Past Meets Present: Creating Historical Analogy with Large Language Models

Nianqi Li[1], **Siyu Yuan**[2*], **Jiangjie Chen**[3†],
Jiaqing Liang[2], **Feng Wei**[4], **Zujie Liang**[4], **Deqing Yang**[2], **Yanghua Xiao**[1*]
[1]Shanghai Key Laboratory of Data Science,
College of Computer Science and Artificial Intelligence, Fudan University
[2]School of Data Science, Fudan University [3]ByteDance Seed [4]MYbank, Ant Group
{nqli23,syyuan21}@m.fudan.edu.cn, shawyh@fudan.edu.cn

Abstract

Historical analogies, which compare known past events with contemporary but unfamiliar events, are important abilities that help people make decisions and understand the world. However, research in applied history suggests that people have difficulty finding appropriate analogies. And previous studies in the AI community have also overlooked historical analogies. To fill this gap, in this paper, we focus on the **historical analogy acquisition** task, which aims to acquire analogous historical events for a given event. We explore retrieval and generation methods for acquiring historical analogies based on different large language models (LLMs). Furthermore, we propose a self-reflection method to mitigate hallucinations and stereotypes when LLMs generate historical analogies. Through human evaluations and our specially designed automatic multi-dimensional assessment, we find that LLMs generally have a good potential for historical analogies. And the performance of the models can be further improved by using our self-reflection method.[1]

1 Introduction

Historical analogy, which draws comparisons between contemporary and past situations, is a vital tool in applied history (Achenbaum, 1983; Guldi and Armitage, 2014; Parsons and Nalau, 2016; Ghilani et al., 2017; Keulen, 2023). These analogies enable a deeper understanding of historical events and facilitate informed decision-making in addressing present difficulties (Bartha, 2013; Axelrod and Forster, 2017). For example, as shown in Figure 1, when the COVID-19 pandemic spread around the world, the influenza pandemic of 1918 emerged as an analogy, aiding in the navigation of the crisis. However, historians have found that individuals, particularly politicians, often misuse historical analogies. They tend to gravitate towards the first analogy that comes to mind, are influenced by superficial similarities, and rarely conduct thorough analyses (Ghilani et al., 2017; Khong, 2020). Furthermore, the creation of historical analogies involves having extensive knowledge of historical events and selecting the appropriate one, which can also be a great challenge. Therefore, exploring large language models (LLMs) (AI-Meta, 2024; OpenAI, 2022, 2023) with the ability to automatically generate historical analogies is of great value.

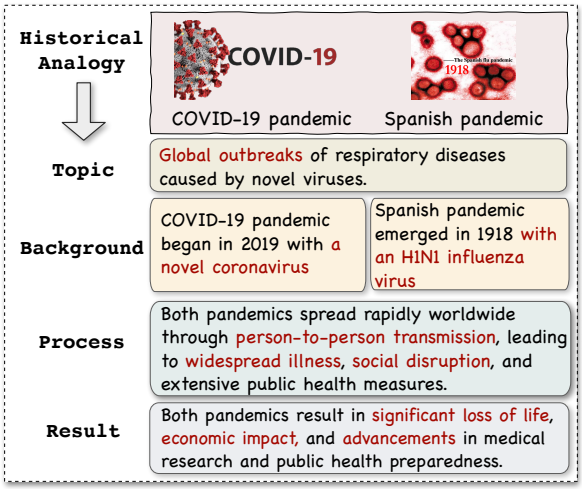

Figure 1: An example of historical analogy: analogizing COVID-19 to the Spanish pandemic based on topic, background, process, and result.

Traditional studies within the AI community have concentrated on recognizing and generating word analogies, *e.g.*, "king is to man as queen is to woman", using word embeddings (Mikolov et al., 2013; Gladkova et al., 2016; Fournier et al., 2020; Ushio et al., 2021) or by training language models (LMs) (Czinczoll et al., 2022; Chen et al., 2022; Yuan et al., 2023b). Recently, with the advance-

*Corresponding author.
†Part of the work done while at Fudan Univeristy.
[1]Resources of this paper can be found at https://github.com/Nianqi-Li/Historical-Analogy-of-LLMs

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3942–3957
July 27 - August 1, 2025

ment of LLMs, some researchers have designed prompts to instruct LLMs to generate free-form analogies (Webb et al., 2022). However, these efforts are limited to the scientific domain (*e.g.*, analogies for atom structure) (Bhavya et al., 2022; Sultan and Shahaf, 2022; Jiayang et al., 2023; Yuan et al., 2023b; Sultan et al., 2024; Yuan et al., 2024) or to everyday scenarios derived from webpages (Wijesiriwardene et al., 2023; Ding et al., 2023; Bhavya et al., 2024), neglecting the exploration of analogies that draw comparisons between contemporary and historical situations, which could provide a comprehensible perspective on history.

In this paper, we explore the concept of historical analogy and introduce a new task, *i.e.*, *historical analogy acquisition*. This task aims to find historical events analogous to current events. Specifically, given an event's name and text description, the ultimate goal is to obtain another historical event analogous to the original event in multiple dimensions, such as cause, process, and result. To test the performance of LLMs on this task, we employ various methods based on two paradigms: 1) dataset retrieval methods, which employ LLMs to retrieve historical events from a specified dataset, and 2) free generation methods, which instruct LLMs to autonomously generate analogous historical events, leveraging the knowledge stored in their parameters. Furthermore, to mitigate the hallucination and stereotyping in generating historical analogies, we propose the self-reflection method, which comprises two LLM-based modules: the Candidate Generator and the Answer Reflector. The Candidate Generator produces potential analogies, while the Answer Reflector offers feedback to refine these candidates to get rid of stereotypes. Additionally, we verify the candidates through Wikipedia API to ensure their authenticity.

For evaluation, we employ both human and automatic methods to thoroughly assess the quality of historical analogies. In human evaluation, we use a manual ranking system to examine historical analogies. To reduce labor, we also introduce automatic metrics designed to evaluate historical analogies across four dimensions: topic, background, process, and result. These dimensions represent the essential components of a historical event (Keulen, 2023). To measure these dimensions, we borrow the idea of Jiayang et al. (2023) to calculate abstract and literal similarities. By integrating these two types of similarities across the four dimensions, our automatic evaluation metrics demonstrate a high correlation with human evaluation.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, our work is the first to explore the historical analogy in the AI community.
- We develop a novel, automatic multi-dimensional metric to evaluate historical analogy from a cognitive perspective, ensuring alignment with human cognition.
- Through extensive experiments, we find that current LLMs have the potential for historical analogies. And by mitigating illusions and stereotypes in LLMs, our proposed self-reflection method can further improve the performance of LLMs in acquiring historical analogies.

2 Related Work

Analogy Making Rooted in classical theories of analogy such as structural mapping (Gentner, 1983; Holyoak and Thagard, 1996), early research in the AI community primarily focuses on generating word analogies (Falkenhainer et al., 1989; Turney and Littman, 2005; Gladkova et al., 2016; Fournier et al., 2020; Ushio et al., 2021; Yuan et al., 2023c) to examine the capabilities of LMs in analogy-making. Recent advancements in LLMs (OpenAI, 2022, 2023; AI-Meta, 2024) have expanded this focus from simple word analogies to the generation of analogies involving more complex entities, including systems (Yuan et al., 2023b), processes (Bhavya et al., 2022; Sultan and Shahaf, 2022; Sultan et al., 2024), paragraphs (Webb et al., 2022; Wijesiriwardene et al., 2023; Ding et al., 2023; Ye et al., 2024; Yuan et al., 2024), measurements (Chen et al., 2024), and stories (Jiayang et al., 2023). Despite these developments, most studies have concentrated on analogies within the scientific domain or everyday scenarios, overlooking the significance of historical analogy (Schuman and Rieger, 1992; Parsons and Nalau, 2016; Ghilani et al., 2017). In contrast, our research is the first to investigate and assess how LLMs can identify historical analogy, offering valuable insights for history and decision-making (Keulen, 2023).

