# Prompting the Past: Exploring Zero-Shot Learning for Named Entity Recognition in Historical Texts Using Prompt-Answering LLMs

**Crina Tudor** and **Beáta Megyesi** and **Robert Östling**
Stockholm University, Sweden
**Correspondence:** crina.tudor@ling.su.se

## Abstract

This paper investigates the application of prompt-answering Large Language Models (LLMs) for the task of Named Entity Recognition (NER) in historical texts. Historical NER presents unique challenges due to language change through time, spelling variation, limited availability of digitized data (and, in particular, labeled data), and errors introduced by Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) processes. Leveraging the zero-shot capabilities of prompt-answering LLMs, we address these challenges by prompting the model to extract entities such as persons, locations, organizations, and dates from historical documents. We then conduct an extensive error analysis of the model output in order to identify and address potential weaknesses in the entity recognition process. The results show that, while such models display ability for extracting named entities, their overall performance is lackluster. Our analysis reveals that model performance is significantly affected by hallucinations in the model output, as well as by challenges imposed by the evaluation of NER output.

## 1 Introduction

Named Entity Recognition (NER), oftentimes also referred to as Named Entity Recognition and Classification (NERC), is in essence a token classification task that aims to extract various types of named entities from a given written source. The choice of how fine-grained we want our analysis to be dictates the number of different labels we want to extract; a coarse-grained analysis would only look at names of people, locations and organizations, for example, while a more fine-grained approach would include dates, events, artifacts, monetary values etc.

While NER is by no means a solved problem in NLP, there have been numerous efforts made to provide tools for modern languages. However, such tools have significant gaps in terms of NER resources (e.g. Jørgensen et al. (2020); Hvingelby et al. (2020)), and many are still ongoing (Ingólfsdóttir et al., 2019), which only highlights the importance of further research in this domain.

At the same time, NER for historical texts faces several unique challenges in its own right. OCR errors are common due to the poor quality of old prints, leading to misrecognized characters and words (Ehrmann et al., 2023). The evolution of language over time, with outdated vocabulary, spelling variations, and different grammar rules, complicates entity recognition, especially since historical texts often lack labeled datasets, making supervised learning difficult. Models trained on modern data struggle with domain transfer to text from antiquated sources, as historical contexts and naming conventions differ significantly. A common example of this phenomenon is toponyms changing through time (e.g. Byzantium, Istanbul, Constantinople); so while we refer to the same geographical location, the name differs, and such changes are oftentimes not linked to each other in databases in order to indicate equivalence. Nonstandardized naming, ambiguity in references, and the need for contextual understanding further hinder accurate recognition. Additionally, historical texts are often multilingual, requiring models to handle archaic language variants from several languages at the same time. These factors, combined with cultural and diachronic variations in entity references, make NER for historical texts a complex and challenging task.

This study is motivated by the proven benefits of prompt-based learning (Le Scao and Rush, 2021). The goal of this paper is to further the development of NERC systems for historical texts. Specifically, we want to explore the potential of prompt-answering LLMs for extracting NEs from historical text in a zero-shot scenario, using historical newspaper data in English, German and French. We



*Proceedings of LaTeCH-CLfL 2025*, pages 216–226

investigate this research avenue in order to counteract the costly nature of creating manually annotated NER datasets from scratch, while also leveraging the potential of prompt-answering LLMs in low resource settings.

In our exploration, we aim to address the following research questions:

- How effective are prompt-answering LLMs in recognizing named entities in historical texts?
- What types of errors do generative prompt-answering models make when extracting named entities in a zero-shot context?
- What effect do hallucinations have on model performance in the context of NER extraction and evaluation?

At the same time, we identify several potential benefits of this work for future research. By enabling the creation of historical social networks, for example, we can uncover and analyze relationships and interactions among individuals across time periods. Additionally, enhancing archival annotation improves the accessibility and usability of historical documents, allowing researchers to extract meaningful insights more efficiently. Such methods facilitate cultural and historical research by automating large-scale annotation, significantly reducing the time and cost associated with manual processes, thereby enabling access to diverse historical narratives.

