VHE: A New Dataset for Event Extraction from Vietnamese Historical Texts

Truc Hoang[1,2], **Long Nguyen**[1,2*], **Dien Dinh**[1,2]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
hoangthuytruc@gmail.com, {nhblong,ddien}@fit.hcmus.edu.vn

Abstract

The event extraction (EE) task, which detects occurrences of specified event types and extracts corresponding event arguments from unstructured data, is crucial for the study of history. However, most existing datasets are not available in Vietnamese. Our work aims to address this data scarcity problem for EE models. In this paper, we introduce a new dataset - Vietnamese Historical Events (VHE) [1] for the EE task in the context of Vietnamese historical documents - a domain characterized by unique linguistic structures, historical references, and cultural nuances. Specifically, our dataset features 35 event types, 9 entity types, and 11 argument roles that pertain to historical events from the Hong Bang dynasty (2879 BC) to the Later Le dynasty in the sixteenth century. To create this dataset, we utilize large language models (LLMs) as data annotators and validate their results through human review. We then conduct experiments on the VHE dataset using both current state-of-the-art event extraction (EE) systems and LLMs, including closed-source models (e.g., GPTs, Gemini) and open-source models (e.g., LLaMA, Phi, Qwen, Gemma). The results reveal their poor performance on historical texts and underscore the numerous challenges faced by existing EE systems, such as the evolution of word meanings over time and ambiguities in sentence structures.

1 Introduction

History is an important field of study that plays a vital role in shaping the identities, values, and futures of individuals and societies (Boros et al., 2022). The proliferation of digital historical documents enables researchers to collect and study information more easily, but it also presents a significant challenge as history continues to unfold and becomes increasingly vast. While the goal of event extraction is to extract organized event knowledge from unstructured text, it also improves the efficiency of information acquisition. Generally, the event extraction task can be decomposed into two subtasks: Event Detection (ED) and Event Argument Extraction (EAE) (Li et al., 2022). The ED task aims to detect event trigger words and classify them into event types, while the EAE task identifies arguments involved in the event and their corresponding roles. Figure 1 shows an example of the event extraction task.

Since event extraction is fundamental to various natural language processing applications (Li et al., 2022), it has attracted many research attention in recent years (Yarmohammadi et al., 2021; Hsu et al., 2022; Peng et al., 2023), building on available datasets such as ACE 2005 (Walker et al., 2006), FewEvent (Deng et al., 2020), MAVEN (Wang et al., 2020), RAMS (Li et al., 2021). However, most existing datasets primarily support high-resource languages like English and Chinese, limiting further research on low-resource languages like Vietnamese. Only one Vietnamese dataset (Nguyen et al., 2024) is available, having been released just a few months ago. Additionally, documents in the existing datasets are typically derived from recent articles, where the use of words differ from their historical usage. Currently, there is only one English dataset (Lai et al., 2021), which focuses on the history domain.

In this study, we introduce VHE, a novel dataset for event extraction from Vietnamese historical texts. VHE supports three tasks: event extraction, event detection, and event argument extraction. We first develop an event schema tailored for Vietnamese historical events. Next, we design prompts to automatically annotate the dataset using large language models (LLMs), including GPT-3.5 and GPT-4o. These annotations are subsequently reviewed by humans to ensure high accuracy and

[*] Corresponding author.

[1] https://github.com/hoangthuytruc/vhe-dataset

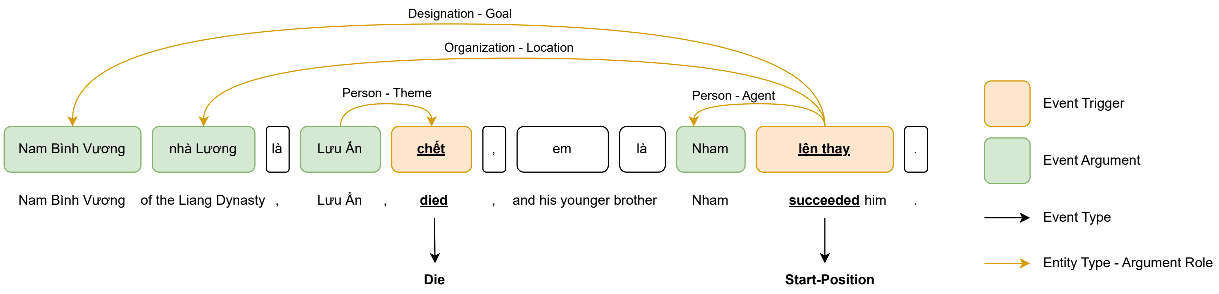

Figure 1: An example of event extraction in the text. It can extract two types of events. The first is the *Die* event, triggered by the keyword *"chết"* with an argument role of *Theme*. The second is the *Start-Position* event, triggered by the keyword *"lên thay"* with three argument roles of *Theme*, *Location*, and *Goal*.

quality. As a result, our dataset includes *4,114* instances containing *5,213* events and *7,423* event arguments. Finally, we evaluate state-of-the-art event extraction models on VHE, including both closed-source and open-source LLMs. Our experiments reveal a significant gap between human performance and that of the models in extracting events from Vietnamese historical texts, highlighting the need for further research in this area.

2 Background and Related Work

2.1 Event Extraction

Event extraction aims to detect occurrences of specified types and extract corresponding event arguments from unstructured data input. The ACE 2005 program (Consortium, 2005) defines an event schema with terminologies that have been widely adopted in event extraction. We outline the key terminologies as follows:

- An **event** is a specific occurrence involving participants
- **Event extent** is a sentence within which an event is expressed.
- **Event trigger** is a word or a phrase that mostly clearly expresses the occurrence of the event.
- **Event argument** are entities that are part of the event.
- **Argument role** is the relationship between an event and its arguments.