Language Model as Knowledge Base Pretrained on extensive datasets, LLMs can implicitly encode a significant amount of knowledge within their parameters (AlKhamissi et al., 2022; Xie et al., 2024; Ju et al., 2024), enabling them to serve

as Knowledge Bases (KBs) (Petroni et al., 2019; Sung et al., 2021; West et al., 2022; Yuan et al., 2023a; Xu et al., 2024). However, relying solely on LLMs for knowledge generation can lead to hallucinations (Rawte et al., 2023; Zhang et al., 2023; Tonmoy et al., 2024), where the content produced seems factual but lacks grounding. To address this issue, some researchers have proposed the retrieval-augmented generation method (Shuster et al., 2021; Gao et al., 2023; Kirchenbauer and Barns, 2024) to mitigate hallucinations by leveraging external KBs. In this paper, we utilize LLMs to identify historical analogy, employing Wikipedia (Vrandečić and Krötzsch, 2014) as an external KB to verify the authenticity of historical events and effectively mitigate hallucinations.

3 Historical Analogy Generation

3.1 A Cognitive View for Historical Analogy

Historical analogy compares contemporary and past situations, offering an accessible view of history and validating policies and decisions, which is a vital tool in applied history (Schuman and Rieger, 1992; Keulen, 2023; Parsons and Nalau, 2016). For example, Margaret Thatcher likened Iraq's invasion of Kuwait to the Munich Agreement, thereby using historical analogy to support their intervention actions in Iraq (Conolly-Smith, 2009). In historical analogy, both events and personalities serve to formulate an argument by analogy, elucidating the present issue. However, research conducted by historians indicates that individuals, particularly politicians, ordinarily use history badly. They often gravitate towards the first analogy that comes to mind, are easily swayed by superficial similarities, and rarely pursue in-depth or extensive analysis (Ghilani et al., 2017; Dobney, 1974; Khong, 2020). Therefore, it is crucial to develop a framework that facilitates the automatic, straightforward, and precise acquisition of historical analogy.

3.2 Task Formulation

Historical analogy acquisition task aims to obtain a historical event for the given event to form an analogy. Given the input event $\mathcal{E}_I$ and its description $\mathcal{D}_I$, the goal is to output the event from history $\mathcal{E}_H$, which is analogous to the input event. Figure 1 presents an example of a historical analogy.

3.3 Data Construction

To comprehensively evaluate the ability of LLMs to acquire historical analogies, we categorize historical analogies into two categories, *i.e.*, popular analogy and general analogy.

Popular Analogy Popular analogies are analogies that are well known to the general public and already have standardized results, often proposed by newspapers, historians, and politicians, such as Figure 1. To obtain these analogies, we manually collect samples of popular analogies from web pages and articles related to historical analogies.[2] Due to the limited number of valid analogies and the presence of misuses or controversies, we end up with 20 test samples that are widely recognized, have standard answers, and show some degree of creativity.

General Analogy Since LLMs may have learned popular analogies during pre-training, we construct general analogy sets with events lacking universally recognized analogies. Specifically, we collect 658 historical events from Google Arts and Culture.[3] These events are categorized into four themes: War, Politics, Culture and Society, and Economy. We select 50 samples each from the first three categories and 10 from the Economy category, creating a balanced general analogical set to assess the LLM's ability to draw historical analogies across different themes. Since there are no standardized answers for general analogies, it is necessary to develop automated evaluation metrics to assess the quality of analogies between analogy events and input events.

3.4 Human Evaluation Metrics

Due to the lack of quantitative criteria for evaluating historical analogies, this paper uses a ranking approach for manual assessment. For $\mathcal{E}_I$, we employ three annotators from the history department to rank the $\mathcal{E}_H$ output from different methods according to the quality of the analogies, using a scale from 1 to n, with higher scores indicating better analogy quality. The frequency of each method being ranked best is also calculated to assess the quality of the analogies. Further details on the human evaluation process are provided in Appendix A.

[2]The online resources are shown in Appendix B

[3]https://artsandculture.google.com/category/event

3.5 Automatic Evaluation Metrics

For Popular Analogies, we can calculate the **Pass@1** based on the standard answers. However, it is not applicable to General Analogies, necessitating the development of broader metrics for automatically evaluating historical analogies quantitatively. Drawing on the historical applied science[4], we develop a multi-dimensional similarity metric (MDS) to evaluate historical analogies automatically.

Dimension Summary In historiography, the universal structure of events encompasses topic, background, process, and result. Therefore, for an event $\mathcal{E}$ and its description $\mathcal{D}$, we utilize GPT-4 to summarize these four dimensions based on $\mathcal{D}$, resulting in $\mathcal{D} = (\mathcal{D}^{\text{Topic}}, \mathcal{D}^{\text{Background}}, \mathcal{D}^{\text{Process}}, \mathcal{D}^{\text{Result}})$. The prompt template is shown in Appendix C.1.

Multi-level Similarity Previous research (Bunge, 1981; Jiayang et al., 2023) indicates that analogies are effective when they share abstract-level similarities, such as themes, central ideas, and processes, rather than identical entities and behaviors (*i.e.*, literal similarity). For abstract similarity, based on the four summarized dimensions, we instruct GPT-4 to rate the abstract similarity between $\mathcal{E}_I$ and $\mathcal{E}_H$ for each dimension on a scale from 1 to 4. The prompt template is shown in Appendix C.1. For literal similarity, we perform the NLTK tokenization (Bird, 2006) on each summary and calculate the Jaccard similarity (Niwattanakul et al., 2013) after removing stopwords. A higher abstract similarity score indicates a better analogy between $\mathcal{E}_I$ and $\mathcal{E}_H$, while lower literal similarity scores indicate more innovation. Thus, the overall multi-dimensional similarity formula is:

$$MDS = \sum_{d \in \mathcal{D}} w^d \cdot sim_{\text{Abs}}(\mathcal{D}_I^d, \mathcal{D}_H^d) \cdot max(\alpha - sim_{\text{Lit}}(\mathcal{D}_I^d, \mathcal{D}_H^d), 0), \quad (1)$$

where $\mathcal{D} = \{\text{Topic}, \text{Background}, \text{Process}, \text{Result}\}$, w_d represents the weight of each dimension, $\mathcal{D}_I^d$ ($\mathcal{D}_H^d$) represents the description of $\mathcal{E}_I$ ($\mathcal{E}_H$) in the d dimension. Given that descriptions are summarized by GPT-4, even identical events may have differing descriptions. Therefore, α serves as a threshold to prevent overly similar analogies.

Effectiveness of Automatic Evaluation To determine w^d and α, and to validate the automatic evaluation, we calculate the correlation coefficient between automatic and human evaluations. We use GPT4 to generate four different analogies for each popular analogy as the evaluation dataset. For the manual assessment, we employ three annotators to rank the four results, with Fleiss's $\kappa = 0.97$. For the automated assessment, we adopt our automatic multi-dimensional similarity metric to rank. For each analogy, we calculate the correlation coefficient between the two sets of rankings, then take the average across different analogies to obtain the final correlation coefficient.