## 2 Background

Earlier work on historical NER has primarily been conducted on monolingual language models and various choices of model architecture and data sources. Moreover, transformer-based models have been gaining significantly more traction. Here, the trend leans towards using off-the-shelf modern LMs, which are later fine-tuned with historical labeled data for the task of NER (Arnoult et al., 2021), but there are also studies experimenting with data sourced entirely from historical text, and fine-tuned on modern labeled data (Tudor and Pettersson, 2024). Moreover, the trend has been to branch out towards multilingual models in order to take advantage of their transfer learning capabilities (Schweter et al., 2022).

The biggest hurdle in the way of designing accurate and high-performing NER systems seems to be the lack of annotated quality data. Ideally, we would want to have large amounts of manually annotated datasets which are curated using expert knowledge. The process of obtaining such data is, however, expensive both in terms of time and resources needed for such endeavors. Furthermore, enormous amounts of data that could be used for annotation reside in libraries and archives, and have yet to be digitized - which is another time-consuming and costly process. While there are significant efforts being made to contribute to this gap in the field, the vast majority are focused around texts from modern sources. Such examples include the Icelandic NER corpus (Ingólfsdóttir et al., 2019), its Norwegian counterpart (Jørgensen et al., 2020), the Swedish SUC (Källgren and Eriksson, 1993; Språkbanken Text, 2024), or the Danish DaNE (Hvingelby et al., 2020).

Naturally, new research directions have come forth, aiming to circumvent the data scarcity issue. The expensive nature of supervised learning prompts for exploration into the capabilities of few-shot learning for LM architectures (Perez et al., 2021). With the recent emergence of prompt-answering models and their impressive few-shot learning abilities (Schick and Schütze, 2021), several studies have attempted to explore their performance on NER (Huang et al., 2020). Moreover, while Schick and Schütze (2021) explore true few-shot learning where there is no development set available for hyperparameter tuning and additional prompt engineering, and highlights its potential for future applications, new research on prompt engineering for few-shot NER is quick to emerge (Liu et al., 2022).

A similar exploration to the one we show in the present paper has been conducted by Arnoult et al. (2021) for Dutch historical text. Their dataset was created based on letters from the Dutch East India Company dating from the 17th and 18th century. In their paper, they compare the performance of monolingual (BERTje, RobBERT) and multilingual (mBERT, XLM-R) language models. The study finds that multilingual models outperform monolingual ones in handling the language variations and cross-lingual transfer needed for historical texts. Overall, both model types benefit from combining historical texts and editorial notes, with multilingual models showing more robustness across various text types.

More recently, González-Gallardo et al. (2023) investigate how language models like GPT-3.5 handle entity recognition in historical documents, highlighting also code-switching between French and Ancient Greek. The study points out that while GPT-3.5 is trained in over 100 languages, it struggles with unrepresented languages such as Ancient Greek. The paper discusses challenges such as the model's difficulty understanding mixed-language texts and the limitations of historical archives that remain inaccessible to models, impacting their performance in recognizing historical entities.

The expensive nature of labeled data for training and evaluation makes the prospect of zero-shot and few-shot learning significantly more appealing for NER research. The basis of our exploration lies in a study conducted by Toni et al. (2022). The paper uses labeled data from the CLEF-HIPE 2020 dataset (Ehrmann et al., 2020), which is an open-access OCR-ed newspaper corpus annotated for NER. The dataset contains Swiss and Luxembourgish newspapers from 1790 to 2010 in English, German, and French. The authors focus on zero-shot NER using T0++ (Sanh et al., 2021), and only use data up to 1950 at the latest in order to keep the focus on the historical aspect of their exploration. Their study shows that, while the model shows some capacity of extracting NEs from the given dataset, dealing with historical text poses additional challenges through spelling variation and OCR errors. They also prompt for further investigation of the capabilities of generative LLMs in this given context.

