Few-Shot Prompting with Large Language Models for Multilingual Person–Place Relation Extraction from Historical Documents

Haojie Cao, Zhongyuan Han* and Xianbing Duan

Foshan University, Foshan, China

Abstract

This work presents the system developed by the MaxFo-Ajie team for the HIPE-2026 shared task, which focuses on person-place relation extraction and classification from multilingual historical documents. Different from conventional task-specific fine-tuning paradigms, the proposed method adopts a few-shot prompting framework based on the proprietary GPT-5.5 model accessed through an API, requiring only a small number of annotated task examples and no parameter updates. This paper designs three few-shot prompting strategies with distinct inference granularities and example presentation formats, including document-level inline prompting, document-level multi-turn conversational prompting, and pair-level inline prompting. Extensive evaluations are conducted to validate the performance of the designed strategies. In the official benchmark evaluation, MaxFo-Ajie ranks 2nd overall in the Accuracy Profile, with run1 achieving a Test A profile score of 0.7001 and placing 3rd at the run level. Furthermore, our method occupies the top three run positions in the Generalization Profile, where run3 ranks 1st with a Test B `MacroRecall_at` score of 0.8163, demonstrating strong cross-lingual generalization performance across German, English, and French, as well as strong robustness on the out-of-domain French literary surprise test set. The experimental results show that API-based LLM prompting, combined with simple post-processing operations, is competitive for low-resource relation extraction on historical documents, while also raising reproducibility, cost, and data-governance considerations.

Keywords

person–place relation extraction, large language models, few-shot prompting, historical documents, multilingual NLP, HIPE-2026

1. Introduction

The automatic extraction of person–place relations from historical documents is a central task in digital humanities and spatial humanities research. Accurately answering the question of *who was where and when* contributes to the reconstruction of biographical trajectories, the analysis of population mobility patterns, and the construction of large-scale historical knowledge graphs, holding direct value for research in history, geography, and cultural heritage preservation.

The HIPE (Identifying Historical People, Events and Places) evaluation campaign has previously been held twice. HIPE-2020 [1] conducted evaluations of named entity recognition, classification, and linking on a French historical newspaper corpus. HIPE-2022 [2] expanded the scope to multilingual corpora encompassing German, English, French, and Luxembourgish, and introduced more complex document types and annotation schemes. HIPE-2026 [3, 4] advances the task to a deeper semantic level, requiring systems not only to identify person–place entity pairs but also to classify their temporal relationships[1]. Specifically, systems must distinguish between two relation types: `at`, indicating whether a person has ever visited or resided at a place prior to the document's publication, and `isAt`, indicating whether the person is located at the place in the immediate temporal context of the document's publication. The `at` relation is classified into three evidence levels: TRUE (clear evidence), PROBABLE (reasonable inference), and FALSE (no evidence or contradicting evidence), while `isAt` is a binary TRUE/FALSE relation.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

caohaojie0322@gmail.com (H. Cao); hanzhongyuan@gmail.com (Z. Han); xian65921@gmail.com (X. Duan)

[1] Task website: https://hipe-eval.github.io/HIPE-2026/.

The task presents several challenges. First, systems must go beyond surface-level co-occurrence analysis and perform genuine textual semantic reasoning to determine whether person–place relations hold. Second, the temporal distinction between `at` and `isAt` requires understanding of verb tense, temporal expressions, and discourse structure. Third, the documents involve three languages (German, English, and French) and originate from historical newspapers with varying degrees of OCR noise. Fourth, the surprise test set is drawn from an entirely different domain—16th–18th century French literary works—posing a severe test of out-of-domain generalization.

In recent years, large language models have demonstrated remarkable zero-shot and few-shot capabilities across numerous NLP tasks [5, 6]. This motivates the central research question of this work: *Can a prompt-based LLM approach, without any domain-specific fine-tuning, achieve competitive performance on the specialized task of historical document relation extraction?*

The contributions of this paper are as follows. We design and evaluate three distinct prompting strategies spanning two levels of inference granularity (document-level and pair-level) and two formats of few-shot example presentation (inline text and multi-turn dialogue). Experiments show that the prompt-based approach ranks 2nd overall in the Accuracy Profile by team best run, with run1 placed 3rd at the run level, and sweeps the top three run positions in the Generalization Profile among 17 participating teams and 45 valid submissions[2]. We further analyze the trade-offs among the three strategies, finding that document-level inference performs better on in-domain data while pair-level inference exhibits advantages in cross-domain generalization.

