QUILL: Quotation Generation Enhancement of Large Language Models

Anonymous ACL submission

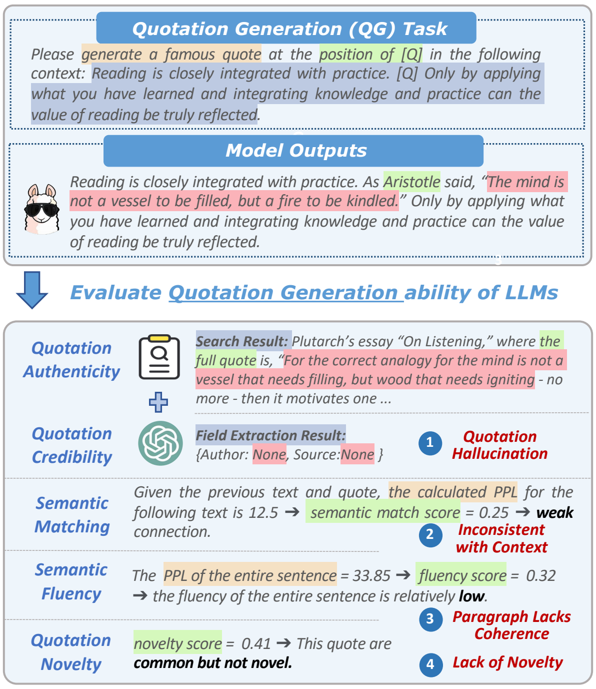
Abstract

While Large language models (LLMs) have become excellent writing assistants, they still struggle with quotation generation. This is because they either hallucinate when providing factual quotations or fail to provide quotes that exceed human expectations. To bridge the gap, we systematically study how to evaluate and improve LLMs' performance in quotation generation tasks. We first establish a holistic and automatic evaluation system for quotation generation task, which consists of five criteria each with corresponding automatic metric. To improve the LLMs' quotation generation abilities, we construct a bilingual knowledge base that is broad in scope and rich in dimensions, containing up to 32,022 quotes. Moreover, guided by our critiria, we further design a quotation-specific metric to rerank the retrieved quotations from the knowledge base. Extensive experiments show that our metrics strongly correlate with human preferences. Existing LLMs struggle to generate desired quotes, but our quotation knowledge base and reranking metric help narrow this gap. Our dataset and code will be released soon.

1 Introduction

Famous quotations (Tan et al., 2015a) are vital in academic and everyday communication. They lend authority to arguments and enhance persuasiveness, as they often stem from historically influential figures whose ideas have endured. Additionally, these quotations elevate the literary and artistic quality of a text, making discussions more engaging. They also facilitate comprehension of complex concepts, enabling readers to grasp ideas efficiently through concise expressions (Vaswani et al., 2023).

The task of Quotation Generation (QG) seeks to produce suitable quotations to deepen the context in large language models (LLMs) (Anil et al., 2023; Achiam et al., 2023; Touvron et al., 2023). However, LLMs encounter significant challenges in this domain, as illustrated in Fig.1. Primarily, the generated quotes frequently fail to correspond to genuine famous quotations and are often inaccurately attributed, a phenomenon termed "Quotation Hallucintion." (Huang et al., 2023; Bang et al., 2023; Guerreiro et al., 2023) Additionally, these quotes don't align with the contextual meaning, resulting in a lack of coherence within the paragraph. Furthermore, LLMs exhibit a tendency to reproduce well-known quotes, which diminishes novelty and restricts creative expression.


Quotation Generation (QG) Task
Model Outputs
Evaluate Quotation Generation ability of LLMs
Quotation Authenticity
Quotation Credibility
Semantic Matching
Semantic Fluency
Quotation Novelty
1 Quotation Hallucination
2 Inconsistent with Context
3 Paragraph Lacks Coherence
4 Lack of Novelty

Please generate a famous quote at the position of [Q] in the following context: Reading is closely integrated with practice. [Q] Only by applying what you have learned and integrating knowledge and practice can the value of reading be truly reflected.

Reading is closely integrated with practice. As Aristotle said, "The mind is not a vessel to be filled, but a fire to be kindled." Only by applying what you have learned and integrating knowledge and practice can the value of reading be truly reflected.

Search Result: Plutarch's essay "On Listening," where the full quote is, "For the correct analogy for the mind is not a vessel that needs filling, but wood that needs igniting - no more - then it motivates one ...

Field Extraction Result: {Author: None, Source:None }

Given the previous text and quote, the calculated PPL for the following text is 12.5 → semantic match score = 0.25 → **weak** connection.

The PPL of the entire sentence = 33.85 → fluency score = 0.32 → the fluency of the entire sentence is relatively **low**.

novelty score = 0.41 → This quote are **common but not novel.**

Figure 1: An example of prevalent issues in Quotation Generation (QR) by LLMs. In QR tasks, LLMs often fabricate sentences, leading to quotation hallucination. Additionally, the generated quotes may not align with the context, resulting in contextual inconsistency and semantic incoherence. Finally, the sentences produced by LLMs tend to be overly common, resulting in a lack of novelty in quotations.