## 3 Method

Our exploration can be seen as a three-step process. The first phase is to run all of our chosen models on the same dataset as the original study described in Toni et al. (2022), which we describe in Section 3.2. The second step is to evaluate and assess the kind of errors that the models are prone to by doing a manual examination of the output of each model. Third and last, we aim to address some of the more common causes of errors in the model output and re-evaluate in order to see how that affects model performance.

### 3.1 Model selection

While Toni et al. (2022) focus on models from the T0 family, specifically T0++, we expand into a more comparative analysis using some of the state-of-the-art prompt-answering LLMs, such as T5, mT5, BLOOMZ and Aya. We limit ourselves to publicly available models of at most 13B parameters, as this approaches the practical limit of most researchers who want to annotate significant amounts of historical text data. We provide more specific information about the models in Table 1. The choice of models is motivated by their capacity for prompt-based learning, as well as their reported performance in zero-shot learning scenarios on other NLP tasks, such as Natural Language Inference, Coreference Resolution or Word Sense Disambiguation. Furthermore, we choose two versions of each model which vary in terms of size - a smaller model of around 3 billion parameters, and a larger version of 10+ billion parameters, wherever applicable. It is important to note here that not all model families have versions that match this requirement exactly, in which case we choose the closest possible variant. The goal here is to see to what extent model size impacts a model's inference capabilities. We summarize all models and their sizes in Table 1.

| ***Model*** | **Parameters** | **Language** |
|---|---|---|
| T0 3B | 3B | English |
| T0 ++ | 11B | English |
| T5 3B | 2.85B | English |
| T5 11B | 11B | English |
| mT5 XL | 3.7B | multilingual |
| mT5 XXL | 13B | multilingual |
| Aya 23 8B | 8B | multilingual |
| Aya 101 | 12.9B | multilingual |
| Bloomz 3B | 3B | multilingual |
| Bloomz 7B1 | 7.07B | multilingual |

Table 1: List of prompt-answering LLMs used, their sizes, along with their main source of training data.

T0 (Sanh et al., 2021) is a prompt-based generative model fine-tuned on multiple NLP tasks and designed to follow instructions directly without needing task-specific fine-tuning. The pre-training for this model is done using a prompt-based setup, meaning that the training examples are converted into prompts using crowd-sourced prompt templates. This particular training setup allows the model to be able to generalize across previously unseen tasks, and it claims to outperform GPT-3 while also being 16 times smaller.

T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2019) is a pretrained generative transformer model that reformulates all NLP tasks as text-to-text tasks, making it highly flexible for various applications like summarization, translation, and classification. The main goal of the T5 architecture is to provide a unified text-to-text format that can easily be transferred across a variety of NLP tasks. The authors evaluate the model on a total of 17 tasks, where T5 either achieves state-of-the-art or competitive results when compared to previous high-performing models.

mT5 (Xue et al., 2020) is a multilingual extension of T5, which was pretrained on data from 101 language. This allows it to handle a wide array of multilingual NLP tasks. The model uses a similar architecture as its monolingual counterpart, and is able to achieve state-of-the-art results on a variety of cross-lingual NLP tasks, such as zero-shot classification or question answering.

BLOOMZ (Muennighoff et al., 2022) is a successor to the original BLOOM (Scao et al., 2023) text generation model. The authors apply Multitask prompted fine-tuning (MFT) to the pretrained multilingual BLOOM to produce fine-tuned variants called BLOOMZ. They find that fine-tuning large multilingual language models on English tasks with English prompts allows for task generalization to other languages that appear only in the pretraining corpus, but that fine-tuning on multiple languages leads to even better performance.