Based on these terminologies, Ahn (2006) proposed dividing event extraction into the subtasks of trigger detection, trigger classification, argument detection, and argument classification. Specifically, trigger identification and trigger classification can be grouped under the event detection task, while argument identification and argument classification fall under the event argument extraction task. **Trigger identification** involves detecting event triggers within an event extent, while **Trigger classification** assigns these identified triggers to specific event types. Similarly, **Argument identification** is to identify all arguments associated with an event type, while **Argument classification** is responsible for assigning these arguments to their corresponding roles. In this paper, we inherit all the above-mentioned settings in both dataset construction and model evaluation.

2.2 Related Work

There are numerous EE datasets across various domains, including the Wikipedia domain (Deng et al., 2021; Li et al., 2021; Pouran Ben Veyseh et al., 2022) and the news domain (Ebner et al., 2020; Tong et al., 2022; Nguyen et al., 2024). Recently, some works have focused on the general domain to encompass a broader range of event types (Deng et al., 2020; Wang et al., 2020; Parekh et al., 2023). In specific domains, datasets like Genia2011 (Kim et al., 2011), MLEE (Pyysalo et al., 2012), and Genia2013 (Kim et al., 2013) have been proposed for biomedical research; CASIE (Satyapanich et al., 2020) for cybersecurity; PHEE (Sun et al., 2022) for pharmacovigilance; EDT (Zhou et al., 2021) for stocks; IndiaPoliceEvent (Halterman et al., 2021) for political events; ChFinAnn (Zheng et al., 2019) for financial data; and BRAD (Lai et al., 2021) for historical events.

3 Dataset Creation Process

Our dataset creation process, illustrated in Figure 2, consists of four main steps: (1) data preparation, (2) event schema construction, (3) data anno-

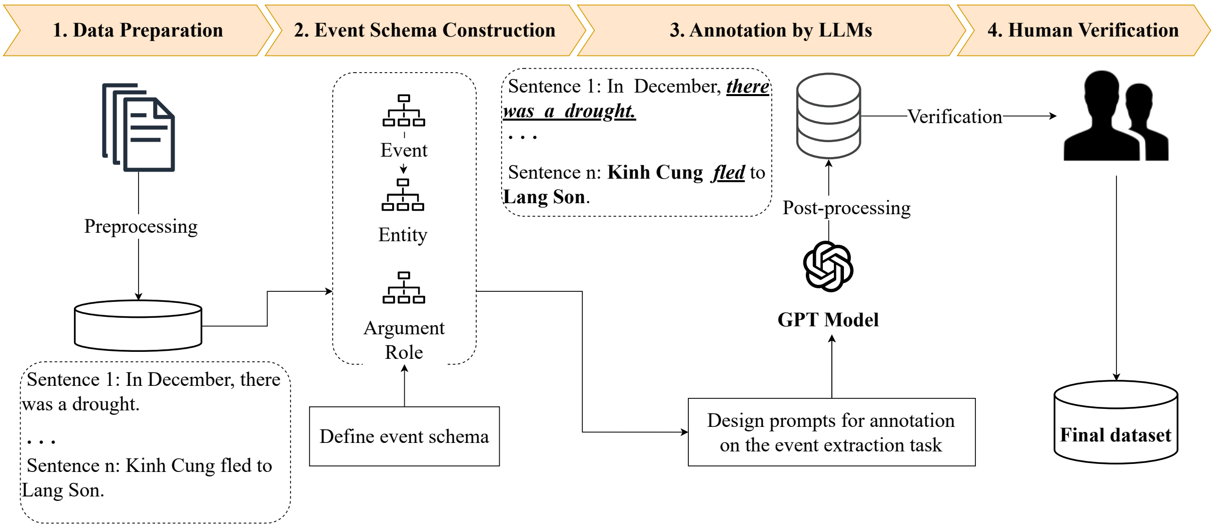
1. Data Preparation
2. Event Schema Construction
3. Annotation by LLMs
4. Human Verification
Preprocessing
Sentence 1: In December, there was a drought.
. . .
Sentence n: Kinh Cung fled to Lang Son.
Event
Entity
Argument Role
Define event schema
Sentence 1: In December, there was a drought.
. . .
Sentence n: Kinh Cung fled to Lang Son.
Post-processing
GPT Model
Verification
Design prompts for annotation on the event extraction task
Final dataset

Figure 2: Our dataset creation process. It consists of four main steps: (1) data preparation, (2) event schema construction, (3) data annotation by LLMs, and (4) human verification, corresponding to four subsections: 3.1, 3.2, 3.3, and 3.4.

tation by LLMs, and (4) human verification. We first describe each of these steps and then provided statistics for the final dataset.

3.1 Data Preparation

We use The Complete Annals of Đại Việt, the oldest official historical text of Vietnam. This book, compiled into 23 volumes, records the history of Vietnam from the reign of King Duong Vuong (2879 BC) to the reign of Le Gia Tong of the Later Le Dynasty (1675). Firstly, text is extracted from the document files, and headers, footers, footnotes, and author comments are removed. We then use VnCoreNLP (Vu et al., 2018) to split texts into sentences and normalize them (removing duplicate spaces, correcting diacritics, etc.). Finally, we obtained a total of 21,001 sentences for the entire dataset.

3.2 Event Schema Construction

The event schema used by the existing datasets is inconsistent because of discrepant assumption about data, different preprocessing steps and the use of external resources (Huang et al., 2024) while extracting includes several tasks such as event detection, event argument extraction, and role labeling. (Lai et al., 2021). Hence, we aim to construct a new event schema with reusable and extendable capabilities that adapt to context.

To begin with, we use the widely adopted event definitions from ACE for event types and the common entity types and semantic roles[2] for argument roles as follows:

Event Types We utilized 33 event types along with an additional event type *Nature* which includes two subtypes: *Natural-Disaster* and *Natural-Phenomenon*, as suggested by our experts. A *Natural-Disaster* event occurs when a natural disaster causes damage to people and property or destroys architectural structures (e.g., earthquake, drought) while a *Natural-Phenomenon* event occurs when an unusual natural phenomenon appears without causing any impact on humans or other entities (e.g., solar eclipse). Table 8 provides the examples of event types in our dataset.

