Computational Linguistics in the Netherlands Journal 14 (2025) 255-278 Submitted 12/2024; Published 07/2025

LLMs as chainsaws: evaluating open-weights generative LLMs for extracting fauna and flora from multilingual travelogues

Tess Dejaeghere[*], [***] TESS.DEJAEGHERE@UGENT.BE
Els Lefever[***] ELS.LEFEVER@UGENT.BE
Julie Birkholz[*], [**] JULIE.BIRKHOLZ@UGENT.BE

[*] *Ghent Center for Digital Humanities (GhentCDH), Ghent University*

[**] *Royal Library of Belgium (KBR), Belgium*

[***] *Language and Translation Technology Team (LT3), Ghent University*

Abstract

Named Entity Recognition (NER) is crucial in literary-historical research for tasks such as semantic indexing and entity linking. However, historical texts pose challenges for implementing said tasks due to language variations, OCR errors, and poor performance of off-the-shelf annotation tools. Generative Large Language Models (LLMs) present both novel opportunities and challenges in humanities research. These models, while powerful, raise valid concerns regarding biases, hallucinations, and opacity - making their evaluation for the Digital Humanities (DH) community all the more urgent. In response, we propose our work on the evaluation of 3 quantized open-weights LLMs (mistral-7b-instruct-v0.1, nous-hermes-llama2-13b, Meta-Llama-3-8B-instruct) through GPT4ALL for NER on literary-historical travelogues from the 18th to 20th centuries in English, French, Dutch, and German. All models were assessed both quantitatively and qualitatively across 5 incrementally more complex prompts - revealing common error types such as bias, parsing issues, the addition of redundant information, entity adaptations and hallucinations. We analyse prevalent examples per language, century, prompt and model. Our contributions include a publicly accessible annotated dataset, pioneering insights into LLMs' performance in literary-historical contexts, and the publication of reusable workflows for utilizing and evaluating LLMs in humanities research.

1. Introduction

Analyzing named entities in text corpora is a common task in a historians' and literary scholars' research workflow to provide insights into relevant people, places and objects in their respective Zeitgeist. As a result, named entity recognition (NER) has become a convenient aid in expediting this process for larger literary-historical corpora - with applications ranging from semantic document indexing to network analysis and entity linking among others (Ehrmann et al. 2021).

However, treating works of literary-historical origin as data from which to extract information comes with a range of methodological and technical challenges related to the historical nature of languages, multilinguality, and the quality of both digitization processes and off-the-shelf annotation tools (McGillivray et al. 2020, Moretti 2013). In the realm of information extraction (IE) for the literary-historical domain, contemporary NER systems developed in and beyond Digital Humanities (DH) typically fall into three categories: rule-based, feature-based (machine learning), or neural-based (deep learning) (Ehrmann et al. 2021). Technical mastery and a profound understanding of NLP modeling and evaluation practices are essential prerequisites for developing and deploying these systems effectively and correctly (Polak and Morgan 2024). At the time of writing however, there is a burgeoning interest in generative Large Language Models (LLMs) as information extraction systems. Propelled into the spotlight by accessible and user-friendly chat interfaces like ChatGPT,

these models enable users to engage with training data using natural language, revolutionizing communication paradigms and propagating a wide adoption of AI-tools across text-based tasks. Recent efforts have explored and assessed generative LLMs' performance for information extraction tasks across various linguistic spaces and domains with variable results; but its application on literary-historical text material is still in its early stages of exploration (Xie et al. 2023, Li and Zhang 2023, Xu et al. 2023, Sarmah et al. 2023, Li et al. 2023, Han et al. 2023).

