

Models with stronger capacity incur fewer token costs during the evaluation process in our method. Costs of GPT4-turbo and LLaMA2-chat in Tab.3 reflect that the cost of evaluating GPT4 on Ref-eval is 26% lower than that of LLaMA. This cost advantage shows in the stage of both Question Generstion and Response Judgement. As shown in Fig.3, the amount of knowledge units that are reduced in each round is different between models, and this also causes the difference in cost between models. Specifically, GPT4-turbo has a significant reduction in the number of knowledge units in each round, which decreases the repetition of questions and judgments, thereby reducing the cost of API calls.

6.3 Evaluation Performance of Knowledge and Knowledge Clustering

In this section, we evaluate the performance of Ref-eval by knowledge and knowledge clustering in Section 3, and provide the effectiveness of knowledge clustering in Ref-eval and the tendency of LLMs to respond to a knowledge cluster.

	Question Generation			Response Judgment			All Consumption		
	#Question	#API-Call	#Token	#Response	#API-Call	#Token	#API-Call	#Token	#Money (US Cent)
Ref-eval	256	0.18	89 + 16	1,622	1.1	342 + 38	1.3	431 + 54	0.59
LLM(ref)	1,423	1.0	216 + 241	1,423	1.0	308 + 31	2.0	524 + 272	1.34
Ref-eval(GPT4)	244	0.17	78 + 11	1,593	1.1	314 + 35	1.3	392 + 46	0.53
Ref-eval(LLaMA)	269	0.19	239 + 18	1,850	1.3	320 + 37	1.5	559 + 55	0.72

Table 3: The resource consumption of calling the API during the actual evaluation process of Ref-eval. The "#API-Call" indicator indicates the average number of API calls required for each knowledge point; the "#Token" indicator indicates the average number of tokens (input + output) required for each knowledge point; and the "#Money" indicator indicates the average cost of calling the API for each knowledge point (in cent).

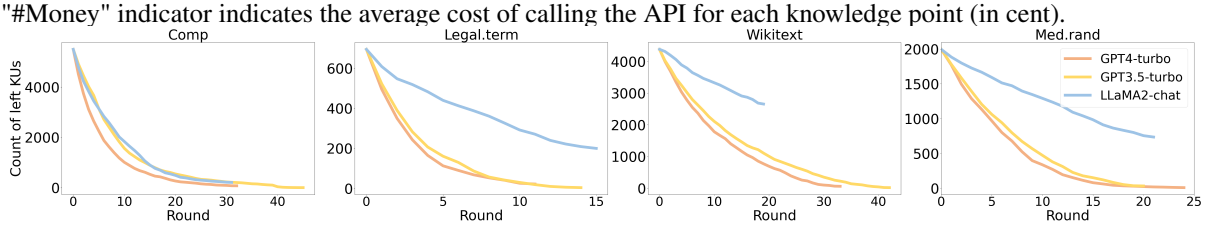


Figure 3: The number of unjudged knowledge points remaining after each iteration of different models on different data sets. The iteration ends when the representation on Ref-eval converges.

Model	GPT4	GPT3.5	LLaMA2	PMC
Comp	**95.4**	92.9	83.1	–
Legal.term	**94.6**	90.8	81.2	–
Wiki	**82.6**	54.5	42.9	–
Med.rand	**95.3**	82.1	73.2	78.5
Med.sim	**97.9**	82.3	72.5	75.7

Table 4: The performance of the model on each dataset, expressed as accuracy, where the bold font indicates the highest accuracy.

Cluster of knowledge with explicit meaning leads to the effectiveness of Ref-eval. As depicted in Figure 5a, the evaluated knowledge is categorized into 13 clusters, most of which exhibit clear and distinct boundaries. From round 30 (Figure 5a) to round 40 (Figure 5b), clusters such as *"Various films, TV shows, operas, and related media productions"* and *"Various STEM topics"* have been evaluated fully by Ref-eval. These clusters feature discrete and well-defined content descriptions that facilitate precise questioning and benefit model responses. In contrast, clusters with ambiguous boundaries, such as *"Various historical, biographical, electoral, media, educational, legal topics"*, present challenges in evaluation, because questions to these clusters with various topics cannot be pinpointed with the same level of clarity and specificity those with clear and distinct boundaries.

