Toponym Resolution: Will prompt engineering change expectations?

Isuri Anuradha Nanomi Arachchige
Lancaster University , UK
i.nanomiarachchige@lancaster.ac.uk

Deshan Koshala Sumanathilaka
Swansea University , Wales , UK
t.g.d.sumanathilaka@swansea.ac.uk

Ruslan Mitkov
Lancaster University , UK
r.mitkov@lancaster.ac.uk

Paul Rayson
Lancaster University , UK
p.rayson@lancaster.ac.uk

Abstract

Deep Learning and, more recently, Large Language Models(LLMs) have revolutionised the field of artificial intelligence and have been successfully employed in many disciplines, capturing widespread attention and enthusiasm. Many previous studies have established that domain-specific Deep Learning models competitively perform with the general-purpose LLMs. However, a suitable prompt which provides direct instructions and background information is expected to yield improved results. The present study, which focuses on Large Language Models for Toponym Resolution, shows that effective Prompt Engineering techniques without fine-tuning or pre-training approaches enable LLMs to surpass Deep Learning models, which is contrary to the initial expectations. After a comparison of open-source and proprietary LLMs and different prompt engineering techniques, the GPT-4o model performs best compared to the other open-source LLMs.

1 Introduction

Deep Learning and large language models (LLMs) have revolutionised the field of artificial intelligence. They have been successfully employed in many disciplines and have captured widespread attention and enthusiasm. Previous studies have established that domain-specific Deep Learning models often outperform general-purpose LLMs (Maatouk et al., 2024; Lu et al., 2024). However, a suitable prompt which provides direct instructions and background information is expected to yield better results (Kamruzzaman and Kim, 2024). Therefore, this study explores the use of pure prompt engineering techniques, without optimising the model weights through pre-training or fine-tuning.

This research discusses prompt engineering techniques such as few-shot learning, zero-shot learning, and Retrieval-Augmented Generation (RAG) to enhance the capabilities of output retrieval in LLMs. We aim to improve the models' ability to generate more relevant outputs for the resolution of toponyms. In previous studies, it has been established that the level of context-related understanding in LLM surpassed the efficacy of traditional rule-based and feature-based statistical machine learning methods(Zhao et al., 2024). A major challenge in this research area is the scarcity of domain-specific labelled datasets, limiting the applicability of traditional feature-based approaches. Therefore, this study focuses on experimenting and evaluating the different prompt engineering techniques for toponym resolution. We evaluated the performance of proposed methodologies through commercial and open-source generative LLMs by providing more refined and contextually specific prompts.

Toponym resolution in Natural Language Processing (NLP) is a highly challenging task, particularly when dealing with spoken language encompassing diverse dialects, accents, and linguistic nuances (Cardoso et al., 2019). The complexity of the task has been increased by the fact that, over time, geographic locations may have been referred to by different names in textual documents. In the context of the spoken text, a single name can refer to two different places in Historical textual contexts. As an example, survivors might use "Auschwitz" interchangeably to describe the town where they were initially brought and the camp where they were imprisoned. Previous studies have used different databases that represent the geographical places' coordinates to disambiguate the toponyms (Gritta et al., 2020; Sá et al., 2022). However, in historical textual contexts such as Holocaust narratives, the specific coordinates of the places relevant to the events are not included in the aforementioned databases in literature because these places are only mentioned within this specific domain. Compared to written texts, the unstructured nature of tran-

Proceedings of Recent Advances in Natural Language Processing,pages 95–104
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-011

scribed texts raises ambiguities in understanding their meaning. This is mainly due to the fact that transcribed texts have particularly inherent complexities such as variations in pronunciation, accents, and dialects, as well as the absence of punctuation. This represents additional complexity when the resolution of toponyms, related to the geospatial paradigm. In Holocaust testimonies, survivors explain what they have witnessed during this period, and frequently, these transcribed testimonies consist of sensitive and traumatic experiences that they have witnessed in different geographical locations. Because of the traumatic nature, often there is no consistency in the naming conventions, which emphasises the need for NER systems to be capable of resolving toponyms.