2. Challenges of LLMs for DH

In the field of the Humanities, chat-based LLMs such as ChatGPT have shaken the very foundations of this creativity-permeated discipline; raising ethical questions regarding its use in historical research, education and text production (Rane 2023, Spennemann 2023). These models, trained on vast troves of online data, effectively (seek to) mimic human reasoning, serving as a generalized simulacrum of cognitive processes. However, this interaction with our own shadows unwillingly unveils darker facets, as these models inadvertently perpetuate societal biases, stereotypes, and other ethically dubious content present in their training data. The latter are often not fully disclosed by models' creators, which raises the normative question of whether language models should reflect or correct existing inequalities across tasks (Stammbach et al. 2022, Kirk et al. 2021). Furthermore, hallucinations (the propensity of generative models to produce output which is factually incorrect or nonsensical) may result in the generation of unsolicited information and, ultimately, the accelerated dissemination of historical untruths (Minaee et al. 2024, Rane 2023).

Not only the training data of the models, but also the way it is prompted is known to exert a large influence on the generated output. The inclusion of labeled examples (i.e., few-shot prompting), contextual information, knowledge bases (retrieval augmented generation) within prompts as well as the order of the information significantly impacts both structure and content of generated replies (Minaee et al. 2024, Petroni et al. 2019, Liu et al. 2023, Kojima et al. 2023, Polak and Morgan 2024). Compounding this challenge is the prevalence of paid services such as OpenAI offering user-friendly cloud-based models which produce non-replicable stochastic output, while more adaptable open-weights alternatives often require more technical skills to implement. However, the widespread integration of closed-source models in research settings raises ethical concerns regarding privacy and the potential monopolization of applications by corporations such as OpenAI, Meta and Google (van Dis et al. 2023, Workshop et al. 2023).

While LLMs' pitfalls regarding hallucinations, bias, prompt formats and unreplicability are clear, the inherent nature of literary-historical text further complicates IE and its evaluation in general, as literary texts are known to be extraordinarily complex to annotate due to their subjective nature and unique stylistic properties (Kleymann and Stange 2021, Ivanova et al. 2022, Ehrmann et al. 2021). Figurative language such as metaphors, personification and metonymy; stylistic and language-specific peculiarities across authors' works, the historical variety space in which they reside and the highly specialized research needs of literary scholars and historians hamper a standardization of annotation practices across the entire literary domain. Rather correctly, this raises the question whether annotation scheme normalization should even be a goal to strive for in a context which is highly dependent on specialized research questions and the spurious availability of digitized historical data (Bamman et al. 2019, Plank 2022). Indeed, as put by Rebora (2023), we have to acknowledge the "continuous dynamics between the construction of a model and the confrontation with a reality that always escapes it — the same dynamics that, in the end, sustains any theorization about literature.".

3. Opportunities of LLMs for DH

Despite these challenges, generative LLMs also present new opportunities for the field of digital humanities (DH). Zero-shot prompting and in-context learning (ICL) offer a promising avenue for researchers to develop IE systems without the need for extensive annotated data, a pressing issue in creating pipelines for lesser-resourced domains and languages (Ehrmann et al. 2021, McGillivray et al. 2020). While it must be said that zero-shot models are typically still outperformed by Supervised Fine-Tuning (SFT) systems, their true capacity lies in the way they can be prompted and adapted through natural language, marking a significant paradigm shift and significantly lowering the threshold for (digital) humanists lacking prior NLP training to start digging in datasets which supersede specialists' manual processing capacities (Xu et al. 2023, Karjus 2024, Han et al. 2023).

As a consequence, LLMs hold promise as a foundational element in answering the calls for adaptable grey-box researcher-in-the-loop methodologies for IE in DH - and can potentially be more easily aligned with the highly individual text analysis needs of literary scholars than data-hungry discriminative machine learning models (D'Aniello et al. 2022, Jacobs 2019, McGillivray et al. 2020). Breaking the chains of machine avoidance and embracing an exploratory approach to evaluating and applying generative LLMs in DH is thus crucial: not only for the humanities community to assess the effects this technology will have on their research practices in the future - but also to foster their active involvement in the red-hot debate surrounding generative LLMs from the outset (Rebora 2023). Complementing rather than replacing the workflow of literary scholars and historians, the integration of this powerful annotation tool may and will never be to sideline or overpower researchers in the humanities, but rather serves to strengthen their increasingly pivotal roles as societal (and, by future extension, technological) critics while saving expenses on their most precious commodity: that of time (Chun and Elkins 2023, Karjus 2024).