The cluster design in Ref-eval reduces the burden of generating questions from the same knowledge units in multiple iterations. Fig.6 illustrates the frequency distribution of repeated occurrences of knowledge units in questions across multiple rounds of iterations. Instances where knowledge units consist of questions but models fail to provide related answers are categorized as "ignored" knowledge units. The occurrence of high-frequency "ignored" units is significantly lower compared to those with lower frequencies. This means that cluster design in Ref-eval provides a low burden of questions generated from the same knowledge units in multiple iterations.

LLM tends to respond to Knowledge units with short and brief content. It is obvious in Fig. 6 that high-frequency "ignored" knowledge units tend to contain extensive information as they often represent multiple entities or lists. This contributes to their high frequency of being ignored, particularly when they are contained in a cluster with fewer informational knowledge units. This can be caused by the characteristics of LLMs, as they are more adept at giving a comprehensive answer, and the less informational knowledge units can be answered comprehensively with less text.

6.4 Evaluation Performance of Different Types of Questions

Fig.4 displays six types of questions with different interrogatives, each yielding varied outcomes in LLM responses.

The complexity of specific knowledge leads to poor performance in answering "who" and "where" questions. It is obvious that all models consistently performed poorly when tasked with answering "who" questions. This difficulty likely stems from the fact that "who" questions often involve specific individuals, historical contexts, or

Comparison of Ref-eval and LLM-as-Judge		
LLM-as-Judge	**Question:** When was Marino railway station in Adelaide opened? **Answer:** Marino Railway Station was opened on October 19, 1885. **Judgement:** Correct× **Reason:** The candidate's answer correctly identifies the location of Marino Railway Station and its opening date.	**Text:** Marino railway station is on the Seaford railway line , in the southern Adelaide suburb of Marino , South Australia Once the terminus of the line that now runs south to Seaford . **Marino railway station opened in 1915 when the line was built from Adelaide to Marino.** The line was extended further south to its current terminus at Seaford in 2014 . The line further south towards Hallett Cove follows the rugged coastline on the line 's western side.
Ref-eval	**Question:** What is the common trait of Marino railway station in Adelaide, . . . , and what years were they opened? **Answer:** Marino Railway Station is located in Adelaide, Australia. It was opened on October 19, 1885, and it is still operational. It serves the Seaford and Tonsley lines of Adelaide's metropolitan train network. **Judgement:** Incorrect✓ **Reason:** The candidate incorrectly states that Marino Railway Station was opened on October 19, 1885, while the reference text indicates it opened in 1915.	
Performance of GPT3.5-Turbo		
Top-baseline	**Question:** In legal terms, what does the term 'damage' specifically refer to? **Answer:** In legal terms, the term 'damage' refers to the harm or injury caused to a person, property, or reputation as a result of the wrongful act of another. . . **Judgement:** Incorrect	**Text:** The word 'damage' traditionally describes a harm to property (hence, 'property damage'), rather than harm to the person (usually referred to as 'personal injury').
Ref-eval	**Question:** In area of law, . . . , and what is the distinction between 'damage' and 'injury'? **Answer:** The distinction between 'damage' and 'injury' is that damage refers to harm caused to property, while injury refers to the victim by certain injury. **Judgement:** Correct	
Human	**Question:** In legal terminology, what kind of object does 'damage' apply to ? **Answer:** "Damage" usually applies to damage to objects. This includes loss or damage to property, land, equipment, etc. . . **Judgement:** Correct	

Table 5: Comparison of Ref-eval and LLM-as-Judge at the top of the table, and performance of GPT3.5-Turbo of different baselines at the bottom of the table.

cultural knowledge, which can be challenging for models to learn accurately. Similarly, performance was also notably lower for questions related to location ("where"). Models may struggle with disambiguation between different individuals sharing the same name, and they may not effectively handle nuanced information about various people. This challenge could be attributed to the diverse contexts and geographical specifics that "where" questions often entail.