Entity Types 9 entity types were selected from the Vietnamese NER tagset,[3] including *Person (PER)*, *Organization (ORG)*, *Location (LOC)*, *Datetime (DTM)*, *Designation (DES)*, *Measure (MEA)*, *Terminology (TRM)*, and *Miscellaneous (MISC)* for other entities. Table 6 provides the definitions of entity types used in our dataset.

Argument Role Types We adopted 11 common argument roles, including *Agent*, *Experiencer*, *Force*, *Theme*, *Content*, *Instrument*, *Beneficiary*, *Source*, *Goal*, *Temporal*, and *Location*. Table 7 provides the definitions of argument role types used in our dataset.

[2] https://web.stanford.edu/~jurafsky/slp3/21.pdf

[3] https://www.clc.hcmus.edu.vn/wp-content/uploads/2016/01/CLC_VN_NER-Tagset.pdf

Entity Types	Percentage (%)
PER	52.0
DTM	19.0
LOC	11.0
DES	10.0
ORG	3.0

Argument Role Types	Percentage (%)
Theme	31.0
Agent	28.0
Temporal	19.0
Content	10.0
Location	5.0

Table 1: Five top-level entity and argument role types in the VHE dataset.

Depending on the context of the text, all these types of entities and argument roles are reused across all event types in our dataset. Appendix C shows more details of the event scheme in VHE.

3.3 Annotation by LLMs

To leverage the information extraction capabilities of LLMs (Ma et al., 2023; Li et al., 2023; Han et al., 2023) and minimize the time required for the annotation process, we designed prompts to automatically annotate events using GPT-4o and verified the results through human review to create a gold dataset.

Based on the predefined event schema, the prompts include the categories of event, entity, and argument role, but do not provide examples. The entire dataset was annotated by two GPT models, including GPT-3.5-turbo and GPT-4o-mini (Brown et al., 2020). We then filtered out all results that did not conform to the event schema or were in the wrong format. As a result, the dataset contains approximately 15,000 instances in total.

3.4 Human Verification

The review process involved two native speakers who were not experts. Initially, they were provided with annotation guidelines and examples for each event type. Each annotator then tested a subset of events to ensure a clear understanding of the guidelines. We subsequently collaborated to discuss and resolve any conflicts, ultimately reaching a consensus on the final dataset.

As the event annotation is complicated, we separated the dataset into 2 subsets to reduce information overload for reviewers. The first subset contained 3,153 events that were assigned the same event type by both GPT models, accounting for about 20% of the dataset. The second subset comprised about 80% of the events annotated by GPT-4o-mini. Initially, reviewers examined the first subset to gain a better understanding of the dataset's context, working independently. Subsequently, they collaborated to review the second subset and produce the gold dataset.

4 Dataset Quality Assessment

To validate the quality of the dataset, we randomly sampled 150 instances from the gold dataset and removed their labels. We then recruited two trained undergraduate students to manually annotate these samples. We utilize Cohen's Kappa (Cohen, 1960) to calculate the inter-annotator agreement (IAA) score between the two annotators for each subtask. The scores obtained were 82.0% for trigger identification, 76.5% for trigger classification, 60.0% for argument identification, and 58.0% for argument classification. Notably, The human performance average scores align with the IAA scores for each subtask. Although the inner-annotator agreement scores of the event argument extraction task are slightly lower, remains within an acceptable range, affirming the consistency and reliability of our dataset.

5 Dataset Analyses

Figure 3 illustrated the distribution of event types in our dataset. We observe that most events from this era focus on three main event types: *Start-Position*, *Attack*, and *Die*. Additionally, the *Justice* event types have relatively few occurrences, and there are no events related to the *Declare-Bankruptcy* event type. Therefore, the inherent data imbalance problem also exists in our dataset. Moreover, we identified ambiguity within VHE, which underscores the need for EE models to address this imbalance and uncover cross-sentence relationships.

Table 1 shows the top five entity and argument types and their proportions in our dataset. The highest proportions include PER (52%) for entity types, and Theme (31%), Agent (28%) for argument role types. Additionally, the argument DTM

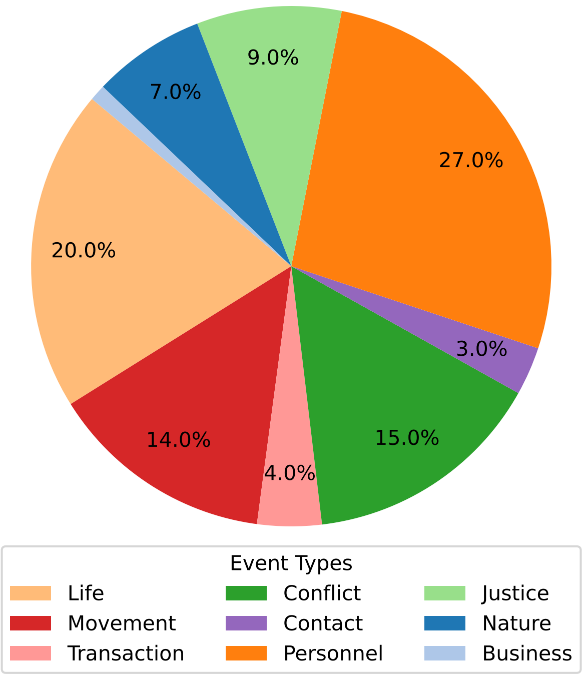

Figure 3: Distribution of event types in the VHE dataset.

and LOC account for approximately 25-30% of the dataset. These proportions are consistent with the most common event types in the dataset. To maintain the real-word distribution in VHE, we do not apply data augmentation or balancing during dataset construction.

6 Experiments

In this section, we first describe our experimental settings, including the models, various types of prompting, and the evaluation metrics used. We then present the performance of LLMs and state-of-the-art event extraction models on our dataset. We evaluate three groups of models: (1) closed-source LLMs, (2) open-source LLMs, and (3) end-to-end models. Finally, we analyze common errors that influenced the evaluation outcomes.