2. Related Work

2.1. The HIPE Evaluation Series

The HIPE shared task series aims to benchmark named entity processing capabilities on historical document collections. HIPE-2020 [1] introduced evaluations of named entity recognition, classification, and linking on a monolingual (French) historical newspaper corpus. HIPE-2022 [2] expanded the scope to multilingual corpora covering German, English, French, and Luxembourgish, and introduced more complex document types and annotation tagsets. Systems in these earlier editions commonly relied on sequence labeling architectures, multilingual Transformer encoders, gazetteers, and entity-linking pipelines adapted to noisy OCR and historical spelling variation. HIPE-2026 [3, 4] represents a further evolution of the series, shifting the focus from entity-level processing to relational reasoning between entities. The task requires not only identifying person–place pairs but also classifying their temporal relationships—a substantially more complex semantic requirement that is harder to solve with local mention-level features alone.

2.2. Large Language Models for Information Extraction

The application of large language models to information extraction tasks has attracted widespread attention. Models such as GPT-4 [7] and other proprietary and open-source LLMs have demonstrated that carefully designed prompts can elicit structured extraction capabilities from models without task-specific training. In the domain of relation extraction, existing research has shown that LLMs can handle multi-hop reasoning, implicit relation detection, and cross-lingual transfer with minimal supervision [8]. The few-shot prompting paradigm—providing a small number of annotated examples within the input context—has proven particularly relevant for historical NLP, where expert annotation is expensive and document collections often differ in language, period, genre, and OCR quality. Compared with fine-tuned encoders, prompt-based LLM systems can be rapidly adapted to a new annotation scheme, but they also introduce dependencies on prompt design, example selection, model versioning, API availability, and output normalization.

[2]The official results include 46 runs in total: 45 valid submissions from participating teams plus 1 random decision baseline submitted by the organizers.

2.3. Relation Extraction in Historical Texts

Relation extraction from historical documents faces a set of unique challenges distinct from contemporary text processing. OCR-induced noise, historical language variation, temporal ambiguity, and limited contextual information collectively increase task difficulty. Traditional methods based on co-occurrence statistics or pattern matching cannot satisfy the requirements of HIPE-2026—systems must reason about whether a person actually visited a place, and whether that visit occurred within the immediate temporal context of the document's publication. Supervised neural approaches can encode richer context, but their performance depends on annotated relation data for each language and domain. This data bottleneck is especially pronounced in digital humanities projects, where collections may be small, heterogeneous, or legally constrained, and where annotation requires historical and linguistic expertise. Our work therefore examines whether an LLM-based few-shot system can provide a competitive baseline without task-specific parameter training.

3. Method

3.1. Model and API Configuration

All three submitted runs use GPT-5.5 [9] via the official OpenAI API. Public documentation for GPT-5.5 does not disclose parameter count or full architectural details; consequently, our system should be understood as a lightweight implementation and deployment framework rather than a lightweight model. The submitted runs used deterministic decoding with temperature set to 0.0, three retries per failed request, and no external retrieval or tool calls during inference. We did not set a top-k parameter in the API request.

3.2. Prompt Construction

Our system operates entirely through LLM API calls, without any domain-specific fine-tuning or external knowledge base invocation, relying solely on the in-context learning capability of the underlying LLM. Formally, given a document d containing a set of person entities $\mathcal{P} = \{p_1, \ldots, p_m\}$ and a set of place entities $\mathcal{L} = \{l_1, \ldots, l_n\}$, the system must classify each pair (p_i, l_j): the `at` relation is labeled as one of TRUE, PROBABLE, or FALSE, and the `isAt` relation is labeled as TRUE or FALSE.