Ref-eval questioning is helpful for LLM to answer reasoning-based questions. Converse to bad performance on questions of "who" and "where", questions involving "how" typically require explanations of reasoning or logic, and all models demonstrated relatively strong performance in this category. Furthermore, questions focusing on "why" also yielded better model performance compared to other types of questions. Ref-eval adds some additional information to each knowledge point in the process of asking questions, and as a result, the model's reasoning ability may improve, and the answer to such questions will become better.

6.5 Case Study

In Tab.5, we show cases of Ref-eval during the evaluation of GPT3.5-turbo and GPT4-turbo. The performances of different models show that:

Ref-eval is more accurate than LLM-as-judge. The top case in Tab.2 shows the cooperation of Ref-eval and LLM-as-Judge, where Ref-eval supplies a reference as the ground truth, and LLM-as-Judge evaluates answers based on the knowledge contained within the LLM. The answer about the opening time of Marino Railway Station is actually "Incorrect" by Wiki. However, LLM-as-Judge provides a "Correct" judgment based on incorrect knowledge in LLM. Ref-eval, in contrast, gives the real judgment of "Incorrect" and provides a credible reason from reference.

Ref-eval improve the performance of LLMs with better in-context learning (Min et al., 2022) ability but poor knowledge memorizing. Tab.2 demonstrates that Top-baseline shows better performance compared to Ref-eval with models like GPT4-turbo and LLaMa2-chat. However, Ref-eval's performance becomes comparable or even superior with models such as GPT3.5-turbo. This shift is primarily due to Ref-eval's capability to

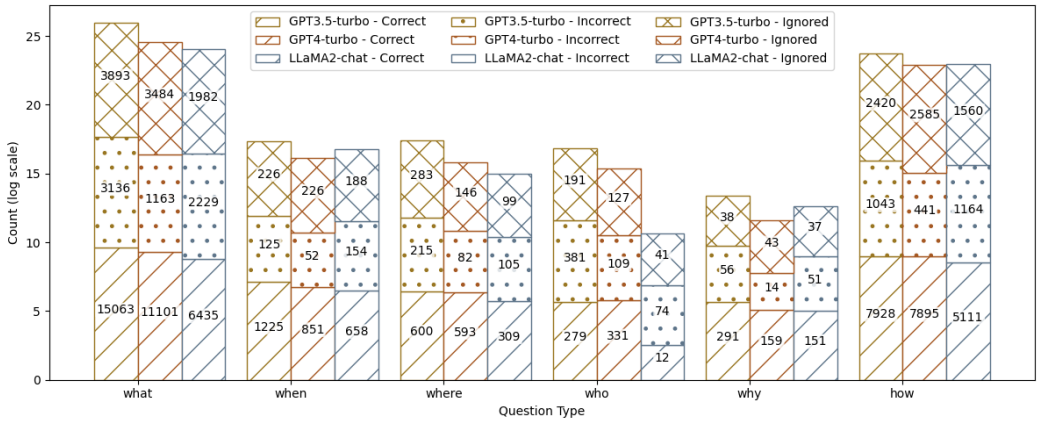
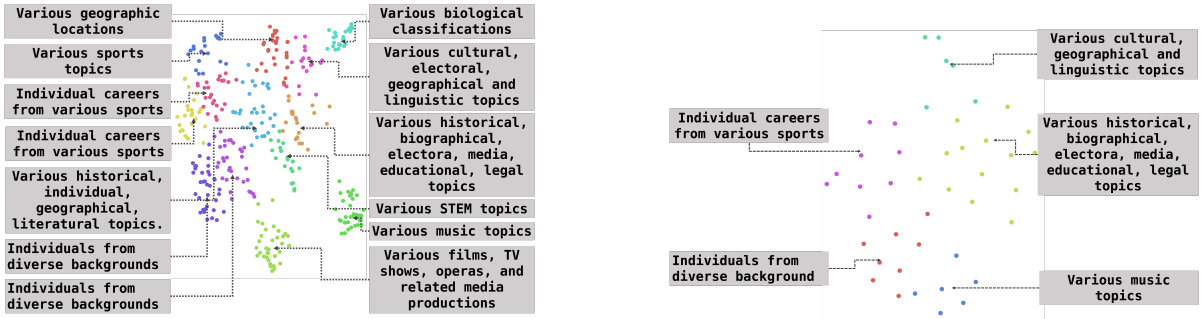


Figure 4: The counts of 6 types of questions of Ref-eval used for different models.