Although the issues of QG task are particularly

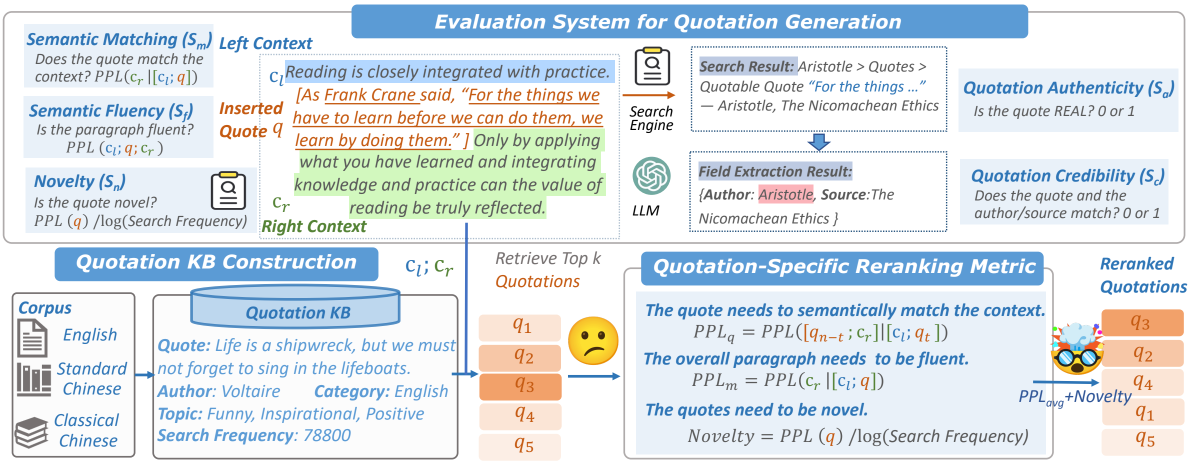

Evaluation System for Quotation Generation
Semantic Matching (S_m)
Does the quote match the context? $PPL(c_r \| [c_l; q])$
Semantic Fluency (S_f)
Is the paragraph fluent?
$PPL\ (c_l; q; c_r\ )$
Novelty (S_n)
Is the quote novel?
$PPL\ (q)\ /\log(\text{Search Frequency})$
Left Context
c_l
Reading is closely integrated with practice.
Inserted Quote q
[As Frank Crane said, "For the things we have to learn before we can do them, we learn by doing them."]
Only by applying what you have learned and integrating knowledge and practice can the value of reading be truly reflected.
c_r
Right Context
Search Engine
LLM
Search Result: Aristotle > Quotes > Quotable Quote "For the things …" — Aristotle, The Nicomachean Ethics
Field Extraction Result:
{Author: Aristotle, Source:The Nicomachean Ethics }
Quotation Authenticity (S_a)
Is the quote REAL? 0 or 1
Quotation Credibility (S_c)
Does the quote and the author/source match? 0 or 1
Quotation KB Construction
$c_l; c_r$
Corpus
English
Standard Chinese
Classical Chinese
Quotation KB
Quote: Life is a shipwreck, but we must not forget to sing in the lifeboats.
Author: Voltaire
Category: English
Topic: Funny, Inspirational, Positive
Search Frequency: 78800
Retrieve Top k Quotations
q_1
q_2
q_3
q_4
q_5
Quotation-Specific Reranking Metric
The quote needs to semantically match the context.
$PPL_q = PPL([q_{n-t}\,; c_r]\|[c_l; q_t\,])$
The overall paragraph needs to be fluent.
$PPL_m = PPL(c_r \|[c_l; q])$
The quotes need to be novel.
$Novelty = PPL\ (q)\ /\log(\text{Search Frequency})$
PPL_{avg}+Novelty
Reranked Quotations
q_3
q_2
q_4
q_1
q_5

Figure 2: The framework for our Quotation Generation (QG) task research. We first establish an evaluation system with 5 evaluation criteria and automatic metrics, then build a quotation knowledge base covering multiple languages, topics and eras, and finally propose a quotation-specific reranking metric to rerank the quotations recalled in the RAG stage and improve the performance of QG tasks.

problematic in LLMs, there is currently no effective solutions. Previous studies (Qi et al., 2022a) were based on representative pre-trained language models such as BERT (Devlin et al., 2019), and it remains under-explored on the problem of quotation hallucination with LLMs. And there is currently no systematic and comprehensive benchmark to evaluate the quotation generation ability of LLMs.

To tackle these challenges, we introduce QUILL for **QU**otation Generat**I**on enhancement of **L**arge **L**anguage Models, a framework integrating an automatic evaluation system and an innovative and effective solution to improve quotation generation performance of LLMs.The framework of QUILL is shown in Fig. 2. QUILL presents a comprehensive benchmark comprising 7 quotation domains and 16 real-world scenarios to evaluate large models' quotation generation abilities systematically, which consists of 5 highly interpretable and rigorous criteria with automatic evaluation metrics (Fig. 1): (1) ***Quotation Authenticity***: Confirm whether the quoted quotes are real quotes from famous people to prevent misquotations or fabrications. (2) ***Quotation Credibility***: Verify whether the quotation satisfies the author or source mentioned in the context (if any) to ensure the credibility of the quoted content. (3) ***Semantic Matching***: Evaluate whether the semantics of the quoted quote align with the context. (4) ***Semantic Fluency***: Evaluate the extent to which the cited quotation affects the fluency of the paragraph. (5) ***Quotation Novelty***: Evaluate the degree of uniqueness of the quote.

Additionally, based on the task's essential characteristics, we introduce an innovative Quotation-Specific Reranking Metric (Karpukhin et al., 2020; Lewis et al., 2021; Chern et al., 2023) to improve model performance in QG tasks. To facilitate the task, we also established a comprehensive and high-quality knowledge database containing up to 32,022 quotes. This database spans both Chinese and English languages, various authors, different eras, and diverse topics, which ensures the wide applicability and generalization of our method. To the best of our knowledge, our work is the first systematic investigation into the automatic evaluation and enhancement of QG performance in LLMs. To summarize, our contributions are mainly four-fold:

1. We establish a holistic and automatic evaluation system for the quotation generation task, consisting of five highly interpretable and rigorous criteria, facilitating both human and automatic evaluation of this task.

2. We construct a comprehensive and high-quality knowledge database containing up to 32,022 quotes, complete with authors or sources.

3. We design a fine-grained quotation-specific metric to rerank the retrieved quotations from the knowledge base.

4. We conduct extensive experiments to verify that our metrics are strongly correlate with human preference and significantly effective in both open-source and closed-source LLMs.

2 Related Work

2.1 Quotation

Previous research on quotations mainly focused on Quote Recommendation (QR) (Tan et al., 2015a). QR task was initially proposed by (Tan et al., 2015a). They proposed a learning ranking framework for the task which integrates 16 features crafted manually. (Wang et al., 2020) utilized an encoder-decoder framework to generate speech responses based on a separate modeling of the history of the dialogue and the current query. (Wang et al., 2021) used semantic matching to encode multi-round dialogue histories using Transformer (Vaswani et al., 2023) and GRU (Cho et al., 2014). However, previous studies overlook the quotation generation capabilities of large models and lack a comprehensive evaluation system or benchmark for assessing performance in quoting famous lines.

2.2 Hallucination

In NLP, hallucinations refer to generated content that is meaningless or misaligned with the source (Filippova, 2020; Maynez et al., 2020). To address this, two main approaches have been proposed: (1) preventing hallucinations during training and generation, and (2) reducing them post-generation. (Manakul et al., 2023) classified methods into black box (no external resources used) and gray box (external resources for validation). Other techniques for alleviating hallucinations include reranking generated sample responses (Dale et al., 2022) and improving beam search (Sridhar and Visser, 2023). Recent mitigation technologies have also shown promise in reducing hallucinations (Mündler et al., 2024; Pfeiffer et al., 2023; Chen et al., 2023; Zhang et al., 2024; Agrawal et al., 2024). Although these methods have partially addressed hallucinations, they have not fully solved the issue, especially for factual quotations and famous quotes.

3 Background

3.1 Task Formulation

Quotation Generation Given a plain text $c = [t_1, t_2, \ldots, t_i, \ldots, t_n]$, the goal of the *Quotation Generation (QG)* task is to generate quotes for the specified insertion point i. The left and right contexts, c_l and c_r, are defined as $c_l = [t_1, t_2, \ldots, t_i]$ and $c_r = [t_{i+1}, \ldots, t_n]$, respectively. In our work, we mainly focus on the ability of the model in quotation generation tasks.

3.2 Preliminaries

Perplexity (PPL) is a crucial metric in natural language processing, reflecting a model's predictive capability on text data and indicating the certainty of its next word prediction. Lower perplexity signifies greater confidence in the model's predictions, demonstrating a stronger ability to generate or understand language. PPL of a language model given a sequence of words $w_1, w_2, \ldots, w_N$ is defined as:

$$PPL\left(P_r \mid P_l\right) = \exp\left(-\frac{1}{s}\sum_{i=t+1}^{N} \log P(w_i \mid w_1, \ldots, w_{i-1})\right) \quad (1)$$

where P_l is the given left paragraph, P_r is the following context needs to be calculated, $P(w_i \mid w_1, w_2, \ldots, w_{i-1})$ is the probability of the word w_i given its left context, and s is equal to $N - t + 1$, which is the length of the following paragraph.