The prompt consists of a system prompt and a user prompt (complete prompt templates are provided in Appendix A). All prompts are written in English, regardless of the document language. The system prompt defines the annotator's role and labeling conventions: `at=TRUE` requires textual evidence that the person visited or resided at the place at some point prior to publication; `at=PROBABLE` applies when evidence is weak or indirect but still reasonable; `at=FALSE` indicates no supporting evidence. For `isAt`, a TRUE label is assigned only when the person's presence is supported by text within approximately one month of the publication date. The prompt further instructs the model to adopt a conservative stance when evidence is ambiguous, preferring PROBABLE over TRUE and defaulting `isAt` to FALSE. This rule was chosen a priori from the annotation definitions and the expected difficulty of temporal evidence, not tuned directly against the official macro-recall scores. Each user prompt includes document metadata (document ID, language, publication date, and source media), the document text truncated to 4000 characters, a set of candidate person–place pairs (each accompanied by entity IDs, surface mention lists, and available Wikidata QIDs), and a JSON output format example.

Few-shot examples are selected from the task training/development material by language, with a maximum of four examples per target language. The example pools contain six preselected documents per language. We manually selected these pools to cover different newspapers, dates, relation labels, and entity-pair densities, while avoiding examples from the test documents. The inference script then uses the first four examples matching the target language. The same example files and ordering are used for all three submitted runs so that run differences reflect inference granularity and few-shot presentation format rather than different demonstrations.

The 4000-character truncation threshold was selected before official submission to keep document-level prompts within a manageable context and output budget when combined with up to four few-shot examples and as many as 16 candidate pairs. On the official unlabeled test files, truncation affects 10/19 German impresso documents, 1/19 English impresso documents, and 7/19 French impresso documents; it does not affect the 30 French surprise documents, whose maximum length is 2044 characters. This design therefore trades off potential loss of distant evidence, especially for German and French newspapers, against lower API cost and fewer context-length failures.

Before settling on the three submitted runs, we tested simpler zero-shot prompts and variants that asked for longer free-text rationales. Zero-shot prompts were less stable on the PROBABLE class, and longer rationales increased the frequency of malformed or incomplete JSON. We therefore submitted the three most reliable configurations: inline document-level prompting, multi-turn document-level prompting, and inline pair-level prompting.

3.3. Inference Strategies

We design three inference strategies that differ along two dimensions—inference granularity (document-level vs. pair-level) and few-shot example presentation format (inline text vs. multi-turn dialogue)—and submit them as three independent runs. Figure 1 illustrates the overall pipeline of the three inference strategies.