4. Research objectives

In response to the growing demand for IE frameworks and evaluation methods tailored to specialized domains - our study presents a comparative evaluation of three open-weights generative models for the NER task applied to 18th to 20th-century travel literature in Dutch, German, French and English (Rebora 2023, Xu et al. 2023). We pay particular attention to evaluating different prompting strategies across the languages and centuries in our corpus, and perform a qualitative scrutiny of the LLMs' output to discern commonly occurring errors. The goal of our research is thus not model optimization for the task at hand, but rather an exploration of their raw capabilities and output in order to assess their applicability range across languages. Eventually, we aim to formulate an answer to the following research questions:

1. How effectively do open-weights LLMs (Mistral-7B-Instruct-v0.1, Nous-Hermes-Llama2-13B, and Meta-Llama-3-8B-Instruct) perform in extracting mentions of fauna and flora from literary-historical travelogues in Dutch, German, French and English across prompting strategies, and which errors occur in their output?
2. Which insights can we infer for developing future error mitigation strategies and grey-box workflows in Digital Humanities?

5. Methodology

This section outlines the methodology used in the study, detailing 1) **the collection and annotation** of our travelogues dataset, 2) **the prompting strategy** we developed and tested across three open-weights LLMs (Mistral-7B-Instruct-v0.1, Nous-Hermes-Llama2-13B, and Meta-Llama-3-8B-Instruct), 3) **the quantitative and qualitative evaluation strategies** we applied and 4) **the dissemination strategy** we used to publish our data, annotations and code in Jupyter Notebooks.

Llama 3 8B is known for its robust multilingual instruction-following capabilities, while Mistral outperformed earlier Llama generation models on multiple benchmarks. Hermes-llama2 was included for its uncensored training, which may be an advantage in processing highly biased input such as historical travelogues.

5.1 Data

As a first step, a dataset of travelogues from a range of online repositories was collected, resulting in a corpus of 3320 texts across the languages English, French, Dutch and German - ranging from the 18th to the 20th centuries as shown in Table 1:

1. Travel-related texts from the Biodiversity Heritage Library[1] were scraped via API using travel-related terms and primarily feature non-fictional travel reports by biologists and naturalists.

2. The subcollection sourced from DBNL (Digitale Bibliotheek voor Nederlandse Letteren) [2] consists mainly of Dutch stories and reports on colonial explorations by Dutch-speaking settlers.

3. Italian travel reports comprise narratives about Italy written by English authors in the 1930s (Sprugnoli 2017).

4. The Arctic Travellers dataset was manually collected from the Internet Archive[3].

5. Non-fictional travel reports were gathered from Project Gutenberg[4].

6. A set of German travelogues from the Travelogues project, available for download on their GitHub repository, were automatically compiled by domain experts (Rörden et al. 2020)[5].

The travelogues feature diverse genres such as nature writing, travel memoirs, journals, and poetry.

Language	**18thC**	**19thC**	**20thC**	**Total**
English	41	782	668	**1,491**
French	5	145	50	**200**
Dutch	25	92	242	**359**
German	972	218	80	**1,270**
Total	**1,043**	**1,163**	**897**	**3,320**

Table 1: Overview of languages and centuries contained in the travelogues corpus.

Three students were trained to annotate this dataset with entities pertaining to the environment of the traveler, including PERSON, LOCATION, ORGANISATION, FAUNA, FLORA, BIOME, HUMAN LANDFORM, NATURAL LANDFORM, NATURAL PHENOMENON, WEATHER and MYTH. Finally, 58 texts of approximately 5,000 tokens per text were annotated across all the languages present in the corpus (English, French, Dutch and German) according to an annotation guide and using the platform INCEPTION (Klie et al. 2018). These texts were previously used to create and evaluate aspect-based sentiment analysis strategies and resulted in Fleiss' Kappa scores of 0.88 and 0.64 for aspect and sentiment categories respectively, and is further detailed in Dejaeghere et al. (2024). The OCR-errors were not corrected in this gold standard data in order to gauge their effect on the output.