4 Evaluation System for QG

The accuracy and rationality of quoting famous quotes are crucial, as they directly affect the credibility and rigor of the content. Therefore, we establish a holistic and automatic evaluation system for QG task evaluation in LLMs, containing five criteria and further design automatic metrics for each criterion (Fig. 1).

4.1 Criteria

Considering the nature of the quotation task itself, we design the following five criteria: (1) ***Quotation Authenticity***: Confirm whether the quoted quotes are real quotes from famous people to prevent misquotations or fabrications. (2) ***Quotation Credibility***: Verify whether the quotation satisfies the author or source mentioned in the context (if any) to ensure the credibility of the quoted content. (3) ***Semantic Matching***: Evaluate whether the semantics of the quoted quote align with the context. (4) ***Semantic Fluency***: Evaluate whether the quoted quote affects the fluency of the original text. (5) ***Quotation Novelty***: Evaluate the degree of uniqueness of the quote.

4.2 Evaluation Metrics

Quotation Authenticity. Authenticity of quotations is crucial as it ensures the reliability and credibility of information (Kington et al., 2021). To verify the authenticity of the quoted celebrity quotes, we first search the quotation database for the information corresponding to the quote. If the database contains the information, we use the corresponding information to make a judgment. If not, we use different search engines (such as Google Scholar[1] and Baidu Scholar[2]) to recall the corresponding search results. Previous studies (Han et al., 2024) have shown that GPT-4o (OpenAI, 2022) has excellent simple extraction capabilities, and the extraction task based on this study only has two fields, author and source. Therefore, we use GPT-4o to extract the corresponding field information, and then compare the results of different search engines. If the field information is different, manual comparison is required. For extraction details and validity of GPT-4o, please refer to Appendix B. Finally, based on the extracted information, we verify whether the quote genuinely originates from the specific celebrity. The final score is defined as follows:

$$S_a = \begin{cases} 1, & \text{if quote is real} \\ 0, & \text{if not real} \end{cases} \tag{2}$$

Quotation Credibility. Generally speaking, in the context of quoting, the source of the quote will be mentioned, such as the author, classic literature, or other sources. Ensuring consistency between the citation and the mentioned author or source is crucial for maintaining contextual coherence and information accuracy (Rami Aly, 2024). In order to confirm whether the citation meets the source restriction mentioned in the context,

we also use GPT-4o to extract citation restrictions in the context and then compares it with the extraction result of the previous indicator. If the source matches, the citation is marked as trustworthy. The final score is defined as follows:

$$S_c = \begin{cases} 1, & \text{if citation meets the context source restriction} \\ 0, & \text{if not} \end{cases} \tag{3}$$

Semantic Matching. Improper quotation may lead to misunderstandings or misinterpretations of the original meaning, thereby weakening the effectiveness and persuasiveness of the argument (Quora, 2020). Perplexity is a common metric in NLP, used to assess a language model's predictive capability for text. Hence, we calculate the PPL of subsequent text based on a given prior text and quotation to evaluate the consistency between the quotation and its context. If the evaluation score is low, it implies that the citation aligns well with the following context in terms of semantics; otherwise, the rationality of the citation should be reconsidered. The formula is as follows:

$$PPL_m = PPL\left(c_r \mid [c_l; q]\right) \tag{4}$$

where c_l stands for the previous text, q stands for the quotation, and c_r stands for the following text.

To simplify computation, we normalize the PPL values to a range between 0 and 1. Given that a lower PPL indicates a higher degree of semantic alignment, we utilize an inverted Sigmoid function. The final calculation formula is as follows:

$$S_m = \frac{1}{1 + e^{k_m(PPL_m - \mu_m)}} * 100\% \tag{5}$$

where μ_m represents the mean value of PPL_m, which is 35.243, and k_m is determined using an empirical formula, yielding a value of 0.053. See the Appendix C for the specific calculation details.

Semantic Fluency. After quotation, it is necessary to ensure that the entire context is fluent and coherent to maintain semantic consistency and logical integrity (Krumm et al., 2020). This study calculates the PPL of the entire context to measure the textual fluency of the overall context after inserting quotations. Lower perplexity indicates smoother overall contextual semantics. The calculation formula for semantic fluency is as follows:

$$PPL_f = PPL_q\left([c_l, q, c_r] \mid \cdot\right) \tag{6}$$

where c_l stands for the previous text, q stands for the quotation, and c_r stands for the following text.

Similarly, for normalizing the PPL values into a range from 0 to 1, the final score for semantic fluency is designed as follows:

$$S_f = \frac{1}{1 + e^{k_f(PPL_f - \mu_f)}} * 100\% \tag{7}$$

where μ_f represents the mean value of PPL_f, which is 16.470, and k_f is determined using an empirical formula, yielding a value of 0.500.

Quotation Novelty. Integrating novel quotations into established ideas enhances originality and personalizes the expression within academic discourse (Savov, 2021). To evaluate the extent to which the quote introduces new ideas or unique perspectives to the original context, we utilize the Bing[3]search engine to determine the number of Search Frequency corresponding to each quotation, applying a log10 transformation to quantify quotation popularity. In addition, to mitigate potential biases in search results, we also incorporate

[1] https://scholar.google.com/

[2] https://xueshu.baidu.com/

[3] https://www.bing.com/

the quoted PPL value for supplementation. As a lower PPL indicates a higher frequency of text occurrence, it is inversely correlated with search frequency. Therefore, the formula is as follows:

$$\text{Novelty} = \frac{PPL(q \mid \cdot)}{log_{10}(\text{Search Frequency})} \tag{8}$$

where Search Frequency indicates the number of search results obtained by searching the quotation in the Bing search engine. In order to map the PPL value to a range of 0 to 1, since higher novelty means higher score, the positive sigmoid function is adopted here, and the final score is as follows:

$$S_n = \frac{1}{1 + e^{-k_n(Novelty - \mu_n)}} * 100\% \tag{9}$$

where μ_n represents the mean value of Novelty, which is 10.67, and k_n is determined using an empirical formula, yielding a value of 0.253.

5 Quotation Knowledge Base

5.1 Dataset Construction

In order to alleviate the problem of quote hallucination in LLMs, we develop a comprehensive bilingual and multi-topic quotation corpus designed to enhance retrieval quotation tasks during the RAG stage. This corpus is structured into three distinct components: the English, the Standard Chinese, and the Classical Chinese. To improve the application scope and practical value of the corpus, we ensure comprehensive coverage of both common and specialized fields and also implement stringent quality control measures. Finally, we obtain a higher-quality corpus with 32,022 entries. The statistics of our knowledge dataset are shown in the Tab. 1. Details regarding the data construction for the English, Standard Chinese, and Classical Chinese corpora are provided in the Appendix A.

Category	Samples	AvgLen	AvgSearchFreq	AvgNovelty
English	16,393	16	2,823,499	6.8
Standard Chinese	7,519	42	150,011	6.3
Classical Chinese	8,110	14	19,017	5.0
Total	32,022	24	997,509	6.0

Table 1: The statistics of our knowledge base. The *AvgLen*, *AvgSearchFreq* and *AvgNovelty* denote the average of the quote length, the frequency of Bing Search engine and the value of Quotation Novlety respectively.