In this study, we study the toponym resolution task utilising the different advanced prompt engineering and augmentation techniques. Although some studies have sought to improve the understanding of geo-entities in natural language using LLMs, most have relied on model fine-tuning and pre-training approaches (Li et al., 2023). To the best of our knowledge, this is the first study to accurately identify and disambiguate toponyms specifically Geopolitical Entities (GPE), Locations (LOC), Concentration Camps (CAMP), Ghettos (GHETTO), and Streets (STREET) in transcribed Holocaust-related texts, relying solely on prompt engineering without model fine-tuning. Additionally, to determine the validity of the methodologies that experimented with the Holocaust Narratives and the Corpus of Lake District Writing (1622-1900), which is well known for geographic locations and Geo-nouns were also employed. While some research has been done on the recognition and tagging of geographical features in Lake district Corpus (Ezeani et al., 2023a,b), it has not gone into depth level to assess whether the toponym is an artificial location such as a bridge, building, or house, or natural location such as river, forest, or sea.

Since our study mainly focuses on the oral testimonies related to the Holocaust, Figure 1 provides real examples of the ambiguities of the NE in the testimonial contexts.

- Referring to the same name for different meanings
- Different names referring to the same place
- Symbols referring to the geographical location

However, when evaluating the data, we discovered similar difficulties in the Lake District corpus.

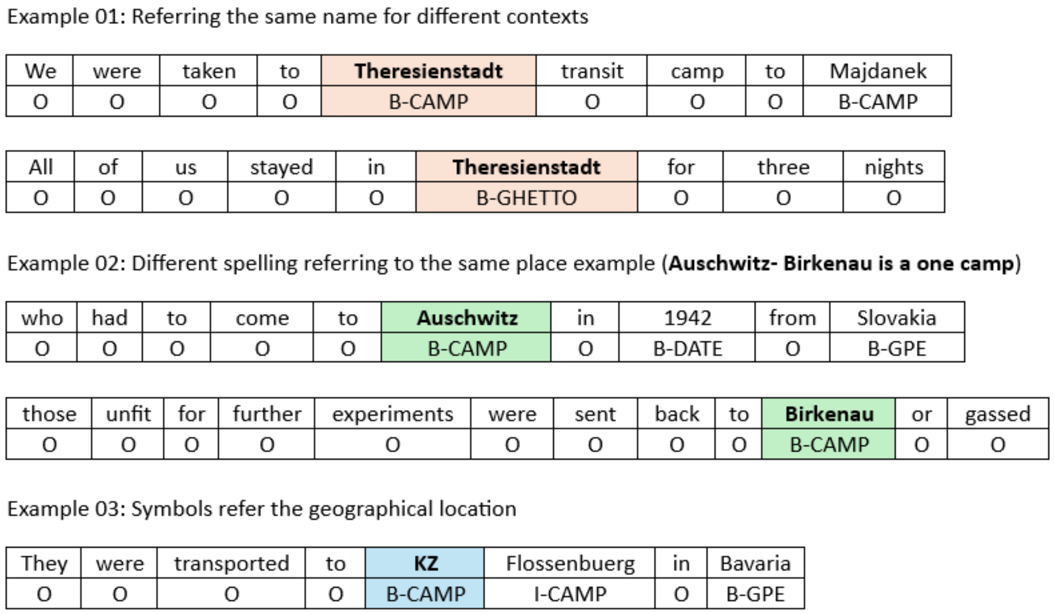
The rest of this paper is organised as follows. We outline previous studies in Section 2. In Section 3 we access our datasets and in Section 4 we describe the methods and Experiments. Section 5 offers a descriptive error analysis for two distinct datasets, and a brief conclusion is provided in Section 6.

2 Related Work

In previous studies, deep learning methodologies have been employed in toponym resolution to model the textual elements by combining bidirectional Long Short-Term Memory (LSTM) units with pre-trained contextual word embeddings (i.e., static features extracted using either the Embeddings from Language Models (ELMo) or the Bidirectional Encoder Representations from Transformers (BERT) methods. A limitation of these studies is that they typically apply only the general named entity tags, such as LOC and GPE, but not the generalised ones for other domain-specific geographical entities, such as Forests, concentration camps, ghettos, streets, etc.

Additionally, several studies have leveraged deep neural network architectures for toponym resolution (Cardoso et al., 2019; Kulkarni et al., 2021). For example, Gritta et al. proposed a network architecture called the CamCoder system, which aims to disambiguate place references by detecting lexical clues within the context surrounding the mention. The authors also introduced a sparse vector representation named MapVec, which encodes prior geographic probabilities associated with locations based on coordinates and population counts (Cardoso et al., 2019). Similarly, (Kulkarni et al., 2021) utilised a combination of context-aware word embeddings (Peters et al., 1802) and a recurrent neural network based on Bidirectional LSTMs (Huang et al., 2015). The above studies have cover not only English but also other languages such as Spanish.