The results show that the best correlation coefficient with the manual results is obtained when the dimension weights are (0.5, 1, 2, 2) and the similarity threshold α is 0.35. In this setting, the Kappa coefficient (Cohen, 1960) is 0.67, and the Pearson (Pearson, 1920) and Spearman correlation coefficients (Spearman, 1961) are 0.72 and 0.73, confirming the reliability of automatic evaluation.

4 Method

In this section, we explore various methods of leveraging LLMs to get historical analogies. These methods fall into two categories: 1) *dataset retrieval methods* and 2) *free generation methods*. The illustration of these methods is shown in Figure 2. The prompt templates for LLMs in each method are shown in the Appendix C.2.

4.1 Dataset Retrieval Method

A common practice to obtain analogous events is to select from existing datasets. In this paper, we use Google Arts and Culture as an event pool for LLMs to retrieve suitable analogies for a specified event. We implement two retrieval strategies:

Direct Retrieval This method embeds the description of the given event and events in the pool using `text-embedding-3-small` (Neelakantan et al., 2022). The event with the highest cosine similarity is then selected as $\mathcal{E}_H$.

Two-stage Retrieval This method first selects the top 10 historical events from the event pool using cosine similarity. Then, given their descriptions, LLM is asked to select the most appropriate analogies from the candidate set.

4.2 Free Generation Method

Due to the growing number of historical events, relying on a fixed dataset for analogies can lead to

[4] https://phi.history.ucla.edu/nchs/historical-thinking-standards/1-chronological-thinking/

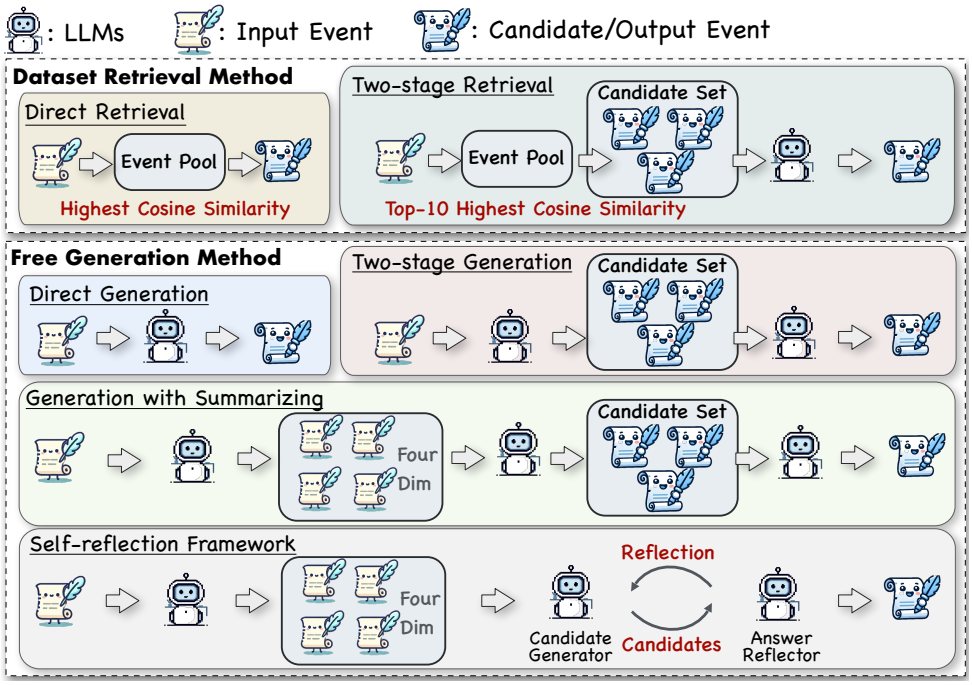

Figure 2: The illustration of different methods for historical analogy identification. We divide these six methods into two categories: dataset retrieval methods and free generation methods.

issues such as high overhead, slow processing, and challenges in updating. Since LLMs have learned extensive knowledge about historical events during pre-training, we can employ LLMs to generate analogous historical events.

Direct Generation Given $\mathcal{E}_I$ and $\mathcal{D}_I$, this method directly asks LLMs to generate the analogous historical event. However, it heavily depends on LLMs' knowledge and can be easily influenced by biases and stereotypes from pre-training.

Two-stage Generation Direct generation can lead to suboptimal results and even produce fictional historical events with hallucinations. To achieve a broader exploration, we ask LLMs to propose 10 candidate events based on $\mathcal{D}_I$. Given the potential for hallucination, each candidate must be verified through Wikipedia to confirm its authenticity. Then, LLMs compare the $\mathcal{D}_I$ and the descriptions of candidate events retrieved from Wikipedia, selecting the most appropriate one as $\mathcal{E}_H$.

Generation with Summarizing As mentioned in § 3.5, the common structure of events encompasses topic, background, process, and result. Thus, we can ask LLMs to summarize the input and candidates into these four dimensions and combine the summaries to form new descriptions to participate in the steps of the two-stage generation. Compared with the original descriptions, the summaries have shorter lengths and more effective information, so LLM can better understand the events and compare the similarities and differences in different dimensions to obtain better analogical results.

Self-reflection Framework Based on the evaluation results in § 5, the generation with summarizing improves output quality but is prone to stereotyping when proposing candidates, and the limitation of only 10 candidates restricts LLM options. Research on the self-reflection method (Shinn et al., 2023; Renze and Guven, 2024; Wang et al., 2024) shows that LLMs can provide feedback and update the unsuitable candidates. Inspired by this, we design two LLM-based modules: the Candidate Generator and the Answer Reflector. These modules collaboratively generate historical analogies through iterative processes. In each iteration, Candidate Generator proposes five candidates based on the $\mathcal{E}_I$'s descriptions of four dimensions. The Answer Reflector then assesses the candidate set. If no candidates are suitable as analogous historical events, the Answer Reflector instructs the Candidate Generator to revise the candidate set for the next iteration. If a suitable candidate is found, the Answer Reflector outputs $\mathcal{E}_H$ and concludes the iteration. Additionally, we also verify each candidate through Wikipedia to confirm its authenticity.