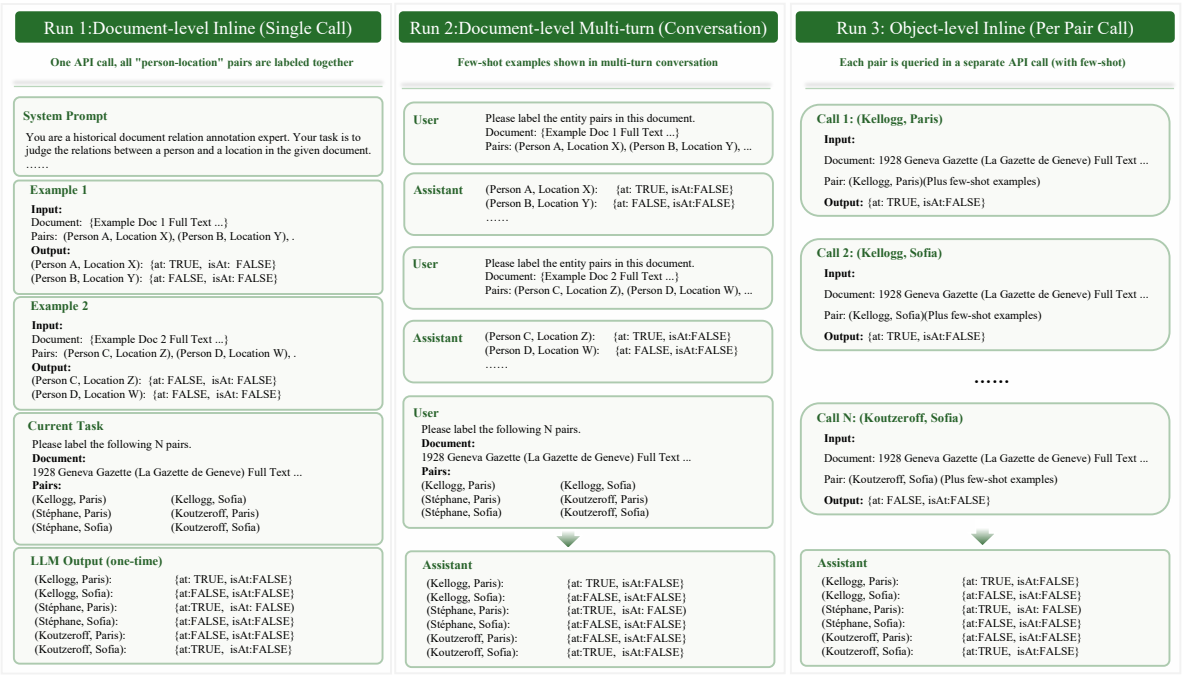

Figure 1: Overview of the three inference strategies. Run 1 and Run 2 process all person–place pairs in a single document-level LLM call with inline and multi-turn few-shot examples, respectively. Run 3 processes each pair independently in separate calls.

Run 1: Document-Level Inline Few-Shot. Few-shot examples are embedded as text blocks within the user prompt, which is structured as system prompt followed by annotated examples in a structured input–output format, followed by the document to be annotated. The LLM processes the entire document in a single inference call, simultaneously producing classifications for all person–place pairs. This strategy enables the model to consider potential inter-pair dependencies within a shared context, while requiring only one API call per document.

Run 2: Document-Level Multi-Turn Few-Shot. Few-shot examples are presented in a multi-turn conversational format, where each example constitutes a complete user–assistant exchange. The model observes multiple rounds of example inputs and their corresponding annotated outputs before receiving

the current document. This format more closely resembles the instruction-following paradigm under which many LLMs are trained, potentially enabling more effective in-context learning, but requires a longer context window.

Run 3: Pair-Level Inline Few-Shot. This approach processes each person–place pair independently: for each pair, the LLM receives the full document text and a single candidate pair, producing a classification for that pair alone. Consequently, each document requires N API calls (where N is the number of candidate pairs). This strategy makes each inference task more focused and constrained, reducing the cognitive load on the model, but at the cost of losing access to inter-pair contextual information and significantly increased API call expenses.

3.4. Output Parsing and Label Normalization

The LLM's raw output undergoes multi-stage processing. The system first attempts to parse the output as a JSON object. It supports bare JSON, JSON enclosed in fenced code blocks (i.e., Markdown triple-backtick blocks sometimes produced by chat models), and bracket-matched extraction from free-form text by taking the substring between the first opening brace and the last closing brace. Parsed labels are then normalized to the canonical label set: for the `at` relation, non-standard values such as "LIKELY", "POSSIBLE", and "MAYBE" are mapped to PROBABLE, with all other unrecognized values defaulting to FALSE; for the `isAt` relation, any non-TRUE value is mapped to FALSE. In cases where API calls fail, output parsing is unsuccessful, or a prediction for a requested pair is missing, the system uses FALSE as the default label and logs the failure for subsequent analysis. In the submitted prediction files, no pair contains the explicit fallback explanation inserted after complete API or parsing failure, suggesting that complete response-level failure was not a major source of the final scores. However, label normalization can still affect class-level performance, especially for the difficult PROBABLE class, and should be considered part of the system rather than a purely mechanical formatting step.

4. Experimental Setup

4.1. Datasets

HIPE-2026 comprises two dataset families. We used the official HIPE-2026 v1.0 files and evaluated outputs with the official scorer distributed in the HIPE-2026 data release[3].

Test A (newspaper test set). Derived from the HIPE-2022 historical newspaper corpus, this dataset constitutes the primary evaluation benchmark. It includes documents in three languages—German (de), English (en), and French (fr)—each containing manually validated person–place relation annotations. Both the `at` and `isAt` relation types are evaluated on this dataset.

Test B (French literary surprise test set). Composed of 16th–18th century French literary works, this dataset is annotated for person–place relations. Only the `at` relation type is evaluated, and it is specifically designed to test out-of-domain generalization—participating systems are not exposed to literary-domain training data.