6.1 Experimental Settings

Models To gain a better understanding of how current models extract events from Vietnamese history texts, we evaluate three groups of models on our dataset: (1) closed-source LLMs, (2) open-source LLMs, and (3) end-to-end models. Since our dataset is in Vietnamese, we consider choosing LLMs that support multilingual capabilities. We use **GPT-3.5-turbo** and **GPT-4o** (Brown et al., 2020) as closed-source LLMs, while **Llama-3.1-8B-Instruct** (Dubey et al., 2024), **Gemma-2-9b-it** (Team et al., 2024), **Phi-3.5-mini-instruct** (Abdin et al., 2024), and **Qwen-2-7B-Instruct** (Yang et al., 2024) are considered for open-source LLMs. For end-to-end EE models, we adopt the pre-trained EE model provided by OmniEvent (Peng et al., 2023), which implements the **Ses2Seq** paradigm (Sutskever et al., 2014) using **mT5** (Xue et al., 2021) as the base model.

Prompting In our experiments, the prompts were designed to perform both event detection and event argument extraction tasks simultaneously. To guide LLMs in generating responses within the scope of predefined event types, we included specific context within the prompts. Each model was evaluated using two prompting techniques: zero-shot and few-shot (2-shot and 4-shot). However, instruction-tuned LLMs (e.g., LLaMA, Gemma, Phi, Qwen) have shown limited robustness to variations in instruction phrasing (Sun et al., 2023). Consequently, we excluded zero-shot evaluations for these models. Appendix B provides an illustration of the prompts used in the evaluation process.

Evaluation Metrics To evaluate the event extraction task, most EE systems and datasets use precision, recall, and F1 scores as key evaluation metrics (Sheng et al., 2021; Yang et al., 2019; Chinchor, 1992). Due to the complexity of event extraction, these metrics are applied independently to each subtask. We report F1 scores for four subtasks: trigger identification, trigger classification, argument identification, and argument role classification. Appendix A provides additional results on our dataset, including all the detailed scores.

6.2 Results

Table 2 presents the performance of the models across four subtasks: trigger identification, trigger classification, argument identification, and argument classification. It is noted that due to the cost of running LLMs, we evaluate closed-source LLMs on a subset of our dataset, which includes 1,300 instances. In contrast, open-source LLMs and the end-to-end EE model are evaluated on the entire dataset.

End-to-End Models vs. LLMs From table 2, it can be seen that the end-to-end model performs poorly on the VHE dataset. Almost all sub-tasks of event extraction achieve less than 20.0 F1, with the event argument extraction task reaching only about 2.0 F1. One reason for this poor performance is

Group	Model	TI	TC	AI	AC
End-to-end models	Seq2Seq + mT5	17.13	6.42	2.03	0.35
Open-source LLMs	Llama-3.1-8B-Instruct (2-shot)	43.37	27.49	21.49	14.68
	Llama-3.1-8B-Instruct (4-shot)	42.57	28.04	**25.53**	17.36
	Gemma-2-9b-it (2-shot)	**48.53**	33.49	20.09	16.86
	Gemma-2-9b-it (4-shot)	47.20	**34.25**	24.39	**19.92**
	Phi-3.5-mini-instruct (2-shot)	17.34	9.29	2.74	1.87
	Phi-3.5-mini-instruct (4-shot)	12.08	6.79	1.56	1.14
	Qwen-2-7B-Instruct (2-shot)	29.36	16.10	10.88	6.12
	Qwen-2-7B-Instruct (4-shot)	19.68	12.47	5.60	3.67
Closed-source LLMs	Gemini-1.5-flash (zero-shot)	38.04	32.03	13.45	7.86
	Gemini-1.5-flash (2-shot)	36.45	30.88	14.43	9.91
	Gemini-1.5-flash (4-shot)	35.20	29.10	13.99	9.46
	GPT-3.5-turbo (zero-shot)	24.45	16.55	4.47	3.51
	GPT-3.5-turbo (2-shot)	28.91	19.25	4.74	3.82
	GPT-3.5-turbo (4-shot)	27.01	18.26	4.93	4.01
	GPT-4o (zero-shot)	18.85	17.39	4.64	3.16
	GPT-4o (2-shot)	**39.11**	**33.85**	13.81	10.80
	GPT-4o (4-shot)	35.18	30.90	**14.53**	**11.18**
Human	Average	75.85	67.97	58.15	41.32
	Inter-Annotator Agreement	82.00	76.50	60.00	58.00

Table 2: F1 scores of the models for four subtasks—Trigger Identification (TI), Trigger Classification (TC), Argument Identification (AI), and Argument Classification (AC)—on our dataset. We also present the average scores from human annotators and the inter-annotator agreement.

that the model has not been trained on any Vietnamese datasets except for the mT5 base model.

Open-source LLMs vs. Closed-source LLMs For open-source LLMs, Gemma-2-9b-it outperforms other models in TI (48.5 F1) and TC (34.2 F1), and its gains in AI (24.3 F1) and AC (19.9 F1) in the 4-shot setting suggest a stronger ability to leverage additional context. In contrast, both Phi-3.5-mini-instruct and Qwen-2-7B-Instruct show declining performance with an increasing number of shots, indicating a potential struggle with handling more contextual information. For example, the highest TI (17.34 F1) and TC (9.2 F1) for Phi-3.5-mini-instruct and the highest TI (29.3 F1) and TC (16.1 F1) for Qwen-2-7B-Instruct are observed under the 2-shot setting.

For closed-source LLMs, GPT-4o (2-shot) demonstrates the best performance in TI (39.1 F1) and TC (33.8 F1) when compared to GPT-3.5-turbo, while Gemini-1.5-flash excels in the zero-shot setting, particularly in TI (38.0 F1) and TC (32.03 F1), outperforming other models in this context.

Overall, most models perform consistently well in the 2-shot setting, though their performance doesn't scale significantly with more shots. Open-source models might be more adaptable for specific use cases where control and customization are crucial, while closed-source models tend to deliver higher performance, especially in scenarios with minimal or no additional context.