5.2 Quality Assessment by Human

After constructing the dataset, we manually check its quality. For each component, we randomly select 100 quotes and engage three annotators to verify their validity. The annotators use search engines [4] to locate references and evaluate both the authenticity of the quotes and the accuracy of their attributed authors and sources. Only quotes that satisfy both criteria are included in the final dataset. The final results are determined through a majority voting process. In the English, Standard Chinese and Classical Chinese components, 99, 97 and 98 quotations respectively met the established criteria. These results confirm the high quality of the dataset, which is derived from trustworthy sources such as published books and reputable citation websites.

5.3 Dataset Statistics

In this part, we compare the statistics of our dataset with existing quotation-related resources, as shown in Tab.2. In contrast, our dataset is the first to consider quotation novelty, covering a broad range of topics and authors while also recording and annotating their sources. Additionally, we have expanded the scale of the quotation dataset, thereby broadening its application scenarios and significance.

Resource	Size	Topic	Author	Multilingual	Novelty
LRQW (Tan et al., 2015b)	3,158	822	762	N	N
QRDW (Ahn et al., 2016)	1,200	-	-	N	N
QuoteR (Qi et al., 2022b)	13,550	-	-	Y	N
Ours	32,022	2,301	9,708	Y	Y

Table 2: The statistics of our dataset with existing quotation-related resources. Multilingual refers to the inclusion of two or more languages, Y denotes Yes, and N denotes No.

6 Quotation-specific Reranking Metric

We propose a fine-grained, end-to-end RAG framework for quotation generation (QG), introducing a quotation-specific rerank metric to improve selection. While semantic similarity only recalls the most semantic relevant top-k quotes, QG demands fluency, context alignment, and novel citations. To address this, we define three evaluative sub-indicators (Fig. 2).

Quotation Matching Quotation matching emphasizes the completion of the quotation itself and its alignment with the subsequent text (MacLaughlin and Smith, 2021). This is accomplished by calculating the PPL of the remaining portion of

[4] https://www.bing.com/

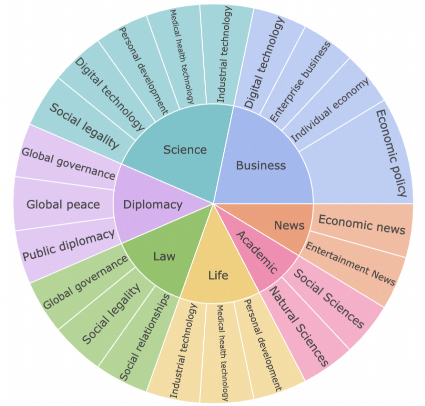


Figure 3: 7 common categories and 21 scenarios details of the evaluation dataset.

Model	S_a	S_c	S_m	S_f	S_n	Avg
Chinese-oriented Models						
ChatGLM3-6B	0.56	0.20	0.72	0.73	0.71	0.58
Qwen1.5-7B-Chat	0.63	0.15	0.72	0.68	0.71	0.58
Qwen1.5-14B-Chat	0.66	0.16	0.72	0.69	0.73	0.60
Qwen1.5-72B-Chat	0.72	0.16	0.71	0.71	0.67	0.60
Deepseek-R1	0.70	0.39	0.72	**0.76**	0.49	0.61
English-oriented Models						
Mixture-7B-v0.2	0.77	0.08	0.70	0.74	0.55	0.57
Llama2-7B-Chat-hf	0.46	0.15	0.73	0.73	0.74	0.56
Llama2-13B-Chat-hf	0.48	0.15	0.74	0.72	0.74	0.56
Llama2-70B-Chat-hf	0.60	0.11	0.69	0.67	0.62	0.55
Close-source Models						
GPT-3.5-turbo	0.62	0.11	0.71	0.72	0.62	0.56
GPT-4o	0.79	0.23	0.71	0.74	0.58	0.61
Ours	**1.00**	**1.00**	**0.75**	0.75	**0.81**	**0.86**

Table 3: Comparison of performance of various models on our evaluation system for QG tasks.

the quotation, given the preceding text and the initial k characters of the quotation. Generally, lower PPL values suggest that the model produces more accurate and coherent quotations. The specific calculation formula is as follows:

$$PPL_q = PPL\left([q_{n-t}; c_r] \mid [c_l; q_t]\right) \tag{10}$$

where n represents the length of the quote, q_t represents the first t characters of the quote, q_{n-t} represents the remaining $n-t$ characters of the quote.

Semantic Matching Semantic matching is concerned with ensuring semantic consistency and logical coherence within the context. This is achieved by predicting the PPL of the subsequent text, given the preceding text and the entire quote. A lower PPL value means that the quotation is more semantically consistent with the following context. The calculation formula is as Equation (4).

Novelty The Novelty metric evaluates the originality of generated quotations. By avoiding repetition and clichés, this metric ensures that content remains fresh and engaging, providing unique perspectives across various contexts. The specific calculation formula is as Equation (8).

To integrate the advantages of the three indicators, we employ a weighted average method, utilizing it as our final quotation-specific rerank metric. This comprehensive indicator seeks to balance semantic matching, fluency, and novelty, thereby enhancing the overall quality of model-generated citations. Finally, after the rerank stage, we select the top-1 quote including author or source information, and add it to the prompt. Then, the model inserts and rewrites quotes dynamically in the context, and ultimately outputs the results we need.

7 Experiments

7.1 Experiment Setup

Evaluation Dataset To construct the evaluation dataset, we select seven key categories: economy, diplomacy, journalism, academia, law, technology, and life. Additionally, 21 frequently cited scenarios are chosen to cover diverse knowledge aspects (Fig. 3). Initially, quotes are collected from each scenario to ensure diversity, richness, and relevance. These quotes serve as keywords for search engine queries, retrieving articles containing them. Relevant contexts are then extracted. To ensure quality, preprocessing includes deduplication, error correction, followed by manual sampling and validation to assess dataset reliability. The final dataset comprises 600 context-quote pairs.

Models We evaluate 9 models ranging from their model sizes and structures, which fall into three categories: Chinese-oriented models, English-oriented models, and Close-source models.

Models for PPL Calculation We employ two advanced language models, Qwen2-7B (Bai et al., 2023) and Llama3-8B (Touvron et al., 2023). These models are used to compute the PPL of the context given the preceding text. Subsequently, the average PPL values calculated by the two models are taken as final PPL values, which balances the judgments of the two models and reduces potential bias introduced by any single model. Since larger models tend to produce lower PPL for the same text, we recommend using the same PPL calculation models in this study when evaluating QG tasks.