Dataset	Method	T_{Abs}	T_{Lit}	T_{All}	B_{Abs}	B_{Lit}	B_{All}	P_{Abs}	P_{Lit}	P_{All}	R_{Abs}	R_{Lit}	R_{All}	MDS
						GPT-3.5-Turbo								
Popular	Direct Re.	2.70	0.15	0.54	2.55	0.14	0.61	2.70	0.09	0.70	2.70	0.09	0.70	3.67
	Two-stage Re.	2.85	**0.13**	0.59	2.45	0.12	0.53	2.65	**0.08**	0.69	2.60	0.10	0.60	3.41
	Direct Gen.	3.10	0.15	0.64	2.64	0.11	0.68	2.94	0.12	0.73	3.15	0.10	0.75	3.97
	Two-stage Gen.	3.25	0.16	0.65	**2.90**	0.12	0.67	2.80	0.13	0.68	2.80	0.12	0.69	3.74
	Summarizing	3.30	0.14	0.67	2.70	0.11	0.63	**3.30**	0.10	**0.82**	**3.19**	0.10	0.76	4.14
	Self-reflection	**3.40**	0.13	**0.71**	2.89	**0.09**	**0.73**	3.09	0.10	0.75	3.09	**0.09**	**0.79**	**4.18**
						Llama3.1-8B								
	Direct Re.	2.70	0.15	0.54	2.55	0.14	0.61	2.70	0.09	0.70	2.70	**0.09**	0.70	3.67
	Two-stage Re.	2.80	**0.11**	0.63	2.45	0.10	0.59	2.60	0.08	0.69	2.44	0.12	0.55	3.38
	Direct Gen.	3.30	0.13	**0.70**	2.69	0.09	0.69	2.90	0.10	0.69	**3.10**	0.10	**0.74**	3.90
	Two-stage Gen.	2.94	0.13	0.64	2.55	0.08	0.67	2.80	**0.08**	0.73	2.80	0.10	0.68	3.81
	Summarizing	3.24	0.14	0.64	2.74	**0.08**	0.74	2.69	0.08	0.70	2.94	0.09	0.73	3.92
	Self-reflection	**3.34**	0.13	0.70	**2.84**	0.08	**0.74**	**3.15**	0.09	**0.81**	2.89	0.10	0.71	**4.13**
						GPT-3.5-Turbo								
General	Direct Re.	3.29	0.18	0.53	3.00	0.13	0.67	2.97	0.11	0.69	2.99	0.12	0.66	3.64
	Two-stage Re.	2.93	0.19	0.51	2.69	0.15	0.58	2.63	0.12	0.60	2.75	0.14	0.58	3.21
	Direct Gen.	2.88	**0.13**	0.62	2.67	**0.10**	0.65	2.63	**0.09**	0.69	2.79	**0.10**	0.68	3.69
	Two-stage Gen.	3.20	0.20	0.57	2.82	0.16	0.63	3.01	0.13	0.70	2.99	0.13	0.67	3.65
	Summarizing	3.49	0.18	**0.64**	3.02	0.13	0.68	3.11	0.12	0.74	3.07	0.13	0.67	3.83
	Self-reflection	**3.52**	0.17	0.61	**3.21**	0.12	**0.73**	**3.16**	0.11	**0.75**	**3.13**	0.12	**0.70**	**3.93**
						Llama3.1-8B								
	Direct Re.	3.29	0.18	0.53	3.00	0.13	0.67	2.97	0.11	0.69	2.99	0.12	0.66	3.64
	Two-stage Re.	3.11	0.16	0.58	2.78	**0.11**	0.64	2.81	**0.09**	0.71	2.73	0.11	0.63	3.60
	Direct Gen.	3.44	0.19	0.60	3.08	0.15	0.67	3.04	0.13	0.72	3.01	0.14	0.66	3.73
	Two-stage Gen.	3.21	**0.15**	**0.63**	2.91	0.11	0.69	2.91	0.10	0.73	2.80	**0.11**	0.66	3.77
	Summarizing	3.41	0.17	0.61	**3.11**	0.11	**0.72**	3.02	0.10	0.74	3.09	0.12	0.70	3.90
	Self-reflection	**3.46**	0.18	0.62	3.08	0.13	0.70	**3.08**	0.11	**0.75**	**3.12**	0.13	**0.70**	**3.91**

Table 1: Results of different methods on Popular Analogies and General Analogies based on ChatGPT and Llama3.1-8B. "Abs" ("Lit") denotes abstract similarity (literal similarity). "T", "B", "P", "R" denote the dimensions of Topic, Background, Process and Result. "MDS" denotes multi-dimensions similarity. The best results are **bolded**, and the second best ones are underlined, both counted to four decimal places.

5 Results

This section evaluates methods for historical analogy acquisition and identifies core challenges, such as stereotypes and differing perspectives. Furthermore, ablation studies reveal the critical components of the framework and validate the potential of LLMs for this task.

5.1 Model Choice

We use the open-source model Llama3.1-8B-Instruct (AI-Meta, 2024) and the closed-source model gpt-3.5-turbo-0125 (OpenAI, 2022) for the main experiment, with the temperature set to 0.1.

5.2 Main Result

Automatic Evaluation Results The results are shown in Table 1 and Figure 3. And Appendix D provides confidence intervals for the results. We find that: *1)* Both Llama and ChatGPT perform better on the Popular Analogy than on the General Analogy. In particular, Direct Generation method achieves a high Pass@1 in Popular Analogy. This discrepancy suggests potential data leakage during the pre-training phase within the Popular Analogy, emphasizing the importance of including General Analogy in evaluations. *2)* Free generation methods outperform dataset retrieval methods significantly, with an average improvement of 0.25. This improvement likely arises because a finite dataset cannot encompass the vast expanse of historical data, making generation from LLMs preferable to retrieval for historical analogies. *3)* The self-reflection method achieves the highest results for both open and closed-source models, indicating that incorporating reflection with feedback can enhance the quality of analogies. *4)* The summarizing method demonstrates notable enhancements over two-stage generation across both models and

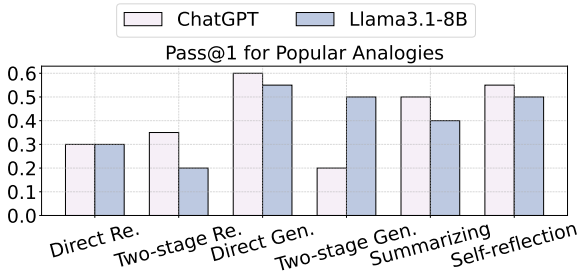

Figure 3: Pass@1 results of different methods on the Popular Analogies.

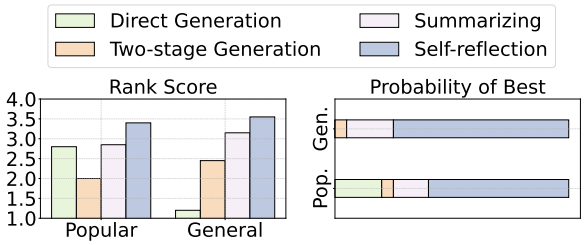

Figure 4: Human evaluation results for free generation methods.

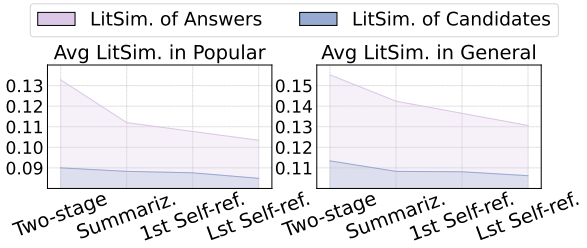

Figure 5: Average literal similarity of candidates and answers for free generation method. "1st" indicates the first round of self-reflection and "Lst' is the last round.

#Can.	Ref.	T_{A11}	B_{A11}	P_{A11}	R_{A11}	MDS
C=1	0.10	0.60	0.70	0.73	0.70	3.86
C=3	0.14	0.61	0.68	0.74	0.71	3.87
C=5	0.11	0.61	0.73	0.75	0.70	3.93
C=10	0.09	0.63	0.70	0.74	0.70	3.89
C=15	0.09	0.61	0.72	0.77	0.69	3.95

Table 2: Results of different candidate set sizes in the self-reflection method. **Ref.** indicates the average number of times the reflection is performed.

datasets, highlighting the effectiveness of dimension splitting in improving historical analogy generation. *5)* Surprisingly, the two-stage method generally underperforms compared to the direct method. This may be attributed to the lengthy and detailed descriptions, making it more difficult for the LLM to choose during the selection process.

Human Evaluation Results To further assess the performance of each method, we conduct a manual evaluation of the four free generation methods based on ChatGPT. The results are presented in Figure 4. In alignment with the automated results, the self-reflection method receive the highest ranking score and the highest percentage of optimal. Also, due to the internal knowledge leakage in LLM, direct generation performs well in Popular Analogies but poorly in General Analogies.