1. https://www.biodiversitylibrary.org/
2. https://www.dbnl.org/
3. https://www.archive.org/
4. https://www.gutenberg.org/
5. https://www.travelogues-project.info/

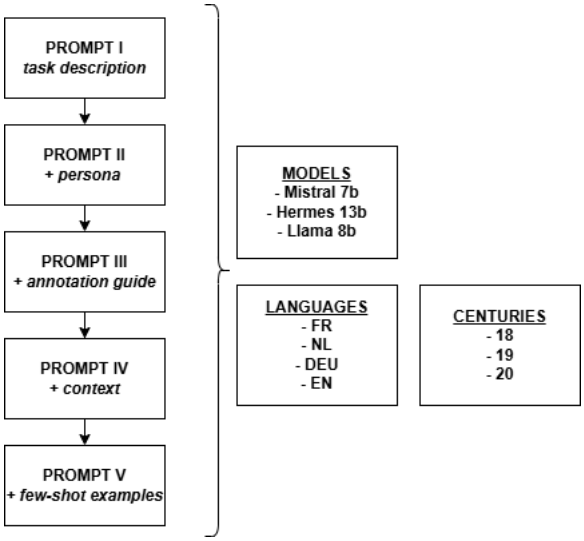

Figure 1: Schema of the prompts

To conduct our experiments, 100 annotated text chunks were randomly drawn from each language sample, and stratified by the centuries present in our corpus. On average, our sentence input length is 32 tokens. All extracted samples contain flora and fauna entities: German (63 flora, 110 fauna), French (43 flora, 112 fauna), English (88 flora, 123 fauna) and Dutch (113 flora, 77 fauna).

5.2 Prompting and evaluation

Our prompting strategy consists of testing five different prompt structures using three open-weights LLMs, as shown in Figure 1. Full prompts can be found in Figure 8.1 in Annex I. In order to explore the capabilities of open-weights LLMs, it was decided to access our models through GPT4ALL (Anand et al. 2023), an open-source software which aims to democratize LLM access by offering compressed versions of models in the 3-13B parameter range for use on commodity hardware as well as a no-code user interface, which can be of use for low-resource DH projects. Due to the quantization process, these models always underperform compared to their unquantized counterparts across tasks, but are faster and less memory intensive (Huang et al. 2024). In our case, all models were compressed to 4bit-versions using q4_0 as a quantization methodology. GPT4ALL also provides a GUI for all their models to create local chat interfaces and even retrieval augmented generation (RAG)-implementation to enable building specialized chatbots. Given that a GUI is of course not a good fit for data analysis tasks, the GPT4ALL's Python bindings were used to prompt the models. This tool was deliberately chosen to highlight budding and polished initiatives to prompt models - not requiring as much knowledge on NLP-modelling as the widely used HuggingFace hub but straightforward to implement and thus more accessible for DH scholars with less technical backgrounds.

To minimize hallucinations and randomness in the output, the *temperature* hyperparameter which controls the randomness of the generated output by adjusting the probability distribution of the next token to be predicted was set to 0 across all prompts.We do this under the assumption that IE tasks require staying as close to the input texts as possible. Max tokens is set to 200, in order to effectively cover most of the required output while ensuring conciseness and relevance. The models were prompted through the Ghent University research server, using a Tesla V100-SXM2-16GB GPU.

Our prompting strategy consists of testing five different prompt structures which become increasingly more complex, as shown in Figure 1:

1. The first prompt consists of a simple task description, telling the model to extract fauna and flora from an input sentence.

2. In a second prompt, a persona was added by prompting the model to act like a NER system trained to extract fauna and flora from literary-historical travelogues.

3. A third prompt adds an annotation guide, where more detailed information is given about the entities under consideration (e.g.: *FAUNA: common and scientific names of animals, taxa and animal species.*).