Method	HR@1	HR@3	nDCG@1	nDCG@3	MRR
Vanilla	0.13	0.43	0.50	0.72	0.35
Supervised					
BM25	0.19	0.50	0.54	0.71	0.39
monoT5 (3B)	0.31	0.61	0.65	0.77	0.48
Unsupervised					
UPR (FLAN-T5-XL)	0.31	0.52	0.63	0.74	0.46
bge-reranker-large	0.32	0.55	0.71	0.82	0.47
LLM API (Permutation Generation)					
GPT-3.5-turbo	0.33	0.61	0.72	0.84	0.50
GPT-4o	0.43	0.63	0.74	**0.88**	0.55
Quotation-specific Reranking Metric					
$\mathrm{PPL_q}$	0.45	0.66	0.71	0.83	0.57
$\mathrm{PPL_m}$	0.34	0.60	0.64	0.77	0.50
$\mathrm{PPL_{avg}}$	0.33	0.60	0.64	0.76	0.50
$\mathrm{PPL_q}$+Novelty	0.34	0.58	0.63	0.73	0.50
$\mathrm{PPL_m}$+Novelty	0.46	0.65	0.70	0.78	0.57
$\mathrm{PPL_{avg}}$+Novelty(ours)	**0.49**	**0.67**	**0.74**	0.79	**0.60**

Table 4: Performance of different rerank metrics on Hit@1, Hit@3, nDCG@1, nDCG@3 and MRR. PPL_q, PPL_m and Novelty are as defined in Section 6, and PPL_{avg} is the average of PPL_q and PPL_m. Best performing rerank method are marked bold.

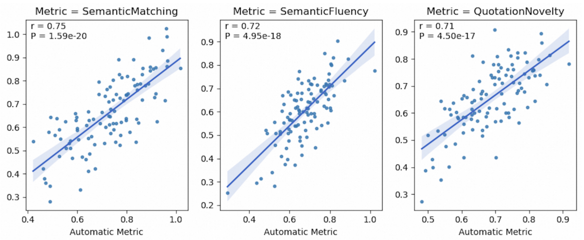


Figure 4: Correlation between our automatic evaluation metrics and human ratings. To avoid overlapping points, random jitters sampled from $N(0, 0.05^2)$ were added to human ratings after fitting the regression.

7.2 Results

We conduct experiments on models of different ranges and sizes on our benchmark, and the results are shown in Tab. 3. For more detailed analysis, please refer to Appendix E.

Severity of Quotaiton Hallucination The results show that more than half of the citations generated by LLaMA2-13B-Chat are not genuine quotes. Furthermore, despite varying parameter sizes, all models demonstrate suboptimal performance on the QR task, especially on the S_c metric. Even the best-performing model, GPT-4o, only scores 0.23 on the S_c indicator, highlighting the critical need to address the quotation hallucination problem.

Performance of Quotation-specific Reranking Metric Notably, our Quotation-specific Reranking method achieves the best results in each indicator, demonstrating the effectiveness of our method. Since our method retrieves the most relevant and appropriate citations from the quotation database, it ensures the authenticity and credibility of the citations. Therefore, both S_a and S_c are equal to 1. In addition, our method can effectively improve the novelty of citations and alleviate the problem of generating common citations with LLMs.

Comparison between Model Sizes We conduct further analysis on different model sizes. Within the same series, larger models tend to show improved performance. This indicates that larger models have richer quotation memory and stronger instruction-following capabilities.

7.3 Ablation Study

Correlations with Human Ratings We randomly select five samples for each scenario from the evaluation dataset, totaling 105 data samples. Due to the varying requirements for background knowledge across different categories (Fig.3), this study specifically invites expert professors in relevant fields to manually score evaluation metrics. Since Quotation Authenticity and Credibility are objective factual metrics, the manual evaluation primarily focused on the remaining three metrics. The process is independently conducted by experts, who are free to consult relevant literature and materials during the evaluation to ensure the reliability and objectivity of the results. Subsequently, we employ correlation analysis to assess the degree of association between various metrics and the overall evaluation results. As shown in Fig. 4, all metrics exhibit high levels of correlation. Specifically, the correlation coefficients are significantly higher than the threshold for statistical significance, indicating that our metric system effectively reflects the actual conditions of the evaluation subjects. For correlation analyses of specific categories, please refer to the Appendix F, where the results also reveal a significant correlation between manual and automated metrics for each category.

Correlation between Evaluation Metrics We present the correlations among the five automatic metrics in Tab. 6. As shown, the correlations between the metrics are all weak. This indicates that the five metrics are mutually independent, making it necessary to evaluate each of them individually in order to obtain a comprehensive view of the citation generation task assessment.

	Naive-0-Shot						Naive-1-Shot						Naive-2-Shot						Naive-CoT					
Model	S_a	S_c	S_m	S_f	S_n	Avg	S_a	S_c	S_m	S_f	S_n	Avg	S_a	S_c	S_m	S_f	S_n	Avg	S_a	S_c	S_m	S_f	S_n	Avg
Chinese-oriented Models																								
ChatGLM3-6B	0.56	0.20	0.72	0.73	0.71	0.58	0.59	0.13	0.72	0.68	0.67	0.56	0.62	0.13	0.72	0.68	0.68	0.57	0.64	0.16	0.71	0.69	0.67	0.57
Qwen1.5-7B-Chat	0.63	0.15	0.72	0.68	0.71	0.58	0.66	0.13	0.72	0.70	0.71	0.59	0.67	0.13	0.71	0.70	0.69	0.58	0.67	0.13	0.72	0.69	0.69	0.59
Qwen1.5-14B-Chat	0.66	0.16	0.72	0.69	0.73	0.60	0.68	0.17	0.72	0.67	0.71	0.60	0.74	0.18	0.71	0.71	0.65	0.60	0.69	0.18	0.72	0.73	0.68	0.60
Qwen1.5-72B-Chat	0.72	0.16	0.71	0.71	0.67	0.60	0.67	0.21	0.72	0.72	0.67	0.60	0.63	0.18	0.72	0.71	0.65	0.58	0.78	0.20	0.70	0.71	0.65	0.61
Deepseek-R1	0.70	**0.39**	0.72	**0.76**	0.49	0.61	0.67	**0.38**	0.71	0.71	0.54	0.60	0.71	**0.38**	0.71	0.72	0.54	0.62	0.77	**0.35**	0.71	**0.74**	0.54	0.62
English-oriented Models																								
Mixture-7B-v0.2	0.77	0.08	0.70	0.74	0.55	0.57	**0.82**	0.17	0.71	**0.75**	0.52	0.59	**0.82**	0.15	0.70	0.75	0.46	0.58	0.77	0.09	0.71	0.73	0.58	0.58
Llama2-7B-Chat-hf	0.46	0.15	0.73	0.73	0.74	0.56	0.46	0.09	0.73	0.71	0.66	0.53	0.44	0.12	**0.73**	0.74	0.67	0.54	0.49	0.14	**0.74**	0.73	0.70	0.56
Llama2-13B-Chat-hf	0.48	0.15	**0.74**	0.72	**0.74**	0.56	0.44	0.10	**0.74**	0.72	**0.74**	0.56	0.50	0.13	0.73	0.68	**0.74**	0.57	0.45	0.10	0.73	0.67	**0.74**	0.55
Llama2-70B-Chat-hf	0.60	0.11	0.69	0.67	0.62	0.55	0.65	0.20	0.71	0.66	0.67	0.58	0.70	0.20	0.71	0.69	0.63	0.59	0.75	0.13	0.71	0.68	0.66	0.59
Close-source Models																								
GPT-3.5-turbo	0.62	0.11	0.71	0.72	0.62	0.56	0.72	0.16	0.71	0.75	0.59	0.59	0.73	0.14	0.71	0.74	0.57	0.58	0.76	0.10	0.71	0.70	0.58	0.57
GPT-4o	**0.79**	0.23	0.71	0.74	0.58	**0.61**	0.75	0.24	0.70	0.74	0.60	**0.61**	0.80	0.23	0.71	**0.76**	0.57	**0.62**	**0.83**	0.22	0.71	0.73	0.60	**0.62**

Table 5: Comparison of performance of various models on our evaluation system for QG tasks in in Naive-0-shot, Naive-1-shot, Naive-2-shot and Naive-cot settings. In these naive experimental setup, our experiment does not employ RAG or rerank metrics. Instead, it relies solely on a specifically designed prompt to enable the models to execute the QG task. The prompt for each setting is detailed in the Appendix G.