4.2. Evaluation Metrics

The official evaluation employs three ranking profiles, as defined in the HIPE-2026 evaluation report:

Accuracy Profile. This is the primary ranking, computed as the mean of `MacroRecall_at` and `MacroRecall_isAt` averaged across all submitted Test A language files. For each relation type, macro recall is the arithmetic mean of per-class recall scores (`at` includes TRUE, PROBABLE, and FALSE; `isAt` includes TRUE and FALSE).

Generalization Profile. Based on the `MacroRecall_at` score on Test B. The `isAt` relation is not evaluated in this profile.

[3] HIPE-2026 data version and official scorer release: https://github.com/hipe-eval/HIPE-2026-data/releases/tag/v1.0.

Efficiency Profile. This profile combines three ranks: accuracy rank, model parameter-count rank, and deployed model-size rank. Its score is their arithmetic mean, so lower is better.

4.3. Submission Configuration

Table 1 presents the configuration of our three submitted runs.

Table 1
Submission configuration of the three runs.

Run	Granularity	Few-shot format
run1	Document-level	Inline
run2	Document-level	Multi-turn
run3	Pair-level	Inline

5. Results and Analysis

5.1. Overall Rankings

Among 17 participating teams and 45 valid submissions (46 runs in total), our system achieved the following results: **2nd place** in the Accuracy Profile at the team level, with run1 scoring 0.7001 on Test A and placing 3rd at the run level, and **1st place** in the Generalization Profile with run3 scoring 0.8163 on Test B. For complete rankings, see the overview-paper result tables in [4].

5.2. Accuracy Profile

The Accuracy Profile ranks all submitted runs by their mean Test A profile score across the three languages. Table 2 presents the top 10 runs in this ranking, following the format of the official evaluation report. Our three runs are ranked 3rd, 7th, and 10th at the run level. By team best run, MaxFo-Ajie ranks 2nd overall; see also the corresponding overview-paper tables in [4].

Table 2
Accuracy Profile ranking on Test A (newspaper test set), top 10 runs. The score is the mean Test A profile score across de, en, and fr.

Rank	Team	Affiliation	Run	Score
1	Spinfo	U. zu Köln	run1	0.7479
2	Spinfo	U. zu Köln	run3	0.7289
3	MaxFo-Ajie	Foshan U.	run1	0.7001
4	Spinfo	U. zu Köln	run2	0.6890
5	whereami	Alexandria U.	run1	0.6880
6	whereami	Alexandria U.	run2	0.6833
7	MaxFo-Ajie	Foshan U.	run2	0.6690
8	Awakened	Politehnica Bucharest	run3	0.6671
9	Awakened	Politehnica Bucharest	run1	0.6584
10	MaxFo-Ajie	Foshan U.	run3	0.6544

46 runs total; random baseline score = 0.4049

Table 3 presents the per-language breakdown of our three runs, with detailed sub-scores for the `at` and `isAt` relations.

The per-language breakdown shows that English is the strongest in-domain setting for all three runs: run1 reaches 0.7305 overall, with both `at` macro recall (0.6645) and `isAt` macro recall (0.7965) higher than the corresponding German and French scores. German and French show different error

Table 3
Per-language scores of MaxFo-Ajie on Test A (newspaper test set). Score = Test A profile score; At-MR = `MacroRecall_at`; At-Acc = `at` accuracy; IsAt-MR = `MacroRecall_isAt`; IsAt-Acc = `isAt` accuracy.

Lang.	Run	Score	At-MR	At-Acc	IsAt-MR	IsAt-Acc
de	run1	0.6720	0.6048	0.7311	0.7391	0.8403
de	run2	0.6485	0.6146	0.7563	0.6823	0.8109
de	run3	0.6335	0.5864	0.7353	0.6807	0.8151
en	run1	0.7305	0.6645	0.7840	0.7965	0.8765
en	run2	0.6988	0.6397	0.7531	0.7579	0.8580
en	run3	0.6802	0.6358	0.7469	0.7246	0.8395
fr	run1	0.6978	0.5617	0.7353	0.8339	0.8824
fr	run2	0.6597	0.5385	0.7101	0.7810	0.8487
fr	run3	0.6495	0.5274	0.7017	0.7717	0.8403