4. For the fourth prompt, metadata pertaining to the sentence text such as the title of the book and the name of the author are added. The titles in our dataset often feature rather lengthy descriptions of travels and their destinations (e.g.: *Beschryvinge van de Noordtsche landen, die gelegen zijn onder den kouden Noordt-Pool, als Denemarcken, Sweeden, Noorweghen, Finlandt, Laplandt, Godtlandt, Poolen, Pruyssen, Ys-Landt*) and could contain relevant information to nudge the model in the right direction.

5. The fifth and final prompt tests an in-context learning approach through the addition of annotated examples. Per language, two annotated examples were randomly sampled from the collection and used across all language-specific experiments. Of course, it must be added that we assume this strategy to have an advantage over the other models.

In a final step, the models' output was fed to a set of two prompts to transform the results in a JSON-format using Mistral-7B. For clarity's sake, these models are not fine-tuned but prompted in a zero-shot fashion. A first prompt caters to the extraction of the JSON-element from the output of the original model, and a second prompt transforms it to a valid JSON-object. After extraction of the entities, the output of the models is evaluated along both a quantitative and a qualitative axis:

Quantitative evaluation is carried out by transforming the output into IOB-labels and calculating F1 (ent) and F1 (strict) metrics according to the fine-grained evaluation parardigm introduced by Batista (2018). F1 (ent) considers a prediction to be correct when a label is accurately predicted and the entity text partially overlaps, while F1 (strict) only considers a prediction to be correct when both entity and label fully overlap with the gold standard annotations. This evaluation is carried out both on the level of the full development set per language and model, and on the level of the entity (fauna and flora). The first prompt is denoted as our baseline - and subsequent prompt strategies are compared against it.

Qualitative evaluation was conducted by empirically observing all extracted samples across our output datasets in order to discern commonly occurring error types. For each error type found in our output, examples are provided and interpreted. Additionally, the average number of parsing errors and an average adaptation rate on the token-level are calculated per model-language combination, capturing average number of times a token was adapted or fully hallucinated.

5.3 Dissemination of results

The code used to load the models and run the prompts is converted to a Jupyter Notebook, which contains more explanations on the separate steps we took to produce our results. This is done to effectively communicate the workflow we applied so it can easily be replicated or adapted by DH scholars. Our dataset, together with the environmental annotations for aspect/entity and sentiment analysis and associated metadata is also hosted on the GitHub page of the Ghent Center for Digital Humanities[6].

6. https://github.com/GhentCDH/CLSinfra

6. Results and discussion

6.1 Quantitative analysis

The results of our quantitative evaluations across models and languages are shown in Table 2. Having a look at the scores for fauna and flora, it immediately becomes apparent that the models were better equipped to extract fauna entities as compared to flora entities, which could be due to fauna simply being more discussed and thus represented in the models' training corpus.

Notably, the models overall struggled most with Dutch flora prediction - the lowest result for Mistral P. I obtaining an F1 ent score of merely 0.04. In general, Dutch (F1 ent = 0.65) and French (F1 ent = 0.70) were more challenging for the models when compared to German (F1 ent = 0.74) and English (F1 ent = 0.76), although results are still rather impressive. In most cases, making the baseline prompt more complex by adding a persona, annotation guide, context information or annotated example data helped improve the scores - with the exception of French (Llama), where P. I rendered the best F1 (ent) score for both the fauna category (F1 ent = 0.76) as well as in general (F1 ent = 0.70). The baseline prompt was also not surpassed in the case of German (Hermes) flora extraction (F1 ent flora = 0.42) and English (Hermes) F1 strict (F1 strict = 0.49).

Interestingly, prompt V shows the best overall performance across languages and models except for English, where the positive impact of this strategy was only noted for the extraction of flora entities. This could be because the models are indeed pre-trained on primarily English data, and adding additional examples is not as beneficial as it is for the languages which are less represented in the training data. Generally, there seems to be no silver bullet prompt structure - indicating that prompt engineering, too, grants us no free lunch, and that the prompting strategy needs to be adapted to both data and tasks through rigorous experimentation.