Aya (Üstün et al., 2024) is a transformer-based generative model that follows the same architecture as mT5. Aya is also a massively multilingual LM that has been trained on over 100 languages. When evaluated on unseen tasks, Aya manages to outperform BLOOMZ by almost 10%.

### 3.2 Dataset

In our exploration, we look at the same dataset as Toni et al. (2022), namely HIPE2020[1], using the same cutoff point (i.e. 1950). The dataset consists of newspaper texts from the 18th to the 20th century in English, French and German, which were manually annotated by human experts.

We focus on the coarse-grained tag set in this corpus, namely persons (PERS), organizations (ORG), products (PROD), time (TIME) and location (LOC). While time, person and location are fairly straightforward entities, the labels for PROD and ORG are harder to define in clear terms, and potentially harder to identify in the annotation process. According to the guidelines used for annotation, ORG can refer to organizations that market products or provides services, press agencies or organizations that mainly have an administrative role. In the case of the PROD label, this consists of either media (newspapers, magazines, broadcasts etc.) or doctrines (such as political, religious or philosophical beliefs).

| Label | Count | Percentage |
|---|---|---|
| PERS | 7618 | 31.92% |
| TIME | 851 | 3.57% |
| LOC | 10711 | 44.88% |
| PROD | 662 | 2.77% |
| ORG | 4022 | 16.85% |
| TOTAL | 23864 | |

Table 2: Count of named entities for each label in the dataset, as well as their corresponding percentage from the total.

The data is split by language and time period, with English containing between 2,202 and 4,697 tokens per time interval, German between 6,735 and 12,829 tokens, and French between 8,550 and 16,874 tokens. We provide the count of all named entities in the gold corpus in Table 2.

### 3.3 Experimental setup

The first step that we take in our exploration is to run all the chosen models on the HIPE2020 datasets using the same setup as the one used by Toni et al. (2022). More specifically, we take the script[2] they use in their experiments and we adjust it in order to fit the requirements of our chosen models. We keep the exact same prompt structure in the initial run of the experiments, as well as the same data and label set. We also use the same evaluation schema, with only minor modifications made to the code[3]. The prompting is done in English across all languages in the dataset. We exemplify with templates in Table 3 (see "Original prompt").

Once we prompt all our models to extract NEs from the given text, we proceed to do a manual analysis of the output of each model. At this stage, we make observations of various peculiarities and types of errors that the models return.

Lastly, we attempt to address some of these common errors and run a comparative evaluation of model performance before and after filtering out misleading phenomena – such as hallucinations – in the output for example.

[1]https://impresso.github.io/CLEF-HIPE-2020/datasets.html

[2]https://github.com/bigscience-workshop/historical_texts/blob/master/NER/parallel-GPUs/NER_parallel-GPUs-fuzzy.py

[3]https://github.com/crina-t/LaTeCH2025

| | |
|---|---|
| **Original prompt** | Input: [SENTENCE] In input, what are the names of [ENTITY TYPE]? Separate answers with commas. |
| **Modified prompt** | Input: [SENTENCE] In input, what are the names of [ENTITY TYPE]? Separate answers with commas without changing the original input text. |

Table 3: Prompt templates according to the original study (top) as well as after being modified to attempt avoiding changes in the original input text (bottom).

## 4 Results

We apply each model to our NER task in a zero-shot setup to assess their baseline performance without extensive customization. We used prompts designed to extract named entities across multiple languages, testing the models' ability to handle common entity types. A manual analysis of the output of each model reveals several systematic types of errors that take a toll on overall model performance.

A common case is models retaining parts of the prompt and regurgitating them as output, instead of outputting parts of the actual input text. For example, out of 50,495 potential entities annotated by T5 3B, over 80% of them contain the words "input" or "in input". The same phenomenon is observed in T5 11B, but to a lesser degree – only 56% of the extracted entities keep the word "input". When looking at its multilingual counterpart, we notice that mT5 displays the same anomaly. Out of all output NEs from mT5 3B, 51% contain at least one occurrence of the word "input", which drops to 49% in the case of mT5 13B.