Transformer-based techniques have recently had a substantial impact on toponym resolution methodologies. The current approaches can be broadly classified into two categories: localisation-based and ranking-based. The localisation-based approach primarily focuses on the direct prediction of geographic coordinates or areas from the given textual input. For instance, Radford's method (Radford, 2021) utilises DistilRoBERTa for end-to-end probabilistic geocoding. Similarly, (Cardoso



Example 01: Referring the same name for different contexts

We	were	taken	to	**Theresienstadt**	transit	camp	to	Majdanek
O	O	O	O	B-CAMP	O	O	O	B-CAMP

All	of	us	stayed	in	**Theresienstadt**	for	three	nights
O	O	O	O	O	B-GHETTO	O	O	O

Example 02: Different spelling referring to the same place example (**Auschwitz- Birkenau is a one camp**)

who	had	to	come	to	**Auschwitz**	in	1942	from	Slovakia
O	O	O	O	O	B-CAMP	O	B-DATE	O	B-GPE

those	unfit	for	further	experiments	were	sent	back	to	**Birkenau**	or	gassed
O	O	O	O	O	O	O	O	O	B-CAMP	O	O

Example 03: Symbols refer the geographical location

They	were	transported	to	**KZ**	Flossenbuerg	in	Bavaria
O	O	O	O	B-CAMP	I-CAMP	O	B-GPE

Figure 1: Sample examples for sentences extracted from the testimonies.

et al., 2022) employ Long Short-Term Memory (LSTM) networks with BERT embeddings to predict probability distributions over spatial regions. In a sequence-to-sequence framework, Solaz and Shalumov (Solaz and Shalumov, 2023) use the T5 Transformer model to translate text into hierarchical encodings of geographic cells. Another notable study by Gomes et al. (Gomes et al., 2024) proposes a method that leverages the adaptation of SentenceTransformer models, initially designed for sentence similarity tasks, for toponym resolution. The authors fine-tune the models on geographically annotated English news article datasets, including Local Global Lexicon, GeoWebNews, and TR-News. One of the major challenges in transformer-based toponym resolution methods is the absence of domain-specific fine-tuning. Pretrained transformer models such as BERT (Devlin et al., 2019) are optimised to generate embedding for tasks like masked language modelling and next-sentence prediction. Therefore, it is plausible that models trained on larger datasets have a greater capacity to identify the correct toponym.

Another significant issue with machine learning-based toponym resolution methods is the geographic bias, which arises due to the imbalance in the geographic distribution of training datasets. (Liu et al., 2022) pointed that models tend to favour locations overrepresented in the training corpora. The scarcity and lack of diversity in geo-tagged datasets further intensifies this bias (Gritta et al., 2018). Our review revealed a notable gap in the current body of research: no studies have employed state-of-the-art pure prompt engineering and augmentation techniques for toponym resolution. Despite advancements in generative language models demonstrating significant potential in other NLP tasks, their application to toponym resolution and disambiguation remains largely unexplored, with room for improvement in this domain.

3 Data and Annotation

In this study, we have used two distinct corpora: 1) Holocaust Survivor Narratives and 2) Lake District Writing corpus.

Holocaust Survivor Narratives Transcribed versions of oral Holocaust testimonies from the Wiener Holocaust Library were selected for the experiments. These data were manually annotated according to the BIO (Beginning-Inside-Outside) tagging scheme. Three human annotators manually annotated the training samples, resulting in an inter-annotator agreement of 0.76. For the annotation process, we employed UBIAI tool. More details about the data used for this study are reported in (Anuradha Nanomi Arachchige et al., 2023). Figure 1 shows the BIO tagging and the annotation style.

Corpus of Lake District Writing The Corpus of Lake District Writing (CLDW) comprises 80 texts and around 1.5 million words that describe the Lake District (Taylor and Gregory, 2022). From the seventeenth century until the early twentieth century, novels, fiction, letters, and diaries were included in the CLDW. Additionally, several travel books were also included in the CLDW, ranging from Thomas West's ”A Guide to the Lakes” (1778) to Black's ”Shilling Guide to the English Lakes” (1900) (Gregory et al., 2018).

4 Methods and Experiments

4.1 Baseline Approaches

Several baseline techniques were developed to identify toponyms using traditional techniques, which were subsequently assessed using the proposed prompt-augmented methods.