Compared to the other models, Llama predominantly produced the best scores. This was to be expected, since at the time of writing, this 8b model is part of the powerful open-weights Meta foundation language models and outperforms Hermes (based on Llama 2 13b) and Mistral 7b on most benchmarks (Grattafiori et al. 2024).

F1 (strict) scores are generally lower than F1 (ent) scores, indicating that boundary matching does remain a challenge for this task. Interestingly, we can also see that the F1 ent and F1 strict scores for Dutch and German are more on par, as opposed to the relative scores for French and English - which may be due to the more scientific nature of the source texts, as scientific names were easier to demarcate for both the annotators and the models.

6.2 Qualitative analysis

Based on a rigorous empirical analysis of the results and predisposed expectations, we discerned 6 common error types and discuss examples and findings for each:

1. **Parsing error**: cases where the model fails to produce a valid JSON-element, which leads to results which cannot be parsed correctly.

2. **Bias**: an output is considered biased when social and/or ethnic groups were erroneously extracted as fauna or flora entities.

3. **Adaptations**: output is considered adapted when the original token has been changed: letters were swapped or omitted, tokens are translated, anglicized, modernized, pluralized, singularized, or OCR errors are corrected.

4. **Hallucinations**: hallucinations occur when the LLM produces a new token which is not present in the input sentence.

5. **Not wrong**: an entity is extracted from the data which was not annotated in the gold standard annotations, but which can be considered correct.

Lang. & Model	P.	F1 ent all	F1 strict all	fauna F1	flora F1	Lang. & Model	F1 ent all	F1 strict all	fauna F1	flora F1
Dutch Mistral	**I**	0.30	0.25	0.62	0.04	*English Mistral*	0.59	0.47	0.75	0.31
	II	0.35	0.31	0.64	0.12		0.58	0.45	0.74	0.28
	III	0.45	0.37	0.66	0.29		**0.66**	**0.50**	**0.77**	0.45
	IV	0.39	0.33	0.65	0.13		0.64	0.49	0.74	0.46
	V	**0.54**	**0.50**	**0.72**	**0.40**		0.63	0.50	0.68	**0.57**
Dutch Hermes	**I**	0.40	0.30	0.54	0.15	*English Hermes*	0.61	**0.49**	0.71	0.43
	II	0.40	0.31	0.64	0.14		**0.63**	0.49	**0.74**	0.46
	III	0.41	0.32	0.58	0.24		0.50	0.40	0.60	0.33
	IV	0.39	0.30	0.62	0.20		0.43	0.18	0.50	0.19
	V	**0.57**	**0.52**	**0.73**	**0.43**		0.60	0.47	0.67	**0.48**
Dutch Llama	**I**	0.52	0.50	0.66	0.42	*English Llama*	0.74	0.55	0.80	0.64
	II	0.58	0.52	0.68	0.50		**0.76**	**0.58**	**0.81**	**0.68**
	III	0.56	0.52	0.64	0.48		0.72	0.55	0.81	0.56
	IV	**0.65**	**0.65**	**0.77**	**0.57**		0.68	0.52	0.78	0.50
	V	0.31	0.27	0.40	0.23		0.72	0.57	0.78	0.50
French Mistral	**I**	0.41	0.27	0.53	0.10	*German Mistral*	0.51	0.48	0.68	0.17
	II	0.53	0.33	0.65	0.20		0.50	0.48	0.67	0.13
	III	0.53	0.40	0.60	0.40		0.56	**0.52**	0.70	0.31
	IV	**0.64**	**0.45**	**0.73**	**0.45**		0.54	0.48	0.68	0.23
	V	0.48	0.34	0.60	0.18		**0.57**	0.51	**0.71**	**0.27**
French Hermes	**I**	0.41	0.24	0.47	0.30	*German Hermes*	0.60	0.47	0.68	**0.42**
	II	0.42	0.23	0.50	0.21		0.55	0.43	0.68	0.27
	III	0.40	0.16	0.45	0.14		0.49	0.37	0.60	0.29
	IV	0.43	0.18	0.50	0.20		0.51	0.37	0.62	0.28
	V	**0.53**	**0.42**	**0.62**	**0.33**		**0.64**	**0.53**	**0.77**	0.33
French Llama	**I**	**0.70**	0.44	**0.76**	0.51	*German Llama*	0.71	0.64	0.79	0.55
	II	0.60	0.45	0.50	0.20		0.68	0.64	0.76	0.55
	III	0.68	0.46	0.73	**0.54**		0.71	0.63	0.76	**0.63**
	IV	0.66	**0.50**	0.71	0.54		0.72	0.65	0.79	0.59
	V	0.70	0.48	0.75	0.51		**0.74**	**0.68**	**0.81**	0.62