5.3 Detailed Analysis

Stereotypes in Historical Analogies Stereotypes in historical analogies are mainly manifested in generating events that focus on the same entities, *e.g.* countries, people, rather than on core ideas. In order to analyse the impact of stereotypes, we count the literal similarity scores of candidates and answers generated by the free generation method, which reflects the degree of stereotyping by measuring the proportion of shared entities. Figure 5 shows the results, which align with the motivations in § 4.2. In both Popular and General Analogy, the candidates and answers of the self-reflection method show the least stereotyping and further decrease with iteration.

Candidate Number and Reflective Rounds To further explore self-reflection method, we first test the effect of different candidate set sizes on it. The results in Table 2 show that increasing the candidate set size improves performance, but gains plateau after five candidates while token consumption continues to rise.

However, the low Ref. shows that only about 10% data executed reflection, indicating that LLMs prefer to accept the current candidate, even when the candidate set is small. To explore the impact of the number of reflection rounds on performance, we further require the LLM to warm up with a few rounds of reflection without output. Table 3 reveals that while a few warmups can slightly improve performance, additional rounds do not continue this trend and may reduce effectiveness due to inappropriate reflection.

Proposing and Selection Capability of LLMs As described in § 4, the methods of two-stage generation, summarizing and self-reflection need to propose a candidate set and select the analogous event $\mathcal{E}_H$ from this set. To prove the effectiveness of LLMs in proposing an appropriate candidate set and selecting the $\mathcal{E}_H$ from the set, we design the

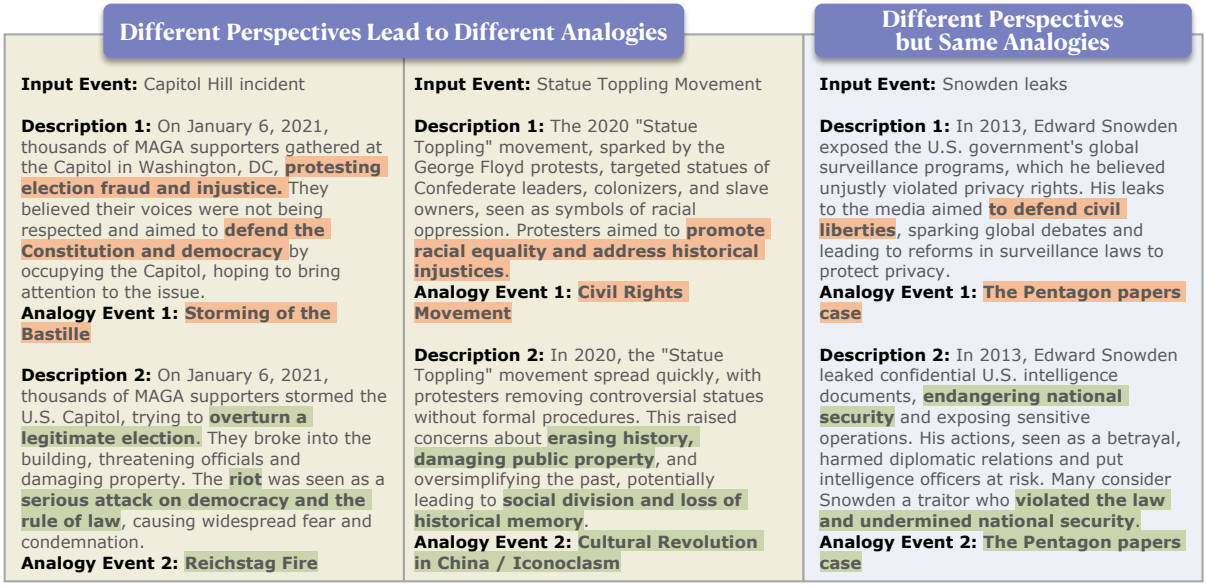

Figure 6: Case studies of historical analogy from different perspectives. Different perspectives often lead to distinct analogies, although a few analogies remain the same due to the ability to interpret the results from multiple viewpoints.

Warmup	T_{A11}	B_{A11}	P_{A11}	R_{A11}	MDS
W=0	0.61	0.73	0.75	0.70	3.93
W=2	0.66	0.72	0.77	0.73	4.03
W=5	0.63	0.70	0.73	0.68	3.85
W=10	0.66	0.70	0.75	0.71	3.94

Table 3: Results for different warmup turns in the self-reflection method.

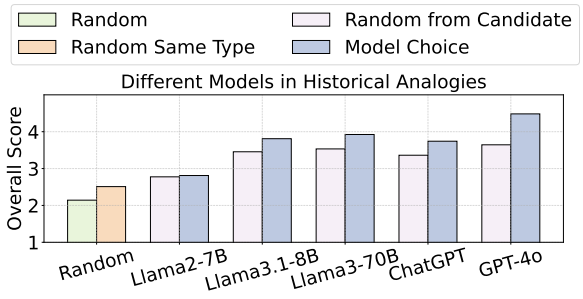

Figure 7: Performance of different models in historical analogies, including proposing candidate sets and selecting analogous event.

three ablated variants in the two-stage generation method for exploration: *1)* random selection from the event pool; *2)* random selection from the event pool within the same theme; *3)* random selection from the candidate set proposed by LLMs.

The results are shown in Figure 7. The evidence provided by ■ < ■ and ■ < ■ confirms that all models, ranging from smaller ones like Llama2-7B (Touvron et al., 2023) to larger ones like GPT-4 (OpenAI, 2023), are capable of generating candidate sets for historical analogies. Furthermore, stronger LLMs, *e.g.*, ChatGPT, GPT-4, demonstrate superior selection performance, as indicated by ■ < ■, showing their effectiveness in selecting the historical event analogous to the input event. However, Llama2-7B shows limited improvement over random selection in generating historical analogies, suggesting that there is room for enhancing the general capabilities of LLMs in this domain.

Historical Analogies from Different Perspectives

Different individuals may describe the same event in various ways. We are interested in determining whether these differing perspectives influence the historical analogies generated from LLMs. To investigate this, we select several controversial events, manually create descriptions from different viewpoints, and utilize a self-reflection method based on ChatGPT to generate historical analogies. Figure 6 presents some typical cases. Our findings indicate that varying descriptions can indeed lead to different analogical outcomes. For instance, the Capitol Hill incident might be analogous to the Storming of the Bastille from the Republican Party's perspective or to the Reichstag fire from the Democrats' perspective. However, different descriptions may also produce the same analogies, since analogous events can also have diverse interpretations. Future research could focus on devel-

oping methods to identify and evaluate historical analogies based on diverse perspectives.

6 Conclusion

In this paper, we explore the concept of historical analogy and examine the ability of LLMs to acquire historical analogies for given events. We create an automatic multi-dimensional similarity metrics to fairly assess the quality of historical analogies, and perform numerous experiments with different models, which show that LLMs have the potential for historical analogies. In addition, we design an optimization method, self-reflection, which breaks from the stereotypes through multiple rounds of reflection and improves the historical analogical performance of the model.

Limitations

First, our evaluation mainly focuses on the accuracy of the analogous historical events, without assessing the reasons provided by the model due to the challenges in automatic evaluation of reasoning. Second, while our evaluation considers four specific dimensions to determine the correctness of historical analogies, it is important to note that in real-life contexts, additional factors such as gender, party affiliation, and motivation might also be considered, particularly by politicians. However, we believe that the evaluation of these additional dimensions could be automated through the application of our proposed evaluation methodology. Although we include historical analogies from various perspectives, assessing the rationality and applicability of these analogies across different perspectives remains challenging.

Ethics Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct.