Rule-enhanced Deep Learning: We chose a customised SpaCy NER model with the named entities as one of the baseline models. We augmented the SpaCy transformer (trf) with rules to extract street names and ghettos. Table 1 lists a few of these specified rules.

Table 1: Examples for the defined regular expression for entity mining.

Entity	Regex Expression	Match
Street	If name followed by street semantically identical word ([A-Z][a-z]*(strasse—straße—straat)—_([A-Z][a-z]*(Street—St—Boulevard—Blvd—Avenue—Ave—Place—Pl)(˙)*))	Hauptstraße
Ghetto	Search on the lexicon consist Ghetto names or either name followed by ghetto [A-Z]w+((-—)*[A-Z]w+)* (g—G)hetto	Anyksciai

N-Gram Approach: We employ the N-gram method as a second baseline model. This approach takes into account the various n-gram window sizes that surround the target word. Then, in order to determine the most prominent entity, we . However, due to the unstructured nature of the dataset, this approach does not fare well.

Part-of-Speech Tags: We generate part-of-speech (POS) tags for each word in the corpus as third baseline. Next, we examine related POS tag combinations that are connected to the target word (Toponym) POS tag. We compute the probability that the target word is the correct toponym by utilising sentences that have the most similar POS tag pairings with it. As a result of oral and transcribed texts being highly unstructured, this approach also proved to be ineffective. Although we could identify common POS tag combinations with the target word, it was challenging to find a sufficient number of instances meeting our threshold. In particular, this method required a minimum of three identical sentences to forecast the label as a certain toponym based on the POS tag structure; however, this requirement was not consistently fulfilled in the dataset.

4.2 Prompt Augmentation techniques for domain-specific settings

To compare general-purpose LLMs' effectiveness in the geo-spatial domain, we used two different prompting strategies: few-shot (FCOT) and zero-shot (ZCOT). When refining the model in ZCOT, we attempted to obtain responses straight from the LLM, whereas in FCOT, we used the labelled Knowledge Base as the retriever. GPT-4o and Llama 3.0(Llama-3-70B-Instruct) were used as the models for the entire assessment. It set up the temperature to 0 and a maximum token limit of 1500 for each output.

4.2.1 Approach 01: Zero-shot COT (ZCOT) prompting

Zero-shot prompting (ZCOT) is the prompt engineering technique where language models generate responses for tasks without any new examples or fine-tuning. Table 2 provides additional details about the prompt example we used in our experiments. In this study we concluded that the ZCOT method is not accurate in identifying geospatial entities like GHETTOS (refer to Table 2).

In the above prompt, {tag meaning} refers to the information related to geospatial entities. We have used the following information to enhance geospatial knowledge within the prompt during the inference process:

- LOC: Locations except countries or cities.
- GPE: Geographical locations such as countries or cities.
- CAMP: Concentration camps (Extermination, Transit, Labour)
- GHETTO: Ghettos, the Jewish quarters in cities.
- STREET: Pathways or roads.

{sentence} refers to the sentence that contains values that need to be tagged with geospatial entities.

4.2.2 Approach 02: Retrieval Augmented Generation (RAG)

In the RAG approach, we shared the geospatial knowledge during the prompting process which

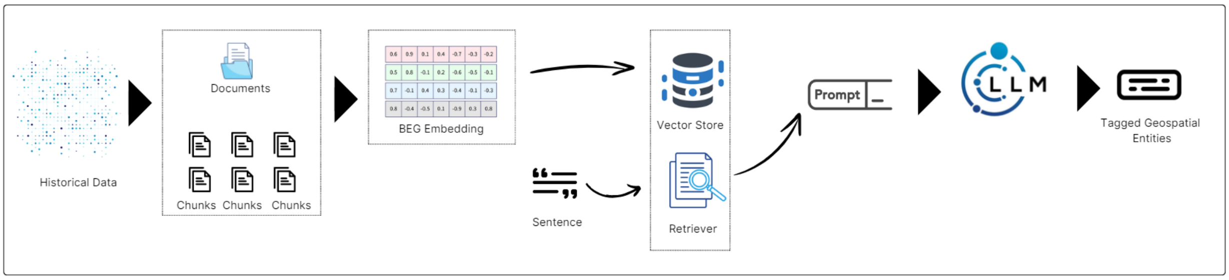


Figure 2: Data-flow of the RAG pipeline.

resulted in better and more accurate responses as this additional knowledge was not available in the training set. The RAG approach is mainly designed using two phases: vector store generation and the retriever with response generation with the selected language models. In this study the following models were used.