This carries over in the case of both versions of the BLOOMZ model as well, but to a different extent. Instead of just keeping parts of the prompt text, the model takes the entire content of the prompt, including the input sentence, and splits it into segments using commas as delimiters. We believe that this could be the case due to the model not properly capturing sentence boundaries, which has been known to cause problems for this particular model family (Muennighoff et al., 2022).

In light of these observations, we are unable to calculate reliable performance scores for these models (F1 $< 1\%$), and we therefore no longer include these 6 models in the rest of our analysis. We focus instead on T0 and Aya, and more specifically T0++ and Aya 101, as larger model versions seem to lead to slight improvements in performance.

### 4.1 Hallucinations

A significant source of errors that we encounter in model output are hallucinations. In the context of LLMs, hallucinations can have different forms and interpretations. However, for our purposes, we define hallucinations as instances where the generated output seems incoherent, irrelevant, or deviates from the given source content, following the categorization provided by Huang et al. (2025).

Consequently, we conduct experiments to see what amount of the extracted entities are not actually part of the sentence given as input, as is the case in examples a) and b) in Table 4. We do this by iterating through all entities in the model output and matching them against the target sentence, removing spaces in order to avoid potential noise. Table 5 shows that about half of the entities extracted by T0++ are not strictly part of the input sentence, while Aya 101 scores a little more than 11% in terms of total hallucinated entities.

In order to see if we can circumvent this issue, we attempt to tweak the original prompt in order to encourage the model to stick to words from the input sentence exclusively (see "Modified prompt" in Table 3). While this does lower the total number of extracted entities, the overall percentage for T0++ increases slightly after this modification. In the case of Aya 101, the change in prompt wording does seem to lower the overall occurrence of hallucinations by about 2.25%.

It is important to mention here that there are nuances in what we count as being a hallucinated entity in our evaluation. A negative result (i.e. entity not in input sentence) can also mean that the model automatically converted the historical spelling to its modern counterpart. Similarly, the model can simply make small edits to the extracted span from the input, which also impedes the evaluation process (e.g. "les conversations particulières" in the original text, but the model extracts "conversation particulières"). In some cases, it can even happen that the model translates the original language into English (e.g. from "Un vin d ' honneur fut offert

| | | |
|---|---|---|
| **a)** | SENTENCE: | A S my enquiries arc extended into the nature of anti - federalifm , and of the motives which acftuate fuch people , I become more convinced , that my deflgn of a general apology for them is very meritorious , and ought to have been made long ago ; and I cannot conceive the reafon why it hath never been publicly attempted , unlefs it be the excreme difficulty of an inveftigation . |
| | PROMPTED FOR: | PERS |
| | OUTPUT: | John Quincy Adams |
| | GOLD: | N/A |
| **b)** | SENTENCE: | After enduring weeks of suffering in the hospital at Moscow , , with no hopes of relief ex 0̆0ac cept in death. |
| | PROMPTED FOR: | TIME |
| | OUTPUT: | 13 and 14 June |
| | GOLD: | N/A |
| **c)** | SENTENCE: | OBITUARY James Hargis James Hargis , one of the most prominent and highly respected citi 0̆0ac zens of this section of the state , died Monday at his home at Granville . |
| | PROMPTED FOR: | LOC |
| | OUTPUT: | Granville |
| | GOLD: | Granville |
| **d)** | SENTENCE: | A . C . MATTEE.SON DEAD Well Known Farmer Passes Away AfUr Much Suffering2̆014Fu 0̆0ac neral on Wedding An 0̆0ac niversary . |
| | PROMPTED FOR: | PERS |
| | OUTPUT: | A. C. Matteson |
| | GOLD: | A . C . MATTEE.SON |