Use of Human Annotations Evaluation on the identified historical analogies from LLMs is implemented by three annotators recruited by our institution. The construction team remains anonymous to the authors. We ensure that the privacy rights of all annotators are respected throughout the annotation process. All annotators are compensated above the local minimum wage and consent to the use of these historical analogies for research purposes, as described in our paper. The annotation details are shown in Appendix A.

Risks The analogy sets used in the experiment, including the popular and general sets, are derived from publicly accessible sources. We have reviewed these analogies to ensure they are free from socially harmful or toxic language. However, we cannot guarantee that they will not offend certain groups. Furthermore, evaluating historical analogies depends on common sense, and individuals from diverse backgrounds may have different perspectives. We use ChatGPT (OpenAI, 2022) to correct grammatical errors in this paper.

Acknowledgements

This work was supported by funding from Ant Group.

References

W Andrew Achenbaum. 1983. The making of an applied historian: Stage two. *The Public Historian*, 5(2):21–46.

AI-Meta. 2024. Llama 3 model card.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Robert Axelrod and Larissa Forster. 2017. How historical analogies in newspapers of five countries make sense of major events: 9/11, mumbai and tahrir square. *Research in Economics*, 71(1):8–19.

Paul Bartha. 2013. Analogy and analogical reasoning.

Bhavya Bhavya, Shradha Sehgal, Jinjun Xiong, and ChengXiang Zhai. 2024. AnaDE1.0: A novel data set for benchmarking analogy detection and extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1723–1737, St. Julian's, Malta. Association for Computational Linguistics.

Bhavya Bhavya, Jinjun Xiong, and ChengXiang Zhai. 2022. Analogy generation by prompting large language models: A case study of instructgpt. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 298–312, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Mario Bunge. 1981. Analogy between systems. *International Journal Of General System*, 7(4):221–223.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955, Dublin, Ireland. Association for Computational Linguistics.

Qing Chen, Wei Shuai, Jiyao Zhang, Zhida Sun, and Nan Cao. 2024. Beyond numbers: Creating analogies to enhance data comprehension and communication with generative ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Peter Conolly-Smith. 2009. "connecting the dots": Munich, iraq, and the lessons of history. *The History Teacher*, 43(1):31–51.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2094–2100, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zijian Ding, Arvind Srinivasan, Stephen MacNeil, and Joel Chan. 2023. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. *arXiv preprint arXiv:2302.12832*.

Fredrick J Dobney. 1974. " lessons" of the past: The use and misuse of history in american foreign policy.

Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63.

Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. 2020. Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, Online. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Djouaria Ghilani, Olivier Luminet, Hans-Peter Erb, Christine Flassbeck, Valérie Rosoux, Ismee Tames, and Olivier Klein. 2017. Looking forward to the past: An interdisciplinary discussion on the use of historical analogies and their effects. *Memory Studies*, 10(3):274–285.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Jo Guldi and David Armitage. 2014. *The history manifesto*. Cambridge University Press.

Keith J Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537, Singapore. Association for Computational Linguistics.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.

Sjoerd Keulen. 2023. Historical analogies: Functions, limitations and the correct use of historical analogies in applied history. *Journal of Applied History*, 5(2):111 – 131.

Yuen Foong Khong. 2020. *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of 1965*. Princeton University Press.

Jason Kirchenbauer and Caleb Barns. 2024. Hallucination reduction in large language models with retrieval-augmented generation using wikipedia knowledge.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Meg Parsons and Johanna Nalau. 2016. Historical analogies as tools in understanding transformation. *Global Environmental Change*, 38:82–96.

Karl Pearson. 1920. Notes on the history of correlation. *Biometrika*, 13(1):25–45.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.

Howard Schuman and Cheryl Rieger. 1992. Historical analogies, generational effects, and attitudes toward war. *American Sociological Review*, pages 315–326.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Charles Spearman. 1961. The proof and measurement of association between two things.

Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. 2024. Parallelparc: A scalable pipeline for generating natural-language analogies. *arXiv preprint arXiv:2403.01139*.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3547–3562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

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

Peter D Turney and Michael L Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60:251–278.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

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

Yifei Wang, Yuyang Wu, Zeming Wei, Stefanie Jegelka, and Yisen Wang. 2024. A theoretical understanding of self-correction through in-context alignment. *arXiv preprint arXiv:2405.18634*.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2022. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. Analobench: Benchmarking the identification of abstract and long-context analogies. *arXiv preprint arXiv:2402.12370*.

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Jankowski, Yanghua Xiao, and Deqing Yang. 2023a. Distilling script knowledge from large language models for constrained language planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4303–4325, Toronto, Canada. Association for Computational Linguistics.

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. 2023b. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2446–2460, Singapore. Association for Computational Linguistics.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2023c. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base. *arXiv preprint arXiv:2305.05994*.

Siyu Yuan, Cheng Jiayang, Lin Qiu, and Deqing Yang. 2024. Boosting scientific concepts understanding: Can analogy from teacher models empower student models? *arXiv preprint arXiv:2406.11375*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Crowd-sourcing Details

We have recruited a team of three undergraduates from the history department. To resolve conflicting annotations, we adopt a voting majority principle to determine the results. Each annotator is compensated at $8 per hour, which surpasses the local minimum wage. Screenshots of the instructions and interface for the annotation of historical analogies are shown in Figure 8.

B Resource of HISANALOY

We manually collect samples of popular analogies from web pages and papers related to historical analogies:

- https://psyche.co/guides/how-should-you-interpret-historical-analogies-in-the-popular-press
- https://scholars-stage.org/sino-american-competition-and-the-search-for-historical-analogies
- https://www.insidehighered.com/opinion/blogs/higher-ed-gamma/2024/03/08/use-and-misuse-historical-analogies
- https://www.sciencedirect.com/science/article/pii/S1090944316301636
- https://medium.com/@ella.ayalon/on-historical-analogies-3f253e52bfbc
- https://brill.com/view/journals/joah/5/2/article-p111_2.xml
- https://journals.sagepub.com/doi/abs/10.1177/1750698017701609
- https://aeon.co/essays/what-thucydides-really-thought-about-historical-analogies
- https://origins.osu.edu/history-news/historical-analogies-handle-care?
- https://www.emerald.com/insight/content/doi/10.1108/SSRP-03-2019-0020
- https://onlinelibrary.wiley.com/doi/epdf/10.1111/0162-895X.00145?saml_referrer
- https://www.washingtonpost.com/outlook/2021/05/03/historical-analogies-covid-fascism-mccarthyism/
- https://slate.com/news-and-politics/2014/07/pick-your-analogy-is-the-middle-east-today-more-like-world-war-i-the-cold-war-or-the-thirty-years-war.html
- https://launiusr.wordpress.com/2012/05/01/the-use-and-abuse-of-historical-analogs/
- https://www.usni.org/magazines/proceedings/2022/may/new-analogy-america-and-china-through-lens-pax-britannica

C Prompt Template

C.1 Prompt Template for Evaluation

The prompt template for GPT-4 to summarize the dimensions of the event is shown in List 1

Listing 1: Instruction templates for GPT-4 to summarize the dimensions of the event.

```
/* Task prompt */
You are an event summary robot. For the
event description input, please combine
your knowledge and summarize it into
four parts: topic, background, process
and result. The summary should be
concise, with each part consisting of
only one sentence and no more than 100
words.
/* Examples */
Input Event: September 11 attacks
{Description of September 11 attacks}
Output: {Topic, Background, Process and
Result of September 11 attacks}
/* Test Data */
Input Event: COVID-19 pandemic
{Description of COVID-19 pandemic}
Output: Topic: global health crises
caused by viruses, resulting in
widespread illness and significant
mortality...
```

The prompt template for abstract similarity scoring is shown in List 2.