profiles. German `at` performance is moderate but stable across runs (0.5864–0.6146), while French has the lowest `at` macro recall (0.5274–0.5617) despite the highest `isAt` macro recall for run1 (0.8339). This suggests that current-location judgments are often recoverable from explicit publication-time cues, whereas the broader historical `at` relation remains harder when evidence is indirect, distributed across the document, or affected by noisy OCR. Across languages, run1 consistently has the best overall score, mainly because document-level context improves `isAt` and preserves enough cross-pair evidence for `at`; run3 is more focused but loses some document-level signal on impresso. Notably, the top-ranked team, Spinfo, employed a substantially larger language model (116.8B parameters), whereas our system achieved comparable performance through API-based inference with GPT-5.5.

5.3. Generalization Profile

The Generalization Profile evaluates systems on Test B, drawn from a domain entirely different from the training data (French literary works). Table 4 presents the top 10 runs. Our three runs occupy the top three positions out of 46 runs, with the best score (0.8163) exceeding the 4th-ranked system by approximately 12 percentage points; see also the corresponding overview-paper tables in [4].

Table 4
Generalization Profile ranking on Test B (French literary surprise test set), top 10 runs. Score = `MacroRecall_at`.

Rank	Team	Affiliation	Run	Score
1	MaxFo-Ajie	Foshan U.	run3	0.8163
2	MaxFo-Ajie	Foshan U.	run1	0.7945
3	MaxFo-Ajie	Foshan U.	run2	0.7712
4	Spinfo	U. zu Köln	run3	0.6984
5	Spinfo	U. zu Köln	run1	0.6910
6	BIU_NLP	Bar-Ilan U.	run1	0.6837
7	Spinfo	U. zu Köln	run2	0.6674
8	whereami	Alexandria U.	run2	0.6665
9	gipplab	U. Göttingen	run2	0.6647
10	Awakened	Politehnica Bucharest	run3	0.6613

46 runs total; random baseline score = 0.3628

5.4. Comparison of Prompting Strategies

Table 5 summarizes the performance of our three strategies across both evaluation dimensions.

Table 5
Performance comparison of the three prompting strategies across Test A (Accuracy Profile) and Test B (Generalization Profile).

Run	Strategy	Accuracy	Generalization
run1	Doc-level inline	0.7001	0.7945
run2	Doc-level multi-turn	0.6690	0.7712
run3	Pair-level inline	0.6544	0.8163

Several observations emerge from this comparison. First, the document-level inline strategy (run1) achieves the highest accuracy on in-domain data, indicating that presenting all entity pairs within a shared context enables the model to exploit inter-pair dependencies, thereby benefiting relation classification. Second, the pair-level strategy (run3) performs best in the generalization test, which we attribute to the task simplification effect of single-pair inference—when the model only needs to judge one pair at a time, the task formulation is more concise and exhibits stronger robustness to domain differences. Third, the multi-turn dialogue strategy (run2) underperforms the inline variant (run1) on both evaluation dimensions, indicating that presenting few-shot examples in a conversational format does not confer consistent benefits for this task.

A noteworthy finding is that the strategy performing best on in-domain data (run1) differs from the one performing best on out-of-domain data (run3). By comparing the predictions of the two strategies, we observe that run1 tends to predict more TRUE labels for the `at` relation, while run3 is relatively more conservative. On the French impresso test set, there are 18 entity pairs where run1 labels `at` as TRUE but run3 does not, compared to only 5 pairs in the reverse direction. On the French surprise set, run1 predicts 122 TRUE, 75 PROBABLE, and 283 FALSE `at` labels, while run3 predicts 119 TRUE, 78 PROBABLE, and 283 FALSE labels, indicating a slight shift from TRUE to PROBABLE. We interpret this result as an interaction between genre and inference granularity. Newspaper texts often state events, visits, diplomatic roles, and locations explicitly, so document-level inference can benefit from wider context. Literary works, by contrast, more often encode place through narration, reported movement, and indirect reference; in this setting, judging a single pair at a time may reduce spurious associations. This is an explanatory hypothesis supported by genre differences and prediction behavior, rather than a full manual error taxonomy, which we leave for future work.