Vector store generation and Embedding: The 'BGE small' model from Huggingface was chosen as the embedding for the study, while Chroma DB was employed to store the vectors related to the labelled geospatial data. To preserve contextual meaning during chunking, a recursive character text splitter from LangChain was incorporated to create the necessary data chunks with 2500 tokens, overlapping 50 tokens. These chunks are stored in the vector store once embedded using the embedding model.

Retriever and prompting: Retrieval QA was utilised to build the retriever, with search_kwargs(k) set to '2' and the search_type set to 'similarity'. The similarity search uses cosine similarity to extract the vectors closest to the input sentence we want to tag. This approach allows us to feed data with a similar labelled context to the model, enriching the response generation task with geospatial knowledge. The designed few-shot COT prompt is employed here, with minor adjustments to fit it into the process. **We observe that the main flaw of this approach is that the retriever uses similarity score assessments to retrieve related data based on the sentence context. As a result, sample chunks without the sought word may be returned.** To address the aforementioned issue, we modified the word-level retriever in the next method to use the Knowledge Base.

4.2.3 Approach 03: Few-shot COT prompting with Knowledge Base

The pre-labelled knowledge is incorporated into the inference process by FCOT to enhance the ZCOT quality. In order to obtain the few-shot prompts required for inference, we stored labelled geographical entities in a knowledge graph. To generate responses, the stored knowledge is utilised. The tree structure of the knowledge graph is designed with the place names as the root node and geospatial entities as the first-level parent nodes. Leaf nodes are designed as the list structure containing sample instances of labelled datasets. The above approach has effectively improved the retrieval time of example phrases required for few-shot learning, which can be performed effectively.

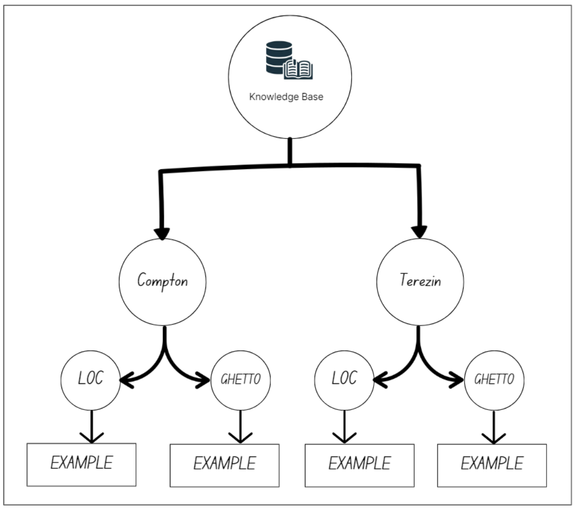


Figure 3: Knowledge base arrangement.

The presence of the target word in the retrieved sentences is considered mandatory for efficient labelling in the few-shot approach. Almost five instances for each entity are retrieved from the KB. If the word is absent, the prompt will function in a zero-shot manner. The detailed workflow of the approach is presented in figure 4.

The amended prompt below shares the additional information retrieved for the second phase.

Table 2: Zero-shot COT Prompt.

Zero-shot COT Prompt-Holocaust Testimonies corpus
Consider the year from 1936-1944. You are going to identify name entity tags for holocaust-specific tags. The list of name entity tags should be {list_of_tags}. Each tag is as follows: {tags_meaning}. Now do the below tasks. 1. Try to identify the most suitable Name entity tag for the word ’NAMEENTITY’ in the GIVEN SENTENCE based on the below criteria: • Analyse the word in front of the ’NAMEENTITY’ tag before you tag. • Understand the complete sentence and try to identify specific factors discussing the word you want to tag. The GIVEN SENTENCE: {sentence}. 2. Return only the GIVEN SENTENCE after assigning the identified tags instead of the word ’NAMEENTITY’. Do not add additional data. Use the following format for the output: ”<Updated sentence with correctly identified name entity tags>”

The below line is introduced as the first chain of thought to the prompt.
Examine the below examples and learn about the appropriate Name entity tags for the words based on the context. Examples are ’{result}’.

In above {result} tag contains the extracted knowledge from the KB. This approach has shown promising results in handling the GHETTO, LOC and CAMP. The detailed analysis of the results is discussed in the Section 5.

The code associated with this research will be made publicly available as part of the supplementary materials accompanying the final version of this paper in case of acceptance for presentation at the conference.