Listing 2: Instruction templates for GPT-4 to score the abstract similarity for the given two events.

```
/* Task prompt */
You are a sentence-level analogy-scoring
 robot. Given the two descriptions,
please judge the quality of the analogy
and give it a score (1-4). The quality
of an analogy only focuses on the
abstract-level similarity, rather than
the literal
similarity.
/* Evaluation Criteria */
1 point: The two descriptions belong to
completely different
topics or fields, have no connection,
and cannot be compared.
```

```
2 points: The two descriptions belong to
 the same general theme, but the
specific situation or aspect they
express is significantly different.
3 points: The two descriptions belong to
 the same topic and
express similar general situations, but
differ somewhat in
details or focus.
4 points: The two descriptions pertain
to the same topic, with
the general situation expressed being
highly similar, and the
concepts and key points closely
overlapping.
/* Test Data */
{Description of COVID-19 pandemic}
{Description of Spanish pandemic}
Score: 3
```

C.2 Prompt Template for Methods

The prompt template of each method is given in List 3

Listing 3: Instruction templates of different methods in historical analogy generation.

```
Direct Generation:
You are a historical analogy bot. For
input events, your goal is to find the
most appropriate event to use for
analogizing with the input.
/* Examples */
Input Event:
coronavirus pandemic: {Description of
coronavirus pandemic}
Historical Analogies Events:
Spanish flu

Candidate Proposals in Two-Stage Method:
You are a historical analogy candidate
proposals robot. For input events,
please consider the summary, background,
 process and results, output n
historical events that are similar in
many aspects above, and return them in
list format.
/* Examples */
Input Event:
coronavirus pandemic: {Description of
coronavirus pandemic}
The 10 historical events that are
similar with input:
["Spanish flu pandemic","Asian flu
pandemic","Hong Kong flu pandemic","AIDS
 pandemic","Ebola outbreak in West
Africa","SARS outbreak","H1N1 influenza
pandemic","MERS outbreak","Cholera
pandemics","Plague pandemics"]

Selection in Two-Stage Method:
You are an analogy robot. For the input
event and the historical event used for
selection, your goal is to find the best
 event that can be used for analogies.
/* Examples */
Input Event:
coronavirus pandemic: {Description of
coronavirus pandemic}
Optional Historical Events:
2022 South Asian floods: {Description of
 2022 South Asian floods}
Croydon typhoid outbreak of 1937: {
Description of Croydon typhoid outbreak}
Spanish flu: {Description of Spanish flu
}
Cold War: {Description of Cold War}
Among the options, the most appropriate
one to use as an analogy for coronavirus
 pandemic is Spanish flu

Candidate Proposals in Self-Reflection:
You're a robot for proposing historical
analogies events. Historical Analogy is
comparsion of a known past event or
person with a contemporary but
unfamiliar event or person in order to
identify common aspects between the two.
 For input events, please consider the
summary, background, process and results
, and output 5 historical events that
are similar in many aspects above, and
return them in list format. If there is
any reflection, please modify the
recommended events based on the
reflection.
/* Examples */
Input Event:
coronavirus pandemic: {Description of
coronavirus pandemic}
The 10 historical events that are
similar with input:
["Spanish flu pandemic","Asian flu
pandemic","Hong Kong flu pandemic","AIDS
 pandemic","Ebola outbreak in West
Africa","SARS outbreak","H1N1 influenza
pandemic","MERS outbreak","Cholera
pandemics","Plague pandemics"]

Selection in Self-Reflection:
You are a historical analogy reflection
robot. Historical Analogy is comparsion
of a known past event or person with a
contemporary but unfamiliar event or
person in order to identify common
aspects between the two. For the input
event and the candidate event set,
please make a comparison, reflect on the
 shortcomings of the candidate set, and
make suggestions for obtaining a better
analogous candidate set. Suggestions
should be succinct and concise, with a
single sentence indicating the direction
 of change for the candidate set.
/* Examples */
===== Case 1
Input Event:
coronavirus pandemic: {Description of
coronavirus pandemic}
Optional Historical Events:
2022 South Asian floods: {Description of
 2022 South Asian floods}
Croydon typhoid outbreak of 1937: {
Description of Croydon typhoid outbreak}
Thought:
The coronavirus pandemic is a global
epidemic, so the themes of 2022 South
Asian floods are completely different.
The Croydon typhoid outbreak of 1937 was
```

Method	$\text{Topic}_{\text{Abs}}$	$\text{Topic}_{\text{Lit}}$	$\text{Topic}_{\text{All}}$	$\text{Background}_{\text{Abs}}$	$\text{Background}_{\text{Lit}}$	$\text{Background}_{\text{All}}$
GPT-3.5-Turbo						
Direct Re.	[3.16, 3.43]	[0.17, 0.20]	[0.48, 0.57]	[2.86, 3.14]	[0.11, 0.15]	[0.63, 0.71]
Two-stage Re.	[2.78, 3.09]	[0.16, 0.21]	[0.47, 0.56]	[2.54, 2.86]	[0.12, 0.17]	[0.54, 0.63]
Direct Gen.	[2.74, 3.04]	[0.12, 0.14]	[0.59, 0.66]	[2.54, 2.83]	[0.09, 0.11]	[0.61, 0.69]
Two-stage Gen.	[3.07, 3.34]	[0.17, 0.23]	[0.52, 0.62]	[2.68, 2.97]	[0.13, 0.19]	[0.58, 0.67]
Summarizing	[3.39, 3.61]	[0.16, 0.21]	[0.59, 0.68]	[2.89, 3.16]	[0.11, 0.16]	[0.64, 0.72]
Self-reflection	[3.42, 3.62]	[0.16, 0.19]	[0.57, 0.65]	[3.10, 3.33]	[0.11, 0.13]	[0.69, 0.77]
Llama3.1-8B						
Direct Re.	[3.16, 3.43]	[0.17, 0.20]	[0.48, 0.57]	[2.86, 3.14]	[0.11, 0.15]	[0.63, 0.71]
Two-stage Re.	[2.97, 3.26]	[0.14, 0.17]	[0.61, 0.68]	[2.62, 2.94]	[0.10, 0.12]	[0.67, 0.74]
Direct Gen.	[3.32, 3.56]	[0.17, 0.21]	[0.55, 0.65]	[2.94, 3.21]	[0.12, 0.17]	[0.63, 0.72]
Two-stage Gen.	[3.07, 3.36]	[0.13, 0.16]	[0.58, 0.67]	[2.77, 3.05]	[0.10, 0.12]	[0.65, 0.73]
Summarizing	[3.27, 3.54]	[0.15, 0.18]	[0.56, 0.65]	[2.98, 3.25]	[0.10, 0.12]	[0.68, 0.75]
Self-reflection	[3.36, 3.57]	[0.16, 0.20]	[0.58, 0.67]	[2.95, 3.22]	[0.11, 0.15]	[0.66, 0.74]

Table 4: Confidence intervals of experimental results on General Analogies, including Topic and Background dimensions.