(a) k=30 (b) k=40

Figure 5: The embedding map of the remaining knowledge units to be evaluated on the Wiki dataset in the process of evaluating GPT3.5-turbo at round **k** (k is different of the two sub-figures), where the representation of knowledge units of the same color is clustered in the same class, and their class names are displayed.

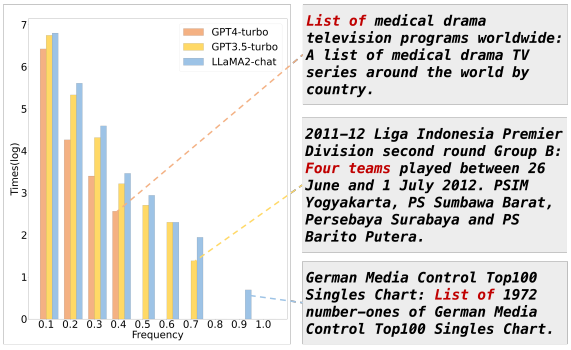


Figure 6: The frequency of knowledge units in multiple iterations of Ref-eval. The X-axis is the frequency of knowledge units, and the vertical axis is the number of knowledge units of a certain frequency.

incorporate richer contextual information derived from related knowledge sources, which facilitates more detailed and accurate question formulation. For instance, in Tab.5, we analyze the term *'damage'*. Ref-eval enriches GPT-3.5-turbo's grasp of 'damage' with contextual clues on 'injury', improving response precision. Top-baseline can struggle without such context, highlighting the need for comprehensive evaluation methods. Notably, we design the prompt of question generation to avoid leaking answers in A.1.

7 Conclusion

In this paper, we introduce the Reference-based LLM-as-Evaluator (Ref-Eval) framework, a novel method for evaluating large language models across diverse domains. Leveraging LLMs' ability to understand and respond to complex queries, Ref-Eval addresses the limitations of traditional evaluation methods. Our experiments on various datasets show that Ref-Eval achieves high consistency with human evaluation, highlighting its effectiveness in assessing model responses. This approach not only offers a scalable and efficient means to evaluate LLMs but also advances the field of model evaluation in knowledge-intensive tasks. Our findings pave the way for the adoption of LLM-based evaluation frameworks and point toward a promising future for language model evaluation.

8 Limitation

While the proposed Reference-based LLM-as-Evaluator framework (Ref-Eval) offers significant advancements in evaluating text generated by language models, it is not without limitations. Ref-Eval's effectiveness heavily depends on the quality and relevance of the external reference materials used. If these references are incomplete or outdated, the framework's evaluations may be compromised. The challenge of integrating large volumes of reference data remains, as even with synthesized knowledge units, the risk of overlooking critical details or context persists. The sensitivity of LLMs to prompt variations can also result in inconsistent evaluation outcomes when dealing with diverse or ambiguously phrased questions. This variability in model responses may affect the reliability of the evaluation results, particularly in scenarios where nuanced understanding is crucial.

9 Ethical Concerns

The Reference-based LLM-as-Evaluator framework (Ref-Eval) introduces several ethical concerns. The external reference data used may include sensitive or controversial content, which could lead to the perpetuation of biases or misinformation. Additionally, handling proprietary or personal information raises privacy and intellectual property concerns.

To address these issues, we implement strict protocols to vet reference data for sensitivity and relevance. We ensure transparency in our data curation process and prioritize ethical standards to safeguard privacy and prevent misuse, balancing the benefits of comprehensive evaluation with responsible data handling.

References

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

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.