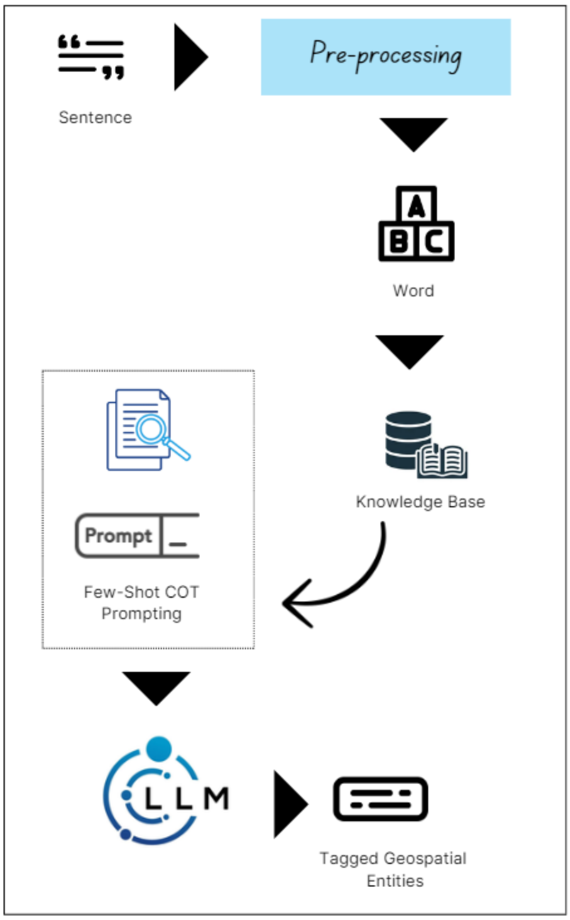


Figure 4: Data-flow of the few-shot COT pipeline.

5 Results and Discussion

Toponym resolution is an essential NLP task when processing oral interviews and testimonies and is at the same time a highly complex task where most computational models fail to pinpoint to the exact geographic references. In this section we first report the evaluation results of the baseline and rule-enhanced transformer methods outlined in section 4. Experiments have been conducted on both datasets also seeking to establish domain-specific adaptability.

Lake District Corpus: As aforementioned, the Spacy model has been used as the rule-based approach and delivered 93.95% recall score by defining the grained tag as PLACE-NAMES (Ezeani et al., 2023b). In previous studies, the list of geographical places and nouns was manually compiled from the lake district corpus and extracted from the corpus based on regex rules defined following a similar approach defined in Table 1. However, this raises another limitation of finding new geographical locations discussed in the corpus which are not in the lexicon. Another limitation in this study was the fact that the entity PLACE-NAMES tag was always misidentified with the LOC and GPE tags, and the evaluation measures were obtained considering all those ambiguities that exist in the text. In our study, we experimented with fine-grained named entities related to geospatial named entities (LOC, GPE, FOREST, RIVER, LAKES and BUILDING).

Holocaust testimonies corpus: The results from the experiments suggest that pragmatic and contextual relationships between words are crucial for differentiating between geographical entities like LOC and GHETTO. Table 4 and Table 5 show that, GHETTO and LOC were misread as GPE, which emphasises how crucial it is for determining the contextual meaning of the words and the way they are used in the testimony.

Table 3: Performance of baseline study using rule-based approach (Spacy Transformer Model).

Entities	Baseline: Spacy-Transformer based		
	Precision	Recall	F1score
LOC	1.0	0.10	0.18
GPE	0.64	0.83	0.72
CAMP	0.77	0.51	0.62
GHETTO	0.00	0.00	0.00
STREET	0.67	0.51	0.58

Table 6 presents the evaluation of the baseline model with ZCOT prompting, RAG approach and the FCOT approach with additional knowledge base. The selected LLMs perform modestly at classifying GPE, CAMP, and STREET entities in the ZCOT prompting approach. However, compared to proprietary and open-source LLMs, the GPT-4o model outperforms the LLAMA model with the same parameter setting. Drawing on standard prompting techniques, the ZCOT approach significantly underperforms at recognising GHETTO, with a notable prediction loss.

As another approach, we combined the RAG pipeline, which targets sentence-level retrievers using the cosine distance. Compared to the baseline approach, recognising the GHETTO tag shows a 0.15 improvement in F1 score, while other entities show a slight improvement in the tagging. This approach has shown that proper retrieval would improve the performance of the tagging process.