Table 4: Examples of model output as extracted by using T0++, alongside the original input sentence, the type of entity requested through prompting, and the corresponding gold standart (where applicable).

| | | **T0++** | | | | **Aya 101** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Original prompt | | Modified prompt | | Original prompt | | Modified prompt | |
| Total | extracted | 76999 | | 74815 | | 93937 | | 85888 | |
| Total | hallucinations | 38240 | 49.66% | 37374 | 49.96% | 10379 | 11.05% | 7556 | 8.80% |
| Of which | PERS | 6975 | 9.06% | 6464 | 8.64% | 1283 | 1.37% | 983 | 1.14% |
| | TIME | 12247 | 15.91% | 11651 | 15.57% | 3775 | 4.02% | 2900 | 3.38% |
| | LOC | 4717 | 6.13% | 5008 | 6.69% | 1091 | 1.16% | 885 | 1.03% |
| | PROD | 8164 | 10.60% | 8236 | 11.01% | 2940 | 3.13% | 1770 | 2.06% |
| | ORG | 6137 | 7.97% | 6015 | 8.04% | 1290 | 1.37% | 1018 | 1.19% |

Table 5: Counts of hallucinated entities for the T0++ and Aya 101 models. We present hallucinations for each label as percentage of the total.

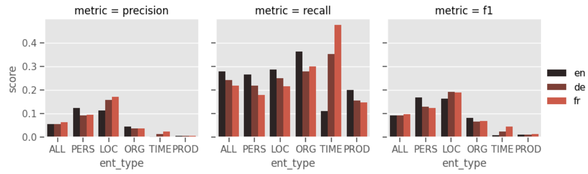
dans la salle des Chevaliers [...]", the model extracts "wine" instead of the original "vin" as an entity).

A hallucinated result could also consist of different parts of the prompt that get marked as entities - such as the entity label itself being extracted as an entity, or other parts of the prompt being kept together with the output, as previously discussed in the beginning of Section 4.

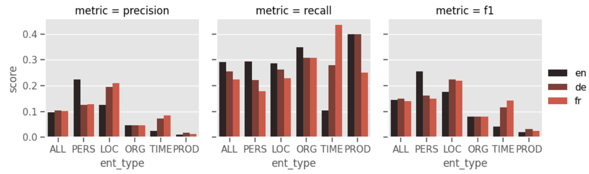
Lastly, we try to filter out these entities which were deemed to be hallucinations, and calculate model performance in terms of precision, recall and F1 score. We present the results for T0++ before and after filtering hallucinations, as well as before and after modifying the original prompt, in Figure 1, and for Aya 101 in Figure 2.

## 5 Discussion

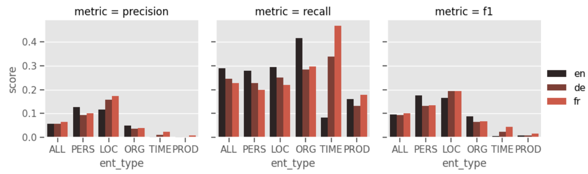
Our results reveal that, while prompt-answering models are able to extract named entities in a zero-