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*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, et al. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18099–18107.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Xiang Li, Yunshi Lan, and Chao Yang. 2024. Treeeval: Benchmark-free evaluation of large language models through tree planning. *arXiv preprint arXiv:2402.13125*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

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.

Tempest A van Schaik and Brittany Pugh. 2024. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2832–2836.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Zhu Xiaoxuan, Xiong Zhuozhi, Zhang Lin, Ye Haoning, Gu Zhouhong, Li Zihan, Jiang Sihang, Feng Hongwei, Xiao Yanghua, Wang Zili, Yang Dongjie, and Wang Shusen. 2023. Codegpt: A code-related dialogue dataset generated by gpt and for gpt. https://github.com/zxx000728/CodeGPT.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Appendix

A.1 Prompt

In Fig.7, Fig.8, and Fig.9, we illustrate three distinct prompts, each designed to complete different tasks within Ref-eval. Fig.7 represents a question generation prompt in step 3 of **Question Generation**. Fig.8 represents a question-answering prompt in step 4 of **Get Response**. Fig.9 represents a response judgment prompt in step 5 of **Judging**.

Each of them includes an instruction, an output format, a notation, and inputs.

A.2 Dataset

Med.rand consists of randomly chosen questions from PubMedQA. **Med.sim**, on the other hand, is a selection from PubMedQA based on the similarity to the paragraph: *"A medical history of arterial hypertension was associated with lower MMSE scores and a higher prevalence of dementia and cognitive decline at baseline. However, intact cognition through the observation period was linked to higher baseline SBP."* The similarity is determined by comparing the embeddings of this paragraph with those of all other paragraphs in PubMedQA.

A.3 Human Baseline

A.3.1 Human Baseline Description

Annotator Selection We selected three annotators with expertise in the relevant fields to ensure the quality of the questions and annotations. All annotators had prior experience in data annotation and a good understanding of the subject matter.

Question Formulation The annotators were instructed to manually formulate questions based on the original text provided in the datasets. They were asked to create questions that would test the comprehension and response-generation capabilities of the models.

Annotation Process The annotators annotated the model responses while having access to the original text. This approach allowed them to assess the accuracy of the model's answers in the context of the given information.

A.3.2 Annotation Scoring

Scoring Criteria Annotations were scored on a binary scale: 0 for incorrect answers and 1 for correct answers. An answer was considered correct if it accurately addressed the question based on the information provided in the original text.

Scores The average scores for each dataset and model are presented in Tab.6:

Inter-Annotator Agreement To ensure the reliability of the annotations, we calculated the inter-annotator agreement using the Fleiss' kappa coefficient. The kappa value was found to be 0.72, indicating substantial agreement among the annotators.

Model	Score 1	Score 2	Score 3	Score 4	Score 5
GPT-4	0.9566	0.9479	0.8269	0.9190	0.9230
GPT-3.5	0.9245	0.9056	0.5094	0.7735	0.8113
Llama2	0.8219	0.8062	0.4002	0.6295	0.6352
MedLlama	0.6367	0.6241	-	-	-