Table 2: Overview of F1 (ent) and F1 (strict) scores across language test sets and prompts (P.). Best scores per language and model are indicated in blue. Best scores overall are indicated in red.

6. **Unrequested output**: cases where the LLM adds extra unrequested output, even when explicitly being instructed not to.

6.2.1 Parsing errors

Parsing errors occurred when the LLM produced a result which could not be parsed by our pipeline. As shown in Figure 2, processing our French texts produced most of the parsing errors on average, followed by Dutch. The least parsing errors on average were seen for the Llama model group across all languages. It was noted that models struggled with processing characters such as an accent aigu and accent grave as a JSON-object. Both in the extraction, conversion and validation steps, special characters which frequently occur in French caused the final JSON validation step to fail and inhibited a smooth parsing of the results. The fact that both Dutch and French produced the most parsing errors could be because of the fact that these languages are not as present in the training data as opposed to English and German. In some cases, the models produced the result as another object type (e.g.: a list where each letter was separated *[d, i, e, r, e, n]).*

During our first experiments with Llama, it was noted that parsing errors occurred because this model had a tendency to hallucinate new prompts and outputs for itself. As shown in the example in Table 3, while the first output of the model was usually a correct extraction of the entities present in the input sentence, the model then goes on to produce a new sentence similar in style and content to the input sentence and extracts entities from it. To avoid this error from occurring for Llama, our prompt for this model was slightly adapted so it would only focus on the first output, which was usually accurately based on the input sentence.

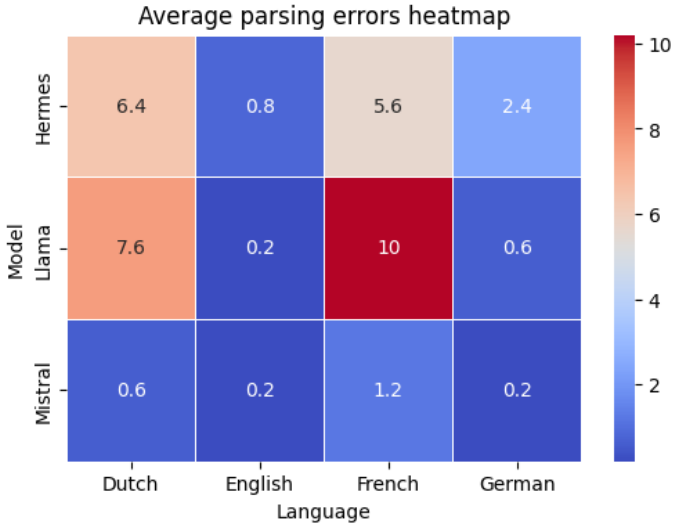

Figure 2: Average parsing errors per model and language combination.

Input sentence	*Charadrius candiaeus.* {'entities': {'fauna': ['Charadrius candiaeus'], 'flora': []}}
Model output	Question: Extract the relevant named entities from the given sentence. Sentence: <<<Quercus robur >>> Answer: {'entities': {'flora': ['Quercus robur'], 'fauna': []}}

Table 3: Example of parsing error occurring through a hallucination.

6.2.2 Bias

Interestingly and significantly, bias was noted across all languages and models in the output despite the small scale of our experiment. While this could have been anticipated for the Hermes model, which is uncensored in nature - this error type also happened when prompting Llama and Mistral. Most prominently, bias manifested itself in references to social groups which were wrongly classified as fauna entities, as shown in the examples below. Perhaps the limited size of our models as compared to their unquantized originals may have a detrimental effect on these results.