Method	$\text{Process}_{\text{Abs}}$	$\text{Process}_{\text{Lit}}$	$\text{Process}_{\text{All}}$	$\text{Result}_{\text{Abs}}$	$\text{Result}_{\text{Lit}}$	$\text{Result}_{\text{All}}$	MDS
GPT-3.5-Turbo							
Direct Re.	[2.84, 3.11]	[0.10, 0.12]	[0.66, 0.74]	[2.86, 3.15]	[0.11, 0.14]	[0.62, 0.70]	[3.50, 3.78]
Two-stage Re.	[2.46, 2.81]	[0.10, 0.14]	[0.55, 0.64]	[2.59, 2.92]	[0.12, 0.16]	[0.54, 0.63]	[3.01, 3.39]
Direct Gen.	[2.49, 2.76]	[0.08, 0.10]	[0.65, 0.72]	[2.66, 2.93]	[0.09, 0.11]	[0.64, 0.71]	[3.54, 3.84]
Two-stage Gen.	[2.87, 3.15]	[0.11, 0.15]	[0.65, 0.74]	[2.87, 3.13]	[0.12, 0.16]	[0.63, 0.70]	[3.45, 3.82]
Summarizing	[2.98, 3.24]	[0.10, 0.14]	[0.70, 0.78]	[2.94, 3.21]	[0.12, 0.15]	[0.63, 0.71]	[3.65, 4.00]
Self-reflection	[3.04, 3.29]	[0.10, 0.12]	[0.71, 0.79]	[3.01, 3.26]	[0.11, 0.14]	[0.66, 0.73]	[3.79, 4.06]
Llama3.1-8B							
Direct Re.	[2.84, 3.11]	[0.10, 0.12]	[0.66, 0.74]	[2.86, 3.15]	[0.11, 0.14]	[0.62, 0.70]	[3.50, 3.78]
Two-stage Re.	[2.68, 2.96]	[0.09, 0.10]	[0.59, 0.66]	[2.59, 2.87]	[0.10, 0.12]	[0.54, 0.62]	[3.46, 3.75]
Direct Gen.	[2.91, 3.18]	[0.11, 0.15]	[0.68, 0.76]	[2.87, 3.14]	[0.12, 0.16]	[0.62, 0.70]	[3.54, 3.90]
Two-stage Gen.	[2.78, 3.05]	[0.09, 0.11]	[0.69, 0.76]	[2.66, 2.94]	[0.10, 0.12]	[0.62, 0.69]	[3.61, 3.91]
Summarizing	[2.88, 3.16]	[0.09, 0.11]	[0.71, 0.78]	[2.95, 3.23]	[0.11, 0.14]	[0.66, 0.73]	[3.75, 4.03]
Self-reflection	[2.94, 3.21]	[0.10, 0.13]	[0.71, 0.79]	[2.99, 3.27]	[0.11, 0.14]	[0.66, 0.74]	[3.75, 4.07]

Table 5: Confidence intervals of experimental results on General Analogies, including Process, Result, and the total score of the multi-dimensional similarity metric.

```
 smaller in scope, while the coronavirus
 pandemic were global influenza
pandemics, so there is no suitable
analogy here and I need to reflect.
Reflection:
Candidate events need to focus on the
epidemic and its impact on a global
scale.
===== Case 2
Input Event:
coronavirus pandemic: {Description of
coronavirus pandemic}
Optional Historical Event:
Spanish flu: {Description of Spanish flu
}
Cold War: {Description of Cold War}
Thought:
The Cold War has nothing to do with the
epidemic. The Spanish flu is also an
epidemic and has had a great impact in
Europe, so it is a qualified analogy for
 the coronavirus pandemic.
Final Answer:
Spanish flu
```

D Stability Analysis of Experimental Results

To demonstrate the stability and reliability of our experimental results, we use bootstrapping (Tibshirani and Efron, 1993) with 1000 resamples on general analogy to compute the 95% confidence intervals for each metric. Our results are shown in Tables 4 and Table 5. The confidence intervals align with the main experiment (Table 1) and show only about a 10% fluctuation. Therefore, our results are stable and trustworthy.

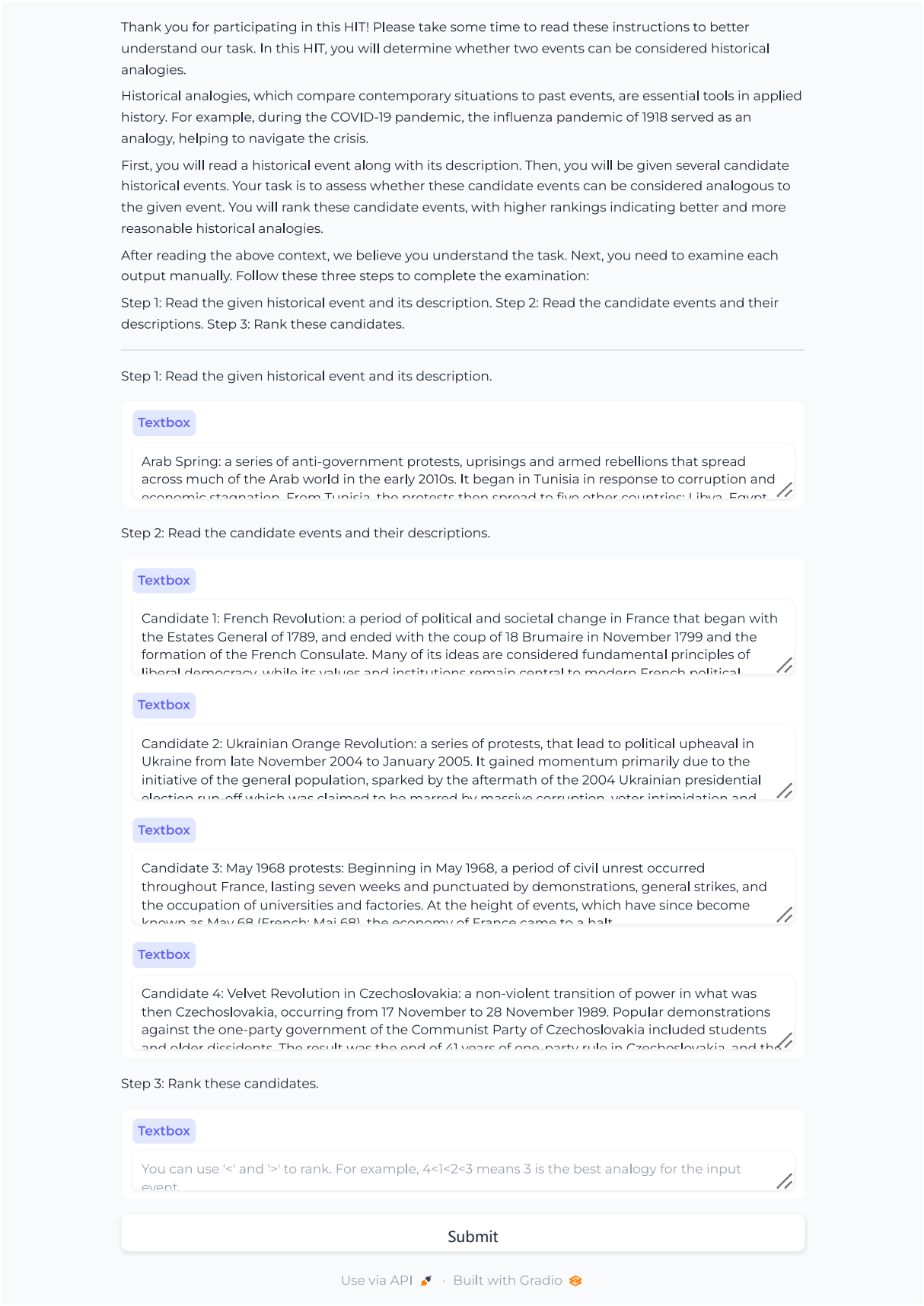

Figure 8: The screenshots of the instructions and interface for historical analogy manual annotation.