The lack of geospatial knowledge is then addressed using a model with the FCOT approach. The word-oriented retriever, which uses a tree structure, is incorporated to extract the most appropriate result. The FCOT approach has shown significant improvements for the GHETTO tag, with an increase of 0.19% in the F1 score, and for the LOC category, with an improvement of 0.08% in the F1 score. Our findings demonstrate that well-crafted prompts, along with a knowledge-sharing approach, can assist the general purpose language models to domain-specific and complex tasks like toponym resolution in the Geo-spacial domain without the need for further pre-training or fine-tuning of LLMs which is cost-effective.

While some studies on disambiguation of toponyms have been carried out, to the best of our knowledge, the present study is the first one using proprietary and open source LLMs in conjunction with augmented prompt engineering techniques. We investigated several techniques to evaluate how well sentence structure similarity identified toponyms and correctly distinguished the provided named entities. However, none of the rule-based (or deep learning) methods produced competitive results. This suggests that with highly unstructured oral transcribed data, structural similarity alone is not sufficient to disambiguate toponyms.

According to our experiments, we integrate the surrounding context of the entity and additional domain-specific information to reduce the ambiguity of the entity as the input prompt and next, augmented prompt engineering techniques execute to identify and disambiguate the toponym within the given domain-specific contexts. Despite the presence of code-mixing, where terms from different languages are within the transcribed texts, these LLMs successfully identified the correct geospatial NEs. This highlights the capability of general-purpose LLMs in handling multilingual and mixed-language scenarios, providing reliable results in both open-source and commercial products.

To improve performance in domain-specific toponym disambiguation, we included advanced prompt engineering techniques as the next phase of our work. Using these advanced prompts lowers computation costs while simultaneously increasing accuracy of overall model. Prompt design can be further optimised to yield more efficient outcomes with less computational requirements, hence simplifying the entire model development process.

As the first study which purely used LLMs to resolve the ambiguities in the toponyms, in the future, this study will be extended to generalise prompts to address the disambiguation of other geographical named entities, including natural landmarks such as rivers, forests, and mountains in the Holocaust testimonies corpus. By expanding the scope to in-

Table 4: Performance comparison between prompt Engineering techniques.(GPT-4o)

Entities	Zero-shot GPT 4o			RAG with GPT 4o			Few-shot COT Prompting		
	Precision	Recall	F1score	Precision	Recall	F1score	Precision	Recall	F1score
LOC	0.57	0.74	0.64	0.54	0.81	0.64	0.63	0.84	0.72
GPE	0.77	0.85	0.81	0.83	0.77	0.80	0.89	0.82	0.85
CAMP	0.95	0.74	0.83	0.90	0.76	0.82	0.88	0.79	0.83
GHETTO	0.62	0.31	0.41	0.59	0.53	0.56	0.61	0.59	0.60
STREET	0.94	0.84	0.88	0.97	0.94	0.95	0.91	0.91	0.91

Table 5: Performance comparison between prompt Engineering techniques.(LLAMA)

Entities	Zero-shot LLAMA			RAG with LLAMA			Few-shot COT Prompting		
	Precision	Recall	F1score	Precision	Recall	F1score	Precision	Recall	F1score
LOC	0.54	0.49	0.51	0.40	0.59	0.47	0.44	0.56	0.49
GPE	0.78	0.78	0.78	0.76	0.79	0.77	0.81	0.75	0.78
CAMP	0.89	0.77	0.83	0.81	0.74	0.77	0.84	0.73	0.78
GHETTO	0.65	0.50	0.57	0.43	0.69	0.53	0.45	0.47	0.46
STREET	0.73	0.91	0.81	0.74	0.92	0.82	0.86	0.90	0.88

clude a broader range of toponyms, we can enhance the model's ability to accurately identify and differentiate between various types of geographical entities. This extension will contribute to a more comprehensive and robust system for geographical named entity resolution, benefiting applications in fields such as geographic information systems (GIS), environmental monitoring, and digital humanities.

6 Conclusion

In this paper, we explored the evolution from traditional methods to state-of-the-art LLMs for toponym resolution in orally transcribed texts, within the context of Holocaust studies. While previous studies have established that domain-specific Deep Learning models often outperform general-purpose LLMs, this study shows that for the challenging task of toponym resolution effective Prompt Engineering incorporating RAG and few-shot learning techniques can help LLMs to surpass high-performing Deep Learning models which is contrary to the initial expectations. In particular, we demonstrate how using labelled data as a knowledge base enriches the inference process, turning few-shot examples into a wealth of information to handle corner cases in geospatial disambiguation. Moreover, leveraging prompts within these models can yield high-quality results at a reduced cost, thereby enhancing the overall feasibility and effectiveness of toponym resolution efforts in this specialised domain.