Metric	S_a	S_c	S_m	S_f	S_n
S_a	1.000	-0.038	-0.018	-0.132	0.077
S_c	-0.038	1.000	-0.033	0.025	0.005
S_m	-0.018	-0.033	1.000	0.070	0.004
S_f	-0.132	0.025	0.070	1.000	0.002
S_n	0.077	0.005	0.004	0.002	1.000

Table 6: Correlation Matrix between Evaluation Metrics

Effectiveness of Reranking Metrics This study delves into the effectiveness of the rerank metric designed in our method and validates it through a series of ablation experiments. We adopt the following metrics: **HR@K**(K=1,3), **NDCG@K**(K=1,3), and **MRR** for comparison. On our benchmark, we compare a range of defined quotation-rerank metrics with state-of-the-art supervised, unsupervised, and closed-source API-based reranking methods. The supervised baselines include: BM25 (Nogueira and Cho, 2019) and monoT5 (Nogueira et al., 2020). The unsupervised baselines comprise UPR (Sachan et al., 2023) and bge-reranker-large (BAAI, 2023). The closed-source API-based baselines include ChatGPT3.5 and ChatGPT4. As shown in Tab. 4, our simple yet effective quotation reranking metrics that demonstrate strong performance across various evaluation criteria. Importantly, both supervised and unsupervised methods underperform compared to our proposed metrics. This indicates that our approach effectively captures the nuances of the QR task, leading to superior citation recommendations. We also conduct a comprehensive case analysis to demonstrate the alignment of our metric with human evaluations, as shown in Tab. 13 in the Appendix.

Comparison between Prompt Strategies We compare various prompting methods for QG tasks, including 0-shot, 1-shot, 2-shot, and Chain of Thought (CoT) (Wei et al., 2023) strategies. As shown in Tab. 5, among the four naive settings, the CoT method outperforms the others . The performance variations among the few-shot settings are not statistically significant, which suggests that the model's in-context learning capability will not substantially enhance its quotation performance. In contrast, the logical reasoning stimulated by the CoT method improves the model's quotation abilities to a certain degree.

8 Conclusion

In this paper, we systematically explore methods to enhance the performance of QR tasks in LLMs. Initially, we establish a holistic and automatic evaluation system consisting of five highly interpretable criteria, facilitating automatic evaluation of this task. Then, we construct a comprehensive and high-quality knowledge database containing up to 32,022 quotes, complete with authors or sources. Moreover, we design a fine-grained quotation-specific metric to rerank the retrieved quotes from the database to improve QG performance. Additionally, we conduct extensive experiments to verify that our metrics are strongly correlated with human preference and significantly effective in both open and close source LLMs.

Limitations

This study highlights several limitations. First, we primarily use PPL to evaluate text fluency. Although PPL is widely applied, it only measures the divergence between the model's and true probability distributions. Future research should integrate additional metrics or human evaluations for a more comprehensive assessment. Second, our analysis is restricted to specific contexts with clear correlations before and after quoted content. While informative, this approach does not cover a wide range of quoting scenarios. Future studies should explore diverse applications for more generalizable insights. Third, this study mainly considers two languages, English and Chinese, and future research should consider adding more languages.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2024. Do language models know when they're hallucinating references? *Preprint*, arXiv:2305.18248.

Yeonchan Ahn, Hanbit Lee, Heesik Jeon, Seungdo Ha, and Sang goo Lee. 2016. Quote recommendation for dialogs and writings. In *CBRecSys@RecSys*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

BAAI. 2023. Bge-reranker-large: A pre-trained model for ranking tasks. `https://huggingface.co/BAAI/bge-reranker-large`.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *Preprint*, arXiv:2305.14908.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Preprint*, arXiv:1406.1078.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *Preprint*, arXiv:2212.08597.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Preprint*, arXiv:2303.16104.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. 2024. An empirical study on information extraction using large language models. *Preprint*, arXiv:2409.00369.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Raynard S Kington, Stacey Arnesen, Wen-Ying Sylvia Chou, Susan J Curry, David Lazer, and Antonia M Villarruel. 2021. Identifying credible sources of health information in social media: principles and attributes. *NAM perspectives*, 2021.

Sabine Krumm, Manfred Berres, Sasa L Kivisaari, Andreas U Monsch, Julia Reinhardt, Maria Blatow, Reto W Kressig, and Kirsten I Taylor. 2020. Cats and apples: Semantic fluency performance for living things identifies patients with very early alzheimer's disease. *Archives of Clinical Neuropsychology*, 36(5):838–843.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Ansel MacLaughlin and David Smith. 2021. Content-based models of quotation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2296–2314, Online. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *Preprint*, arXiv:2305.15852.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

OpenAI. 2022. Introducing chatgpt.

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmt5: Modular multilingual pre-training solves source language hallucinations. *Preprint*, arXiv:2305.14224.

Fanchao Qi, Yanhui Yang, Jing Yi, Zhili Cheng, Zhiyuan Liu, and Maosong Sun. 2022a. QuoteR: A benchmark of quote recommendation for writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 336–348, Dublin, Ireland. Association for Computational Linguistics.

Fanchao Qi, Yanhui Yang, Jing Yi, Zhili Cheng, Zhiyuan Liu, and Maosong Sun. 2022b. Quoter: A benchmark of quote recommendation for writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 336–348.

Quora. 2020. What happens if you make too many citation mistakes in your research paper?

Samson Tan George Karypis Rami Aly, Zhiqiang Tang. 2024. Learning to generate answers with citations via factual consistency models. *arXiv preprint arXiv:2406.13124v1*.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2023. Improving passage retrieval with zero-shot question generation. *Preprint*, arXiv:2204.07496.

Pavel Savov. 2021. Measuring the novelty of scientific papers.

Arvind Krishna Sridhar and Erik Visser. 2023. Improved beam search for hallucination mitigation in abstractive summarization. *Preprint*, arXiv:2212.02712.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2015a. Learning to recommend quotes for writing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2015b. Learning to recommend quotes for writing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). [Online; accessed 2024-10-22].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Lingzhi Wang, Jing Li, Xingshan Zeng, Haisong Zhang, and Kam-Fai Wong. 2020. Continuity of topic, interaction, and query: Learning to quote in online conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6640–6650, Online. Association for Computational Linguistics.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2021. Quotation recommendation and interpretation based on transformation from queries to quotations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 754–758, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2024. The knowledge alignment problem: Bridging human and external knowledge for large language models. *Preprint*, arXiv:2305.13669.