Models vs. Human Performance Across all metrics, human performance vastly outstrips that of both open-source and closed-source models. The closest models achieve less than 30% of human performance in TI (75.8 F1) and TC

(67.9 F1), with even larger gaps in AI and AC. Among the models, Gemma-2-9b-it (open-source) and GPT-4o (closed-source) achieve the highest scores, but they still fall far short of human-level accuracy, particularly in more nuanced tasks like Argument Identification and Classification.

Summary Despite advances in model capabilities, a substantial gap remains between machine performance and human expertise. Most models performed better in trigger identification and classification than in argument identification and classification. Notably, there is a significant gap between the event detection and event argument extraction tasks. This highlights numerous research opportunities for future work on the VHE dataset. Appendix A show details of evaluation results.

6.3 Analyses

Through the manual checking, we find that ther errors mainly inlude:

Span Error Since LLMs generate human-like responses, they often extract event triggers and arguments that are longer than those found in the gold dataset. For instance, in the sentence *"Sai quân đánh úp phá được tướng Tây đạo nguỵ là quận Nhai, quận Cao ở Nhật Chiêu thuộc Bạch Hạc bắt được 40 chiếc thuyền và 7 con voi. (The dispatched troops launched a surprise attack and defeated the Western Route rebel generals, Quận Nhai and Quận Cao, at Nhật Chiêu in Bạch Hạc, capturing 40 boats and 7 elephants)"*, the event trigger *"đánh úp (surprise attack)"* is sufficient, rather than *"đánh úp phá (surprise attack and defeated)"*. Additionally, LLMs have also automatically rephrase sentence which cause a failure of event trigger. For example, in the sentence *"Tháng 11, cho Nguyễn Danh Thế kiêm chức Đô ngự sử. (In November, Nguyễn Danh Thế was concurrently appointed to the position of Chief Censor.)"*, the phrase *"cho kiêm chức (appointed)"* was assigned to the event trigger while the entity *"Nguyễn Danh Thế"* was automatically omitted.

Linguistic Structures The dataset is derived from the oldest historical texts, which employ numerous linguistic structures that differ from those found in modern texts. Many subjects, as well as cross-references, are implied rather than explicitly stated, leading to ambiguities in meaning. For example, in the sentence *"Hôm ấy, Hữu tướng Hoàng Đình Ái sai thuộc tướng đánh bắt được, đem chém, bắt được 4 tên đồ đảng giải đến cửa dinh, cũng chém cả. (That day, the Right General Hoàng Đình Ái ordered his subordinate officers to attack and capture the enemy, who was then executed. Four members of the rebel group were also captured and brought to the headquarters, where they were all executed.)"*, the event trigger *"chém (executed)"* activates the *Execute* event in which the entity *the enemy*, affected by the event, is omitted and the entity *4 tên đồ đảng (Four members of the rebel group)* was assigned to an argument role of this event instead.

Entity vs. Event Argument Confusion There might be confusion between what constitutes an entity in NER and an event argument in event extraction tasks. For example, the argument mention *"chùa Thiên Quang, Thiên Đức (Thiên Quang, Thiên Đức pagodas)"* is automatically interpreted as *"chùa Thiên Quang (Thiên Quang pagoda)"* and *"chùa Thiên Đức (Thiên Đức pagoda)"*. Moreover, in historical texts, entities might be ambiguous or outdated, leading to challenges in accurate argument annotation.

Error Types In the post-processing of LLM-annotated events, we identify four types of errors related to event types, entities, and argument roles: Incorrect types, Undefined types, Incorrect format, and Other errors, which include issues like unannotated spans, unexpected information, and irrelevant context.

7 Conclusion

In this paper, we propose VHE, a new event extraction dataset focused on historical texts in Vietnamese. We conduct a thorough evaluation of state-of-the-art end-to-end model as well as LLMs on VHE. The results indicate that the event extraction from historical texts remains challenging, and VHE may facilitate further research in this area.

In the future, we intend to extend our work in several ways. First, we plan to enlarge our dataset with additional annotated documents. Second, we aim to expand the event schema to include event relations. Third, we will develop an end-to-end model for Vietnamese historical events.

Limitations

In this work, we make efforts to reduce the gap between high-resource and low-source languages in the event extraction. However, due to limitations in human resources, it is challenging for us to obtain a larger amount of labeled data. Additionally, there is a possibility that some events annotated by LLMs may be overlooked. Furthermore, as history is a complex domain, our knowledge may not encompass all taggable events from the dataset. We will continue to maintain and update our proposed dataset for future research.

Acknowledgements

We thank Xanh Ho for her invaluable support which have greatly contributed to this work. Her expertise and guidance were instrumental in shaping the direction of this research.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, and et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Emanuela Boros, Luis Adrián Cabrera-Diego, and Antoine Doucet. 2022. Experimenting with unsupervised multilingual event detection in historical newspapers. In *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries: 24th International Conference on Asian Digital Libraries, ICADL 2022, Hanoi, Vietnam, November 30 – December 2, 2022, Proceedings*, page 182–193, Berlin, Heidelberg. Springer-Verlag.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Nancy Chinchor. 1992. MUC-4 evaluation metrics. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

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

Linguistic Data Consortium. 2005. Ace (automatic content extraction) english annotation guidelines for events. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 151–159, New York, NY, USA. Association for Computing Machinery.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Hui Chen, Huaixiao Tou, Mosha Chen, Fei Huang, and Huajun Chen. 2021. Ontoed: Low-resource event detection with ontology embedding. *CoRR*, abs/2105.10922.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. 2021. Corpus-level evaluation for event QA: The IndiaPoliceEvents corpus covering the 2002 Gujarat violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4240–4253, Online. Association for Computational Linguistics.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *Preprint*, arXiv:2305.14450.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12804–12825, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The Genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria. Association for Computational Linguistics.

Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400, Online. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *Preprint*, arXiv:2304.11633.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and Philip S. Yu. 2022. A survey on deep learning event extraction: Approaches and applications. *Preprint*, arXiv:2107.02126.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Thi-Nhung Nguyen, Bang Tien Tran, Trong-Nghia Luu, Thien Huu Nguyen, and Kiem-Hieu Nguyen. 2024. BKEE: Pioneering event extraction in the Vietnamese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2421–2427, Torino, Italia. ELRA and ICCL.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023. Omnievent: A comprehensive, fair, and easy-to-use toolkit for event understanding. *Preprint*, arXiv:2309.14258.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8749–8757.

Jiawei Sheng, Shu Guo, Bowen Yu, Qian Li, Yiming Hei, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2021. Casee: A joint learning framework with cascade decoding for overlapping event extraction. *Preprint*, arXiv:2107.01583.

Jiuding Sun, Chantal Shaib, and Byron C. Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *Preprint*, arXiv:2306.11270.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Preprint*, arXiv:1409.3215.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, and et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A

Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus ldc2006t06. https://catalog.ldc.upenn.edu/LDC2006T06.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. 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 337–346, Hong Kong, China. Association for Computational Linguistics.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online. Association for Computational Linguistics.

A Detailed Results

Table 3 demonstrates the detailed evaluation results for trigger identification, trigger classification, argument identification, argument classification respectively.

B Prompts for LLMs

Table 4 and 5 illustrate the prompts we use for testing the ability of LLMs in event extraction task.

C Details of the Event Schema

Tables 6 and 7 illustrate the definitions of entity and argument role types, respectively, while Table 8 and 9 provide examples of each event type in the VHE dataset.

Model	TI			TC			AI			AC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Gemini-1.5-flash (zero-shot)	53.85	29.41	38.04	44.94	24.89	32.03	28.75	8.78	13.45	16.81	5.13	7.86
Gemini-1.5-flash (2-shot)	59.22	26.33	36.45	49.20	22.50	30.88	28.35	9.68	14.43	19.41	6.65	9.91
Gemini-1.5-flash (4-shot)	42.14	30.23	35.20	35.39	24.70	29.10	27.47	9.38	13.99	18.69	6.34	9.46
GPT-3.5-turbo (zero-shot)	43.46	17.01	24.45	27.95	11.75	16.55	19.42	2.53	4.47	15.12	1.98	3.51
GPT-3.5-turbo (2-shot)	51.58	20.08	28.91	32.46	13.68	19.25	19.84	2.69	4.74	16.02	2.17	3.82
GPT-3.5-turbo (4-shot)	49.73	18.55	27.01	31.53	12.86	18.26	22.14	2.77	4.93	18.14	2.25	4.01
GPT-4o (zero-shot)	76.09	10.76	18.85	70.59	9.92	17.39	39.89	2.46	4.64	27.18	1.68	3.16
GPT-4o (2-shot)	66.10	27.77	39.11	56.56	24.15	33.85	38.11	8.43	13.81	29.69	6.60	10.80
GPT-4o (4-shot)	70.25	23.46	35.18	62.05	20.57	30.90	41.75	8.79	14.53	31.99	6.77	11.18
Llama-3.1-8B-Instruct (2-shot)	39.95	47.44	43.37	23.04	34.07	27.49	24.65	19.05	21.49	16.78	13.06	14.68
Llama-3.1-8B-Instruct (4-shot)	36.27	51.54	42.57	22.34	37.65	28.04	25.94	25.14	25.53	17.67	17.06	17.36
Gemma-2-9b-it (2-shot)	44.49	53.38	48.53	29.07	39.49	33.49	24.21	17.16	20.09	20.29	14.42	16.86
Gemma-2-9b-it (4-shot)	42.44	53.18	47.20	29.34	41.14	34.25	28.21	21.48	24.39	23.08	17.51	19.92
Phi-3.5-mini-instruct (2-shot)	24.49	13.42	17.34	12.93	7.25	9.29	10.82	1.57	2.74	7.45	1.07	1.87
Phi-3.5-mini-instruct (4-shot)	20.85	8.50	12.08	11.76	4.78	6.79	10.34	0.84	1.56	7.64	0.62	1.14
Qwen-2-7B-Instruct (2-shot)	23.92	38.01	29.36	12.78	21.76	16.10	15.00	8.54	10.88	8.47	4.80	6.12
Qwen-2-7B-Instruct (4-shot)	17.91	21.82	19.68	11.59	13.50	12.47	12.79	3.58	5.60	8.45	2.34	3.67
Seq2Seq + mT5	28.81	12.19	17.13	7.68	5.51	6.42	11.82	1.11	2.03	1.89	0.19	0.35
Annotator 1	86.14	72.50	78.73	69.16	57.81	62.98	62.64	49.66	55.40	52.72	41.91	46.70
Annotator 2	79.41	67.50	72.97	64.35	57.81	60.91	40.65	39.64	40.14	36.36	35.54	35.94

Table 3: Precision (P), Recall (R), and F1 scores for four subtasks, including Trigger Identification (TI), Trigger Classification (TC), Argument Identification (AI), and Argument Role Classification (AC) on VHE.

Zero-shot prompt for Event Extraction

Instruction
Your task is to extract all events mentioned in a list of texts. If any event does not belong to the event types listed below, or if you are unsure, just ignore it.
Input format: text-id: text.

Context
An event has four parts: the event type, which includes the type of event and its corresponding subtype; the event trigger, which is a word or phrase that most clearly expresses the occurrence of the event; the event arguments, which are entities involved in the event; and the argument role, which defines the relationship between the event and its arguments.

Event types: {event 1, ..., event n}
Entity types: {entity 1, ..., entity n}
Argument roles: {role 1, ..., role n}

Output Indicator
Output format: A list of strings, where each string represents an event. Each event includes the following components separated by pipes: text-id | event-type | event-trigger | event-arguments. Each event argument follows the format: argument - entity type - argument role, and multiple event arguments are separated by commas.
No explanation in output.

Input Data
Text: {text}

Table 4: Zero-shot prompt template used for evaluating LLMs' performance on the event extraction task.