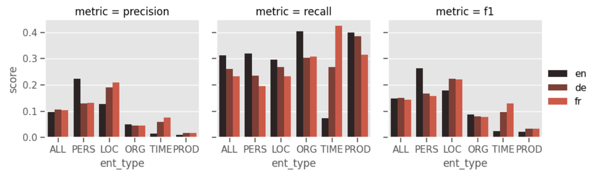

(a) Original prompt




(b) Original prompt, filtered

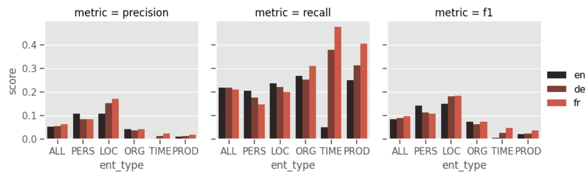



(c) Modified prompt

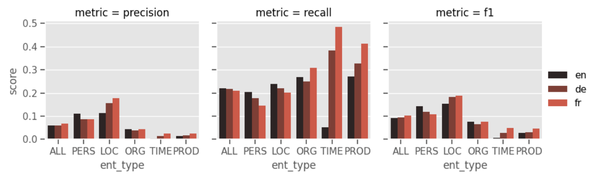



(d) Modified prompt, filtered

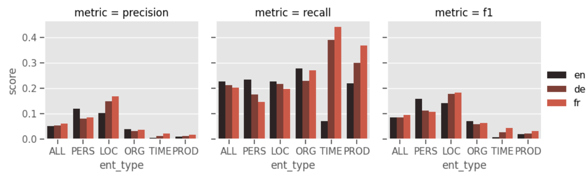
Figure 1: Results for T0++, using the original prompt and our modified version, both before and after filtering hallucinated entities.



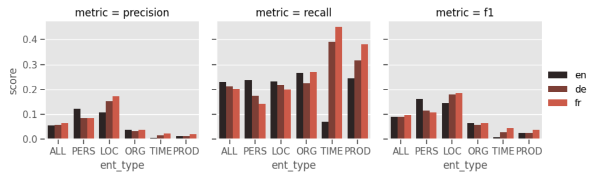

(a) Original prompt




(b) Original prompt, filtered




(c) Modified prompt




(d) Modified prompt, filtered

Figure 2: Results for Aya 101, using the original prompt and our modified version, both before and after filtering hallucinated entities.

shot setting, their overall performance is significantly below what is considered state-of-the-art. This is in part due to errors in the source text, hallucinations produced by the model, or the general difficulty in evaluating NER systems (Fort et al., 2009), especially in a historical and multilingual context (Ehrmann et al., 2020).

Frequent OCR errors introduce unpredictable variations in the spelling of "gold" words, including inconsistencies in spacing, letter placement, and diacritics. T0 automatically corrects these during its predictions, which hinders our ability to match its answers accurately with the corresponding tokens in the sentence. This is exemplified in sentence d) in Table 4, where the model automatically corrects the formatting issues introduced during the OCR process.

Another hurdle in the way of effective NE extraction and evaluation is the frequent occurrence of hallucinations in the model output. Filtering out hallucinated entities does lead to an increase of around 5% in overall F1 score for T0++ (see Figure 1), and to a lesser extent in Aya 101 as well (see Figure 2). However, the overall results are still around the same ranges as before, which only highlights the difficulty of evaluating NER spans accurately, as well as the model's tendency to overgenerate rather than not provide an output at all. This is made evident by examples a) and b) in Table 4, where the model outputs entities that match the requested label, but which are not part of the input sentence.

Moreover, the relatively uniform distribution of hallucinations among labels supports the assumption that T0 models tend to produce non-empty outputs, and therefore over-generate rather than provide a blank answer or no answer at all (Toni et al., 2022). The same phenomenon has been observed across all investigated model families, including T5, mT5 and BLOOMZ.

It is also important to note that Aya 101 achieves higher recall scores than T0++ for French and German, likely due to the fact that it was trained on multilingual data as opposed to English exclusively. Therefore, while the model might not be able to label the entities correctly, it is more likely to extract entities in languages other than English.

The overall effect of prompt engineering and filtering of hallucinations is not to be overlooked either. Both of these approaches lead to small improvements in model performance, which prompts for further exploration in this direction.

## 6 Conclusions and Future Work

In this paper, we explore the zero-shot capabilities of prompt-answering LLMs for NER on historical text.

Our study shows that, while prompt-answering LLMs display some capacity to automatically extract NEs, they do not reach satisfactory enough results for further use (e.g. reliable automatic annotation of archival text). Moreover, we also highlight the models' tendency to produce output even in scenarios where it generates false positive results, and we draw attention to the extensive amount of hallucinations produced by the models. Lastly, we attempt to explore the effect that hallucinations have on model performance by conducting a comparative evaluation after filtering them from model output.