Appendix

A Details of Quotation Knowledge Base

This chapter further analyzes the data details in the quotation corpus, which is divided into three languages: English, Standard Chinese, and Classical Chinese, all classified by topic and author. The number of topics and authors for each language is shown in Tab.7.

Language Type	Topic	Author	Total
English	1,216	6,624	16,393
Standard Chinese	228	2,377	7,519
Classic Chinese	869	876	8,110

Table 7: The specific topics, authors, and total count of the quotation corpus.

In addition, we also conduct analysis on the proportion of different topics in each language in the corpus, as shown in Fig. 5 - 6. for specific topics and proportions.

English Corpus To construct the English quotation corpus, we extract approximately 27,260 quotes covering different topics from the *BrainyQuote*[5], *A-ZQuote*[6] and *Goodreads*[7] websites, categorizing them by topic and author.

Classical Chinese Corpus Considering the representativeness and novelty of the Chinese corpus, we first collect some famous citations from *Gushiwen*[8]. Subsequently, given the limited number of citations, we utilize LLM to conduct a meaningful selection of the collected poems from *BaiduHanyu*. For instance, the seven-character quatrains in Tang poetry can be divided into two citations. Furthermore, to enhance the generalization of themes, we employ LLM to summarize the topics of the quotes. Finally , we collect over 9,233 citations with its poems, author and topics, including various genres such as Tang poetry and Song lyrics.

Standard Chinese Corpus Regarding the Standard Chinese quotation corpus, we gather 13,453 quotes from the *Guzimi*[9] and *Mingyancidian*[10] websites, similarly categorized by topic and author.

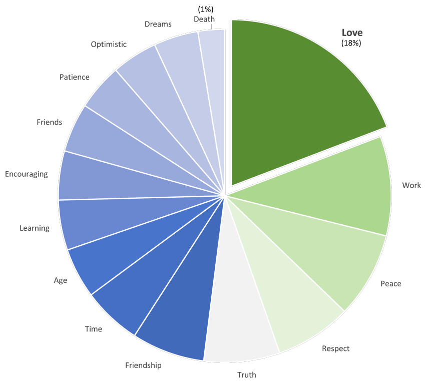


Figure 5: The specific topic distribution of the English quotation corpus.

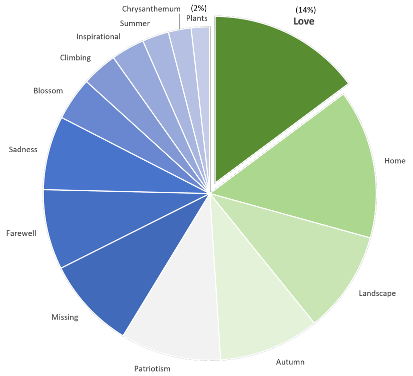


Figure 6: The specific topic distribution of the Classic Chinese quotation corpus.

Dataset Evolution The corpus collected from various websites has two limitations: (1) Semantic redundancy, (2) Excessive length. To address these, we first used the Jaccard similarity coefficient to reduce semantic redundancy. Then, we applied a length restriction and removed extreme cases based on the quotation perplexity metric.

B Effectiveness of GPT-4o Extraction

Previous studies have demonstrated that GPT-4o exhibits superior performance in simple information extraction tasks under zero-shot settings (Han et al., 2024). In our context, we primarily extract authors and sources, which involves only two fields, thus categorizing this as a simple information extraction task. Therefore, we believe that GPT-4o is capable of achieving excellent extraction performance. We also conduct experiments to validate the extraction effectiveness of GPT-4o in our task. We extract 100 citations containing authors or sources from a citation knowledge base in three different languages and use these as keywords to search on the

[5] https://www.brainyquote.com/
[6] https://www.azquotes.com/
[7] https://www.goodreads.com/
[8] https://www.gushiwen.cn/
[9] https://www.juzimi.com.cn/mingyan/
[10] http://mingyan.juzicidian.com

Bing search engine. In the returned search results, GPT-4o is utilized to extract the authors or sources of the citations. Ultimately, we assess the matching degree between the fields extracted by GPT-4o and the annotated fields in the knowledge base. The specific results are as follows:

Language Type	Extraction accuracy
English	97%
Standard Chinese	95%
Classic Chinese	98%

Table 8: Extraction and verification results of ChatGPT

C Details of the Inverted Sigmoid Function

To map the calculated Perplexity (PPL) values to a range of [0, 1], this study employs the Sigmoid function, which not only maps the scores to [0, 1] but also handles positive extreme values in the data. For the two key parameters of the Sigmoid function, k and μ, the calculation methods used in this study are as follows:

For μ: The Sigmoid function outputs 0.5 when $x = \mu$, where the slope is at its maximum. Typically, μ is set to the median or mean of the data, ensuring that the middle values are mapped to 0.5. In this study, we choose the mean of the data as the value for μ.

For k: The slope parameter k controls the "compression degree" of the mapping. A larger k results in a steeper Sigmoid curve, which is suitable for data with a concentrated distribution. In contrast, a smaller k results in a gentler curve, making it more appropriate for data with a wide range or extreme outliers. This study calculates k based on an empirical formula as follows:

$$k = \frac{ln(9)}{Q_{95} - Q_5} \tag{11}$$

where $ln(9) \approx 2.2$, corresponding to the span of the Sigmoid function from 0.1 to 0.9, Q_5 represents the 5% digit of the data, and Q_{95} represents the 95% digit of the data.

D Details of Evaluation Dataset

We also conducted manual analysis on the Evaluation Dataset, selecting 275 quotes from numerous context-quote pairs, dividing into Chinese and English, which categories and scenarios details are shown in Fig. 3. After statistics, there are 204 Chinese samples and 71 English samples, with a total of 144 Chinese and English authors.

E More Analysis of Experimental Results

Due to space limitations, we provide more experimental results analysis in this section.

Comparison between Model Types The performance comparison between the Chinese-oriented group and the English-oriented group on the Chinese-English benchmark reveals no significant differences, suggesting that the model's quotation ability is not language-dependent. Overall, the current opensource small to large-scale models exhibit a relatively small performance gap compared to close-source models, indicating the universality of the issue of quotation hallucination in LLMs.

F Details of Human Evaluation Metrics

We randomly selected 5 samples for each scenario from the evaluation dataset, totaling 105 data. Since different scenarios have different requirements for background knowledge, this study specially invited professional professors in related fields to manually score different categories of data. The scoring process was completed independently by experts, and relevant literature and materials were freely available during the review process to ensure the reliability and objectivity of the scoring results. In addition, we also further analyzed the correlation analysis of each specific category. The results are shown in the Table. 9. It can also be seen from the results that the manual indicators and automatic indicators of each category are also significantly correlated.

Metric	Overall	Science	Business	News	Academic	Life	Law	Diplomacy
Authenticity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Credibility	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Matching	0.75	0.74	0.72	0.74	0.67	0.70	0.83	0.71
Fluency	0.72	0.71	0.69	0.71	0.64	0.67	0.80	0.68
Novelty	0.71	0.70	0.68	0.70	0.63	0.66	0.79	0.67

Table 9: Metric evaluation results across different categories.