Table 6: Human scores for different models

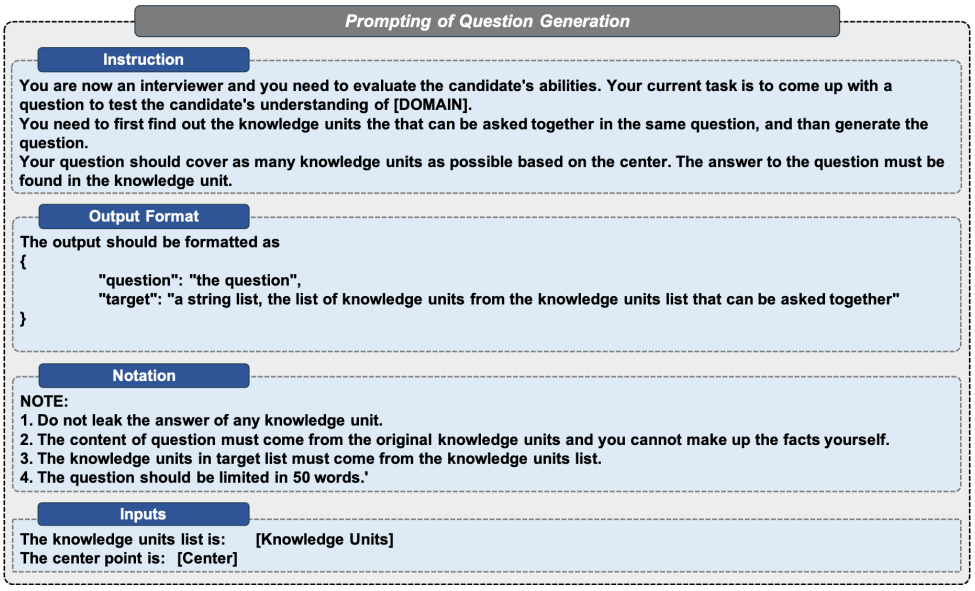


Figure 7: Prompt for question generation.

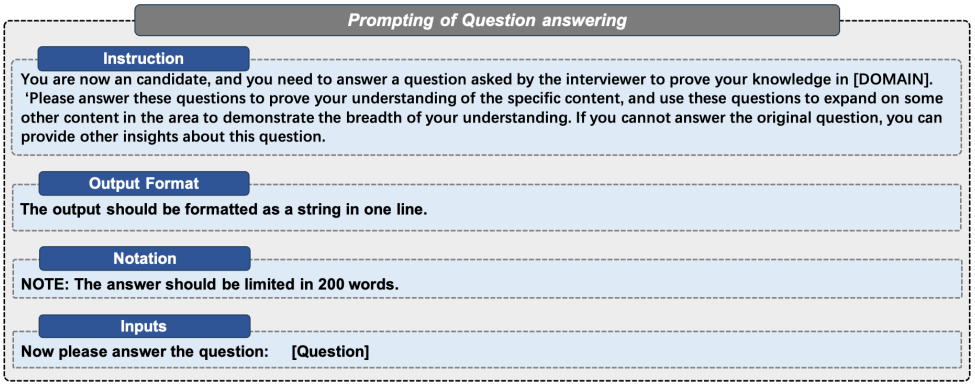


Figure 8: Prompt for question answering.

Prompting of Judging

Instruction

You are now an interviewer and you need to evaluate the candidate's understanding of the domain: [DOMAIN].
Your current task is to judge whether the answer from the candidate is correct or related according to the reference text.
There are 3 types:
”Unrelated" means "The candidate's answer is completely irrelevant to the question and the reference text, and does not provide any relevant information.”
”Correct" means "The candidate's answer partially or completely aligns with the reference text. The candidate's answer may contain some omissions, incompleteness or even some details, facts or dimensions that are not mentioned in the reference text, but overall demonstrates the candidate's understanding of the knowledge and their ability to accurately apply and explain relevant concepts, facts, or principles.”
”Incorrect" means "The candidate admits that he/she does not know the answer" or "The candidate's answer contradicts, is completely incorrect compared to the reference text. It may contain factual errors, misunderstandings, or incorrect inferences that are inconsistent with the reference text. Pay attention that never make the mistake of thinking that facts you don’t know are incorrect! "
The candidate's answer may provide some details, facts or dimensions that are not mentioned in the reference text: You can just assume these details or facts are correct. Because the reference cannot not provide exact factual support to prove that these details or facts are indeed incorrect. However, if all dimensions or details provided by candidate are not mentioned in the reference text, the output type should be "unrelated”.

Output Format

The output should be formatted as
{
"type": "unrelated, incorrect or correct",
"reason": "the reason why you give the type, if the type is incorrect, please point out the exact error from candidate\'s answer and the exact correct answer from the reference text.”
}

Notation

NOTE:
1. Please output according to the output format in one line.
2. The reason should be limited in 100 words.

Inputs

The answer from the candidate is: [Answer]
The reference text is: [Text]

Figure 9: Prompt for response judgment.