The main contribution resulting from this approach is enhancing the understanding of LLMs' limitations and capabilities in historical NER tasks, providing valuable insights for improving model reliability. Our findings advance historical NER research by broadening the model comparison, extensive error analysis, testing prompt modifications, and addressing hallucination issues.

In future work, we would be keen to investigate the effects of prompt engineering on few-shot NER for historical text, with the hope of benefiting from the proven advantages of prompt-based learning (Le Scao and Rush, 2021). Adjusting the way we feed our prompts into the model can also affect the overall model performance, as previously shown in Liu et al. (2022). Since the model has the tendency to over-generate, and at times it provides an answer extracted form the prompt rather than the input text itself, it could potentially be more beneficial to treat prompting as a two-step process, where we first provide the model with the prompt, and then input the text we want to work with as a secondary step.

Another possible avenue for research is to look into what would be the minimum amount of data or examples required for few-shot or zero-shot learning in historical NER tasks using LLMs without having to compromise on performance. Lastly, since it is common practice for current state-of-the-art models to be released in "families" consisting of various sizes of the same ground architecture, it could also be relevant to experiment with how more variation in parameter size affects the capabilities of such prompt-answering LLMs – including, but not limited to, the model families already mentioned in this paper. A final way forward would be to ensure that the LLM used has seen sufficient amounts of historical text and, if possible, NER examples in historical texts during training.

This study highlights the potential of generative models in improving access to and the analysis of historical texts, aiding in digital humanities efforts, as well as in archival and historical research, while also drawing attention to some of their potential pitfalls.

## Acknowledgments

The project is financed by the Swedish Research Council, partially by DECRYPT - Decryption of Historical Manuscripts (grant 2018-06074), and partially by The Swedish Graduate School of Digital Philology (grant 2022-06343). The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

## References

Sophie I. Arnoult, Lodewijk Petram, and Piek Vossen. 2021. "Batavia asked for advice. Pretrained language models for Named Entity Recognition in historical texts.". In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Litera-*

*ture*, pages 21–30, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310, Cham. Springer International Publishing.

Karen Fort, Maud Ehrmann, and Adeline Nazarenko. 2009. Towards a methodology for named entities annotation. *ACL-IJCNLP 2009 - LAW 2009: 3rd Linguistic Annotation Workshop, Proceedings*.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G. Moreno, and Antoine Doucet. 2023. Yes but.. Can ChatGPT Identify Entities in Historical Documents? *Preprint*, arXiv:2303.17322.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2020. Few-shot named entity recognition: A comprehensive study. *Preprint*, arXiv:2012.14978.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Svanhvít Lilja Ingólfsdóttir, Sigurjón Thorsteinsson, and Hrafn Loftsson. 2019. Towards high accuracy named entity recognition for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 363–369, Turku, Finland. Linköping University Electronic Press.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Gunnel Källgren and Gunnar Eriksson. 1993. The linguistic annotation system of the Stockholm - Umeå Corpus project. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Andy T. Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. QaNER: Prompting Question Answering Models for Few-shot Named Entity Recognition. *Preprint*, arXiv:2203.01543.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Preprint*, arXiv:2105.11447.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg

Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmBERT: Historical Multilingual Language Models for Named Entity Recognition. *Preprint*, arXiv:2205.15575.

Språkbanken Text. 2024. Sucx 3.0.

Francesco De Toni, Christopher Akiki, Javier de la Rosa, Clémentine Fourrier, Enrique Manjavacas, Stefan Schweter, and Daniel van Strien. 2022. Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0. *Preprint*, arXiv:2204.05211.

Crina Tudor and Eva Pettersson. 2024. People and places of the past - named entity recognition in Swedish labour movement documents from historical sources. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 185–195, St. Julians, Malta. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.