G Details of Naive Setting Prompts

For the naive experimental settings, we also disclose its prompt in detail, see Tab. 10 for Naive-0-Shot, Tab. 11 for Naive-1-Shot, and Tab. 12 for Naive-Cot setting.

/ Task prompt */*
Suppose you are a literary scholar and are familiar with many famous people's quotes. You are required to populate contextualised quotes based on user input text within the specified [Q] symbols.

/ Output requirements */*
1. The famous quotes must be quotes from a famous person in history or in the present, Please output the quote in English.
2. The quote should be closely related to the context, so that the context is more reasonable, smooth and beautiful.
3. If there is a specified author in the context, the famous quote must be given according to the corresponding restrictions.
4. Output Formate: "quote".
5. Only output the quote, NO MORE INFORMATION!
6. The number of quote should be 5 to 30 words.

/ Input */*
—INPUT—
{**Query**}
—OUTPUT—

Table 10: The details of the prompt for Naive-0-Shot setting.

/ Task prompt */*
Suppose you are a literary scholar and are familiar with many famous people's quotes. You are required to populate contextualised quotes based on user input text within the specified [Q] symbols.

/ Output requirements */*
1. The famous quotes must be quotes from a famous person in history or in the present, Please output the quote in English.
2. The quote should be closely related to the context, so that the context is more reasonable, smooth and beautiful.
3. If there is a specified author in the context, the famous quote must be given according to the corresponding restrictions.
4. Output Formate: "quote".
5. Only output the quote, NO MORE INFORMATION!
6. The number of quote should be 5 to 30 words.

/ Example */*
—INPUT—
.[Q], said by Confucius in Analects of Confucius - Wei Linggong. So is reading. Hard reading is the foundation, good reading is the key. In order to read effectively, you also need to make use of its "tools".
—OUTPUT—
"To do a good job, you must first sharpen your tools."

/ Input */*
—INPUT—
{**Query**}
—OUTPUT—

Table 11: The details of the prompt for Naive-1-Shot setting.

H QUILL Application

In this study, we conduct a comprehensive case analysis to demonstrate the efficacy and alignment of our reranking metric with human evaluations. As shown in Tab. 13 in the Appendix, we focus on several key models for comparison: the supervised BM25 and our own reranking metric, which combines average perplexity (PPL$_{avg}$) with novelty. Additionally, we manually sort and annotate the top-5 quote list initially recalled, serving as a benchmark for comparison. The findings reveal that our metric exhibits a higher correlation with human sorting than the other methods, underscoring its broad applicability and effectiveness. See the Appendix for a detailed comparison of the unsupervised UPR, the closed-source model GPT-3.5-turbo, and our approach.

I NDCG Formulation

In our experiment, in order to get the relevance between quote and query, we first use GPT-4o to score the relevance and get the complete relevance list after manual sampling. Hence, given m candidate quotes $Q = \{q_1, q_2, \cdots, q_m\}$, the nDCG@k is defined as follows:

$$\text{nDCG}(k) = \frac{\text{DCG}(O_{\text{real}}, k)}{\text{DCG}(O_{\text{ideal}}, k)} \qquad (12)$$

$$\text{DCG}(O, k) = \sum_{i=1}^{k} \frac{Rel_i}{\log_2(1+i)} \qquad (13)$$

where O_{ideal} and O_{real} represent the score list given by the ideal ranking relevance and the real ranking

/* Task prompt */
Suppose you are a literary scholar and are familiar with many famous people's quotes. You are required to populate contextualised quotes based on user input text within the specified [Q] symbols.

/* Output requirements */
1. The famous quotes must be quotes from a famous person in history or in the present, Please output the quote in English.
2. The quote should be closely related to the context, so that the context is more reasonable, smooth and beautiful.
3. If there is a specified author in the context, the famous quote must be given according to the corresponding restrictions.
4. Output Formate: "quote".
5. Only output the quote, NO MORE INFORMATION!
6. The number of quote should be 5 to 30 words.

Please think step by step then return the result!!!

/* Examples */
1: —INPUT—
.[Q], said by Confucius in Analects of Confucius - Wei Linggong. So is reading. Hard reading is the foundation, good reading is the key. In order to read effectively, you also need to make use of its "tools".
—OUTPUT—
"To do a good job, you must first sharpen your tools."
2: —INPUT—
.[Q]. As an ancient civilisation and a responsible power, it has always been China's pursuit to help the world. By guiding the direction of the world's changing circumstances with Chinese concepts, Chinese-style modernisation will advance and expand in benign interaction with the world, and will also strengthen the power for world peace and provide opportunities for the development of all countries.
—OUTPUT—
"Already wanting to be established, we should be established; already wanting to achieve, we should achieve."

/* Input */
—INPUT—
{**Query**}
—OUTPUT—

Table 12: The details of the prompt for Naive-CoT setting.

relevance respectively, Rel_i denote the relevance score of the quote q_i.

Method	Literal Sentence	Recalled List	Metric Rerank	Human Rerank
BM25	Education empowers individuals to transform their lives and contribute to societal progress. [Q]. It fosters critical thinking, innovation, and social responsibility. By providing access to knowledge, education breaks down barriers and creates opportunities. It is a key driver of positive change and development.	Education is a human right with immense power to transform. On its foundation rest the cornerstones of freedom, democracy and sustainable human development.	Education is the transmission of civilization.	Education is the most powerful weapon which you can use to change the world
		Education is the transmission of civilization	Education is a human right with immense power to transform. On its foundation rest the cornerstones of freedom, democracy and sustainable human development	Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family.
		Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family	Education is the most powerful weapon which you can use to change the world	Education is a human right with immense power to transform. On its foundation rest the cornerstones of freedom, democracy and sustainable human development
		Education is the most powerful weapon which you can use to change the world	Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family	Education is the transmission of civilization
		The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the goal of true education	The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the goal of true education	The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the goal of true education
Ours	Education empowers individuals to transform their lives and contribute to societal progress. [Q]. It fosters critical thinking, innovation, and social responsibility. By providing access to knowledge, education breaks down barriers and creates opportunities. It is a key driver of positive change and development.	Education is a human right with immense power to transform. On its foundation rest the cornerstones of freedom, democracy and sustainable human development	Education is the most powerful weapon which you can use to change the world	Education is the most powerful weapon which you can use to change the world
		Education is the transmission of civilization	Education is a human right with immense power to transform. On its foundation rest the cornerstones of freedom, democracy and sustainable human development	Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family
		Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family	Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family	Education is a human right with immense power to transform. On its foundation rest the cornerstones of freedom, democracy and sustainable human development
		Education is the most powerful weapon which you can use to change the world	Education is the transmission of civilization	Education is the transmission of civilization
		The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the goal of true education	The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the goal of true education	The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the goal of true education

Table 13: The examples of recalled candidates reranked via different rerank metrics and human evaluation. The indicators [Q] denotes the insertion positions of the given context. A darker shade of green indicates a higher rank bestowed by humans. See the Appendix for a detailed comparison of the unsupervised UPR, the closed- source model GPT-3.5-turbo, and our approach.