Few-shot prompt for Event Extraction

```
### Instruction ###
I will provide you with some examples of event extraction, your task is to extract all events mentioned in a list of texts. Note that these examples do not cover all event types in the texts, so please extract any events that match the types listed below. If an event does not belong to the specified types or if you are unsure, just ignore it.

### Context ###
An event has four parts: the event type, which includes the type of event and its corresponding subtype; the event trigger, which is a word or phrase that most clearly expresses the occurrence of the event; the event arguments, which are entities involved in the event; and the argument role, which defines the relationship between the event and its arguments.

Event types: {event 1, ..., event n}
Entity types: {entity 1, ..., entity n}
Argument roles: {role 1, ..., role n}

Example 1:
Input:
s1: Xưa cháu ba đời của Viêm Đế họ Thần Nông là Đế Minh sinh ra Đế Nghi, sau Đế Minh nhân đi tuần phương Nam, đến Ngũ Lĩnh lấy con gái Vụ Tiên, sinh ra vua [Kinh Dương Vương].
s2: Vua Vũ chia chín châu thì Bách Việt thuộc phần đất châu Dương, Giao Chỉ thuộc về đấy.
s3: Mùa thu, tháng 9, ngày rằm, giờ Mão, có nhật thực.
Output:
s1 | LIFE.BE-BORN | sinh ra | Đế Minh - PERSON - AGENT, Đế Nghi - PERSON - THEME
s1 | LIFE.MARRY | lấy | Đế Minh - PERSON - AGENT, con gái Vụ Tiên - PERSON - THEME
s1 | LIFE.BE-BORN | sinh ra | Đế Minh - PERSON - AGENT, Kinh Dương Vương - PERSON - THEME
s3 | NATURE.NATURAL-PHENOMENON | nhật thực | Mùa thu - TIME - TEMPORAL, tháng 9 - TIME - TEMPORAL, ngày rằm - TIME - TEMPORAL, giờ Mão - TIME - TEMPORAL

Example 2:
Input:
s4: Nhâm Tuất, năm thứ 1.
s5: Tháng 3, ngày mồng 6, đúc xong ấn báu.
Output:
Not found.

... (n-shot)

Text:
{sentence 1: text 1}
...
{sentence n: text n}
Output:
```

Table 5: Few-shot prompt template used for evaluating LLMs' performance on the event extraction task.

No.	Entity Type	Description	Count
1	Person (PER)	Name of a specific person or family	6901
2	Date time (DTM)	Time or a specific period of time	2606
3	Location (LOC)	Names of land according to political or geographical border (city, province, country, international regions, oceans…	1449
4	Designation (DES)	Position or title of a specific person.	1371
5	Organization (ORG)	Names of organizations, offices or companies	443
6	Miscellaneous (MISC)	Other entities	345
7	Terminology (TRM)	Word-combinations having special meanings depending on the contexts are used in respective specialties. They include: science, technique, military, politics, religion…	141
8	Measurement (MEA)	Measurement, quantity of things (other than money) in a standard unit.	106

Table 6: Entity types used in the VHE dataset. **Count:** count of annotated entities.

No.	Argument Role Type	Description	Count
1	Theme	The participant most directly affected by an event.	4148
2	Agent	The volitional causer of an event.	3769
3	Temporal	The time the event occurred	2608
4	Content	The proposition or content of a propositional event.	1357
5	Location	The location the event occurred	652
6	Goal	The destination of an object of a transfer event.	458
7	Beneficiary	The beneficiary of an event.	196
8	Source	The origin of the object of a transfer event.	128
10	Instrument	An instrument used in an event.	33
11	Force	The non-volitional causer of the event.	7
12	Experiencer	The experiencer of an event.	6

Table 7: Argument role types used in the VHE dataset. **Count:** count of annotated arguments.

Event Type	Event Subtype	Example	Count
LIFE	BE-BORN	Tháng 3 , ngày mồng 5 , cháu chúa Trịnh Tạc ra đời , đó là con trai thứ của Bình quận công. (In March, on the 5th day, the grandson of Lord Trịnh Tạc was born, who was the second son of Duke Bình.)	117
	MARRY	Tháng 3 , ngày mồng 7 , gả công chúa Bình Dương cho châu mục châu Lạng là Thân Thiệu Thái . (In March, on the 7th day, Princess Bình Dương was married to Thân Thiệu Thái, the chieftain of Châu Lạng.)	90
	DIVORCE	Đến khi Lĩnh bị giết, Thuyên cũng bỏ vợ . (When Lĩnh was killed, Thuyên also abandoned his wife.)	4
	INJURE	Thạc đoạt lấy cờ tiết của Lượng, Lượng không cho, Thạc bèn chặt tay trái của Lượng , Lượng nói: "Chết còn không tránh, chặt cánh tay thì làm gì?". (Thạc seized Lượng's flag, which Lượng refused to give up, so Thạc cut off Lượng's left arm. Lượng said, "Even death cannot be avoided, what's the use of cutting off my arm?".)	21
	DIE	Tuần bèn giết hết những kẻ không chịu chết theo, rồi gieo mình xuống sông mà chết. (Tuần then killed all those who refused to die with him, and then threw himself into the river to die.)	890
MOVEMENT	TRANSPORT	Quân Lương tan vỡ chạy về Bắc . (The troops of Lương were defeated and fled north.)	389
TRANSACTION	TRANSFER-OWENERSHIP	Châu Vi Long (nay châu Đại Man) dâng ngựa trắng bốn chân có cựa . (Châu Vi Long (now Châu Đại Man) offered a white horse with four legs and spurs.)	179
	TRANSFER-MONEY	Vua rất hiểu ông, sai người ban đêm đem 10 quan tiền bỏ vào nhà ông . (The king highly valued him, sending someone at night to place 10 quan of money in his house.)	29
BUSINESS	START-ORG	Tháng 6 , lập Quốc học viện . (In June, the National Academy was established.)	31
	MERGE-ORG	Trước đây, châu Nam Mã thuộc nước Ai Lao, sau vì mộ đức nghĩa nhà vua mà quy thuận . (Previously, Châu Nam Mã belonged to the country of Ai Lao, but later it submitted due to the king's virtue.)	4
	DECLARE-BANKRUPCTY	N/A	0
	END-ORG	Năm ấy nhà Chu mất . (That year, the Zhou dynasty fell.)	18
CONFLICT	ATTACK	Mùa thu , tháng 7 , ngày mồng 5 , nước Ai Lao lại làm phản, đánh vào Mường Viễn . (In autumn, on the 5th day of the 7th month, the country of Ai Lao rebelled again and attacked Mường Viễn.)	857
	DEMONSTRATE	Thái bảo Phù quận công Trịnh Lịch , Thái phó Hoa quận công Trịnh Sầm , hận vì bất đắc chí, liền nổi quân làm loạn . (The Grand Protector of Phù Duke Trịnh Lịch and the Grand Tutor of Hoa Duke Trịnh Sầm, frustrated by their failures, raised troops to revolt.)	7
CONTACT	MEET	Thời Thành Vương nhà Chu [1063-1026 TCN] , nước Việt ta lần đầu sang thăm nhà Chu (không rõ vào đời Hùng Vương thứ mấy), xưng là Việt Thường thị, hiến chim trĩ trắng. (During the reign of King Cheng of the Zhou dynasty [1063-1026 BC], our country of Viet made its first visit to the Zhou (uncertain which reign of the Hùng Kings), calling itself Viet Thường thị and offering white pheasants.)	119
	PHONE-WRITE	Mới rồi nghe nói vương có gửi thư cho tướng quân Lâm Lư hầu , muốn tìm anh em thân và xin bãi chức hai tướng quân ở Trường Sa. (Recently, it was heard that the king sent a letter to General Lâm Lư Hầu, seeking to find close relatives and requesting to remove the two generals in Chương Sa.)	67