5.5. Efficiency Considerations

As our system uses external LLM APIs rather than locally deployed models, it did not rank prominently in the Efficiency Profile. This ranking is better suited to teams employing local small-scale models (e.g., 3B parameters), which inherently possess smaller computational footprints and publicly reportable model sizes. From a practical standpoint, the API-based approach requires no local GPU infrastructure and imposes minimal deployment overhead, which can be attractive to research teams with limited hardware access. This advantage should nevertheless be qualified: API inference incurs non-trivial monetary cost, run3 scales with the number of candidate pairs rather than documents, and external APIs may be inappropriate for collections subject to confidentiality, copyright, or data-protection restrictions.

6. Discussion

6.1. Strengths of the Prompt-Based Approach

The experimental results reveal several strengths of the prompt-based approach for historical document relation extraction. First, the method requires little task-specific annotated data and no task-specific parameter training, making it easier to adapt than supervised fine-tuning pipelines. Second, the multilingual capabilities of modern LLMs enable a single system to process German, English, and French documents without language-specific components. Third, the strong generalization performance

indicates that LLMs possess robust transfer capabilities that can bridge the domain gap between historical news text and literary works. For digital humanities research, this suggests a practical way to prototype relation extraction across heterogeneous collections before investing in large annotation campaigns. It may also support exploratory workflows in which scholars test alternative relation definitions, compare genres or periods, and prioritize documents for human review. In settings where proprietary external APIs are not appropriate, the same prompt-and-normalization framework could be transferred to locally deployed open-weight models, with expected trade-offs in accuracy, cost, and maintenance effort.

6.2. Limitations

The approach has several limitations. Reliance on external API calls causes inference costs to scale linearly with the number of requests, with the pair-level strategy (run3) being particularly expensive. Truncating document text to 4000 characters may result in the loss of contextual information at distant entity mentions, particularly for German and French impresso documents. Although temperature 0.0 reduces sampling variance, exact reproducibility is not achievable with a proprietary API model whose serving stack and model version may change over time. Furthermore, the system's dependence on large-scale proprietary models limits its applicability in settings where data privacy, copyright, institutional policy, or cost constraints preclude the use of external APIs.

6.3. Future Work

Several directions merit further exploration. A hybrid strategy—performing document-level coarse filtering followed by pair-level fine-grained inference for uncertain pairs—could potentially combine the strengths of both approaches. Chain-of-thought prompting [6] may guide models toward more structured reasoning, particularly for temporally ambiguous cases, although such reasoning traces would need to be controlled to preserve output validity. Locally deployed medium-scale models (7B–14B parameters) could provide a middle ground between accuracy and efficiency. Finally, incorporating external knowledge sources such as Wikidata could be implemented as retrieval-augmented generation: entity QIDs and candidate aliases could be used to retrieve concise biographical or geographical facts, which would then be inserted into the prompt to support disambiguation and temporal reasoning without replacing textual evidence from the document.

7. Conclusion

This paper presented the system submitted by the MaxFo-Ajie team to the HIPE-2026 shared task on person–place relation extraction from multilingual historical documents. Our approach is based on few-shot prompting of large language models, without domain-specific fine-tuning and with only a small number of task examples used in context. We designed and evaluated three prompting strategies differing in inference granularity and example presentation format. In the official evaluation, MaxFo-Ajie ranked 2nd overall in the Accuracy Profile among 17 participating teams, with run1 placed 3rd at the run level on Test A, and our three runs swept the top three positions in the Generalization Profile out of 46 total runs, with run3 ranked 1st on Test B. Experimental results demonstrate that prompt-based LLM inference with lightweight post-processing constitutes a competitive and generalizable approach to relation extraction in historical document settings, particularly suited to scenarios where labeled data is scarce or cross-domain robustness is required, while still requiring care around API cost, privacy, and reproducibility.

Declaration on Generative AI

During the preparation of this manuscript, the authors utilized OpenAI ChatGPT tools for manuscript translation, as well as grammar and spelling revisions. Following the use of this service, all content was

carefully reviewed and revised by the authors. The authors assume full responsibility for all contents of this published paper.

Acknowledgments

This work is supported by the National Social Science Foundation of China (24BYY080).