Limitations

The study is exclusively centred on GPT-4o and Llama3 70B models without any training or fine-tuning. The characteristics of the data used in these models' initial training may have directly impacted the outcomes. Further refinement could enhance the final results by fine-tuning the model with oral and transcribed data. Due to the datasets' constraints and hardware resource limitations, we utilised a limited number of records for the evaluation process, which warrants further exploration.

Table 6: Performance comparison between prompt Engineering techniques. (GPT-4o)

Entities	Baseline			RAG with			Few-shot COT Prompting		
	Precision	Recall	F1score	Precision	Recall	F1score	Precision	Recall	F1score
Forest	0.67	0.70	0.68	0.74	0.62	0.67	0.73	0.69	0.71
River	0.43	0.37	0.39	0.56	0.64	0.58	0.52	0.61	0.56
Lakes	0.57	0.46	0.51	0.66	0.74	0.69	0.78	0.81	0.79
Building	0.62	0.78	0.69	0.81	0.79	0.79	0.85	0.79	0.81

Ethics

The data used for this study were publicly available on the Wiener Holocaust Library webpage and GitHub. Since human annotation occurred during the dataset creation process, all ethical approvals were obtained and cleared.

References

Isuri Anuradha Nanomi Arachchige, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, HIP '23, page 85–90, New York, NY, USA. Association for Computing Machinery.

Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2019. Using recurrent neural networks for toponym resolution in text. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part II 19*, pages 769–780. Springer.

Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2022. A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS International Journal of Geo-Information*, 11(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ian Gregory, Erum Haris, Anthony Cohn, John Stell, Tim Cole, Joanna Taylor, David Bodenhamer, Neil Devadasan, Erik Steiner, Zephyr Frank, and Jackie Olson. 2023a. Towards an extensible framework for understanding spatial narratives. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, GeoHumanities '23, page 1–10, New York, NY, USA. Association for Computing Machinery.

Ignatius Ezeani, Paul Rayson, and Ian N Gregory. 2023b. Extracting imprecise geographical and temporal references from journey narratives. In *Text2Story@ ECIR*, pages 113–118.

Diego Gomes, Ross S Purves, and Michele Volpi. 2024. Fine-tuning transformers for toponym resolution: A contextual embedding approach to candidate ranking. In *GeoExT@ ECIR*, pages 43–51.

Ian Gregory, Christopher Donaldson, Andrew Hardie, and Paul Rayson. 2018. Modeling space in historical texts. In *The Shape of Data in Digital Humanities*, pages 133–149. Routledge.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics. *Language resources and evaluation*, 54:683–712.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018. What's missing in geographical parsing? *Language Resources and Evaluation*, 52:603–623.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Mahammed Kamruzzaman and Gene Louis Kim. 2024. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. *arXiv preprint arXiv:2404.17218*.

Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. 2021. Multi-level gazetteer-free geocoding. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 79–88.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023. Geolm: Empowering language models for geospatially grounded language understanding.

Z. Liu, K. Janowicz, L. Cai, R. Zhu, G. Mai, and M. Shi. 2022. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3:9.

Ruei-Shan Lu, Ching-Chang Lin, and Hsiu-Yuan Tsao. 2024. Empowering large language models to leverage domain-specific knowledge in e-learning. *Applied Sciences*, 14(12).

Ali Maatouk, Kenny Chirino Ampudia, Rex Ying, and Leandros Tassiulas. 2024. Tele-llms: A series of specialized large language models for telecommunications. *arXiv preprint arXiv:2409.05314*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 1802. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*, 42.

Benjamin J. Radford. 2021. Regressing location on text for probabilistic geocoding. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 53–57, Online. Association for Computational Linguistics.

Breno Dourado Sá, Ticiana Coelho Da Silva, and José Antônio Fernandes de Macêdo. 2022. Enhancing geocoding of adjectival toponyms with heuristics. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 37–45.

Yuval Solaz and Vitaly Shalumov. 2023. Transformer based geocoding.

Joanna E Taylor and Ian N Gregory. 2022. *Deep Mapping the Literary Lake District: A Geographical Text Analysis*. Rutgers University Press.

Zehui Zhao, Laith Alzubaidi, Jinglan Zhang, Ye Duan, and Yuantong Gu. 2024. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, 242:122807.