Table 8: Examples of event types used in the VHE dataset. Event triggers are highlighted in orange and event arguments are highlighted in green.

Event Type	Event Subtype	Example	Count
NATURE	NATURAL-PHENOMENON	Tháng 2 , ngày Đinh Dậu mồng 1 , có nhật thực . (In February, on the 1st day of Đinh Dậu, there was a solar eclipse.)	266
	NATURAL-DISASTER	Mùa hạ , tháng 4 , hạn hán . (In summer, in April, there was a drought.)	99
PERSONNEL	START-POSITION	Cháu là Hồ lên nối ngôi . (The grandson Hồ ascended to the throne.)	1329
	END-POSITION	argumenttextitMùa đông, tháng 10 , ngày Nhâm Ngọ , Đàn Hòa Chi bỏ quan về. (In winter, in October, on the day of Nhâm Ngọ, Đàn Hòa Chi left his position and returned home.)	123
	NOMINATE	Đến đây, Quý Ly tiến cử ông ta . (At this point, Quý Ly recommended him.)	40
	ELECT	Bề tôi nhà Minh lại tôn lập Vĩnh Lịch Hoàng Đế . (The officials of the Ming dynasty again revered Emperor Vĩnh Lịch.)	40
JUSTICE	ARREST-JAIL	Phiên tướng Thái Nguyên là Thông quận công Hà Sĩ Tứ đem quân bản xứ đi đánh, bị giặc bắt được. (The provincial general Thái Nguyên, Duke Hà Sĩ Tứ, who led local troops, was captured by the enemy.)	258
	RELEASE-PAROLE	Vua bằng lòng, tha cho Chế Củ về nước (Địa Lý nay là tỉnh Quảng nam). (The king agreed and pardoned Chế Củ, allowing him to return home (now Địa Lý, Quảng Nam province).)	34
	TRIAL-HEARING	Xuống chiếu cho quan Đình uý xét tội Lợi . (Issued a decree for the Inspector of the Capital to investigate Lợi's crimes.)	26
	CHARGE-INDICT	Nguyễn Vĩnh Tích hặc tội , cho là đáng phải biếm chức. (Nguyễn Vĩnh Tích accused of [a crime], deeming it worthy of being demoted..)	15
	SUE	Em Đỗ Khắc Chung là Đỗ Thiên Thư kiện nhau với người, tình lý đều trái. (Đỗ Khắc Chung's brother, Đỗ Thiên Thư, was in dispute with someone, with both the facts and reasoning against him.)	7
	CONVICT	Tử Dục hết lẽ, phải thú tội . (Tử Dục, having exhausted all reasons, had to confess to his crimes.)	5
	SENTENCE	Tư không châu Phục Lễ Đèo Mạnh Vượng có tội, cho tự tử . (The Chancellor of Châu Phục Lễ, Đèo Mạnh Vượng, was guilty and was allowed to commit suicide.)	37
	FINE	Công bộ hữu thị lang Trịnh Công Đán bị phạt 30 quan tiền vì bỏ phơi mưa nắng những gỗ, lạt của công. (The Minister of Works, Trịnh Công Đán, was fined 30 quan for neglecting to protect public wood and rattan from the weather.)	10
	EXECUTE	Chém Hồ bả ở phường Diên Hưng . (Executed Hồ Bả in Diên Hưng district.)	76
	EXTRADITE	Tháng 5 , nhà Thanh sai Phạm Thành Công và Mã Văn Bích mang sắc dụ đến cửa Nam Quan, bảo bắt giải lũ giặc biển Dương Nhị , Dương Tam . (In May, the Qing Dynasty sent Phạm Thành Công and Mã Văn Bích with an edict to the South Gate, ordering the capture and return of the pirate leaders Dương Nhị and Dương Tam.)	2
	ACCQUIT	N/A	0
	APPEAL	Vì tám người cùng họ như Lê Khắc Phục và công chúa Ngọc Lan làm đơn khẩn thiết van xin vua nói phép ban ơn, nên có lệnh này. (Because eight people with the same surname, such as Lê Khắc Phục and Princess Ngọc Lan, earnestly petitioned the king for leniency, this order was issued.)	2
	PARDON	Tháng 3 , tha tội chết cho Nguyễn Sư Hồi . (In March, the death penalty was commuted for Nguyễn Sư Hồi.)	46

Table 9: Continuation of Table 8.