References

[1] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers, in: Working Notes of CLEF 2020–Conference and Labs of the Evaluation Forum, 2020.

[2] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents, in: Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum, volume 3180, 2022.

[3] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, S. Clematide, Overview of HIPE-2026: Person–Place Relation Extraction from Multilingual Historical Texts, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science (LNCS), Springer, 2026.

[4] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, E. Boros, S. Gabay, S. Clematide, Extended Overview of HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS, 2026. doi:10.5281/zenodo.20344461.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, volume 35, 2022.

[7] OpenAI, GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[8] S. Wadhwa, S. Amir, B. Wallace, Revisiting relation extraction in the era of large language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023.

[9] OpenAI, GPT-5.5 system card, https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf, 2026. Accessed 2026-06-30.

A. Prompt Templates

This appendix provides the prompt templates used in our experiments. All three runs share the same system prompt and user prompt structure; they differ only in few-shot presentation format (inline vs. multi-turn) and inference granularity (document-level vs. pair-level).

A.1. System Prompt

Listing 1: System Prompt

```
You are an expert relation extraction annotator for historical multilingual documents. Given a document
    and candidate person-location pairs, assign:
- at in {TRUE, PROBABLE, FALSE}
- isAt in {TRUE, FALSE}

Definition hints:
- at=TRUE if text supports that person has been/was at that place at some time before publication.
- at=PROBABLE if evidence is weak/indirect/ambiguous but plausible.
- at=FALSE if unsupported.
- isAt=TRUE only if presence is supported around publication time (roughly within one month). Else FALSE
    .
Be conservative with TRUE when evidence is weak.
Return JSON only.
```

A.2. User Prompt Template (Inline Mode)

Listing 2: User Prompt Template (used in Run 1 and Run 3)

```
Annotate all pairs in this document.
Output JSON only and exactly one object with key 'predictions'.
Keep explanations concise (<= 20 words).
If uncertain, prefer at=PROBABLE and isAt=FALSE.

FEW-SHOT EXAMPLES:
EXAMPLE 1 INPUT:
{
 "document_id": "...",
 "language": "de",
 "date": "1938-11-13",
 "media": {
  "publication_title": "...",
  "time_period": "...",
  "source_type": "..."
 },
 "text": "...",
 "pairs": [
  {
   "index": 0,
   "pers_entity_id": "...",
   "pers_mentions_list": ["..."],
   "loc_entity_id": "...",
   "loc_mentions_list": ["..."],
   "pers_wikidata_QID": null,
   "loc_wikidata_QID": "Q..."
  }
 ]
}
EXAMPLE 1 OUTPUT:
{
 "predictions": [
  {
   "index": 0,
   "at": "FALSE",
   "isAt": "FALSE",
   "at_explanation": "...",
   "isAt_explanation": "..."
  }
 ]
```

```
}

[additional examples up to 4 per language]

INPUT:
{
 "document_id": "...",
 "language": "en",
 "date": "1950-06-10",
 "media": {
  "publication_title": "...",
  "time_period": "...",
  "source_type": "..."
 },
 "text": "[truncated to 4000 characters]",
 "pairs": [
  {
   "index": 0,
   "pers_entity_id": "...",
   "pers_mentions_list": ["Joseph Roddy"],
   "loc_entity_id": "...",
   "loc_mentions_list": ["NEW YORK"],
   "pers_wikidata_QID": null,
   "loc_wikidata_QID": "Q1384"
  }
 ]
}

OUTPUT FORMAT EXAMPLE:
{
 "predictions": [
  {
   "index": 0,
   "at": "TRUE|PROBABLE|FALSE",
   "isAt": "TRUE|FALSE",
   "at_explanation": "short reason",
   "isAt_explanation": "short reason"
  }
 ]
}
```

A.3. Differences Across Runs

Run 1 (Document-Level Inline). Few-shot examples are embedded in the user prompt as text blocks. The model receives one call per document and predicts all person–place pairs jointly.

Run 2 (Document-Level Multi-Turn). Few-shot examples are presented as preceding user–assistant turns, and the target document is sent as the final user turn.

Run 3 (Pair-Level Inline). Prompt structure is the same as Run 1, but each call includes only one person–place pair, resulting in N calls per document.