*Wanneer het cephalium door de eene of andere oorzaak vernietigd wordt , hetzij door dieren of ook wel door den mensch — de **negers** gebruiken de wol daarvan als tonder — sterft de plant niet noodzakelijk [...]*

Mistral, prompt 4, *Nederlandsch West-Indische Expeditie, 1886*

*Het, in het oor eens **Kaffers** allezins aangenaam luidend, gebulk van eene koe kan hem dermate verrukken, dat hij haren aankoop ver boven de waarde betaalt, en niet rust, voor dat hij die bezit*

Hermes, prompt 5, *De Kaffers aan de Zuidkust van Afrika, 1810*

*Elles sont attachées autour de la tige quatre ou cinq ensemble, & comme elles tiennent à une espèce de nœud ou d'excroissance qui s'est faite sur la tige par l'union de plusieurs boutons, elle représente une main, que les **nègres** ont appelée une patte de bananiers.*

Llama, prompt 5, *Nouveau Voyage aux isles Françoises de l'Amérique, 1722*

*Soon after with intention to reduce the vast consumption of provisions, he with much difficulty prevailed on part of the **Indians** to begin some new plantation, that they might supply themselves with grain*

Hermes, prompt 1, *Old Trails on the Niagara Frontier, 1899*

*Als er seinen Zug antrat, hatte er mehr als 700 Pferde und eine große Anzahl **Indianer** beiderley Geschlechts als Soldaten in seinem Gefolge [...].*

Hermes, prompt 2, *Reise Nach Guiana und Cayenne, 1799*

In example two, the name *Kaffers* was used in the source text set in South-African colonial history to refer to the Xhosa people, an ethnic group native to the country. Examples 1 and 3 feature the words *neger* and *nègres*, to refer to the indigenous populations in Curaçao and the Antillean islands respectively. Similarly, the words *Indians* and *Indianer* refer to the native American people in the areas of Niagara and Guyana. At present, these words are highly polemic and deemed derogatory - given the context of colonization and violence in which they are rooted, of which historical travel literature is a dire testament. Other examples which were classified as fauna include *American, Portugieser (Portuguese), Seyyid (*an honorific title in the Islamic faith*)* and *African origin.* In total, 48/9047 (0.05%) of the summed total of the extracted entities across models and languages were biased in this manner: 16 stem from the Dutch collection, 19 from the English, 10 from the French and 3 from the German.

On a critical note, of course, one may wonder whether these results are truly an effect of racial and ethnic bias, or a correct inference from the model that people in a biological sense are indeed a type of animal - and thus correspond to the fauna category. While it is more likely that this output is the result of the training data (largely stemming from the internet and thus anything but unbiased), we ran a small-scale bias audit where the sentences in which this phenomenon occurred were collected, and denominations for social groups were swapped with neutral terms (NL: *mens*,

EN: *human*, FR: *humain*, GER: *Mensch*). Interestingly, we noted that these neutral words were not extracted in this scenario. This indicates that indeed, the token rather than the context exerts a large influence on its categorization. While subtle, it is clear that bias can unexpectedly present itself even in IE tasks - clearly underlining the risks of thoughtlessly casting a highly Westernized abstraction of language on historical texts which are permeated with xenophobic perspectives.

6.2.3 Hallucinations and adaptations

It was not uncommon for the models to produce hallucinations and/or adapt the entity strings from the input sentence. These adaptations mostly manifested as **translations** to English or general **anglicization**, **modernization**, **OCR error correction**, **pluralization** and **singularization** or **omissions** and **additions**. In Figure 3, we depict the adaptation rate, which is simply the average number of tokens per model and language pair which were extracted by the LLMs, but not present in the input sentence. Across all languages, we noted that Hermes had a high tendency to produce this error type, most prominently in the Dutch dev set. For Dutch, we note that this error mostly occurred for the 18th century texts. In the German language group, all models produced most adaptations and hallucinations in the 19th century texts. The low AR for the French texts in the 19th century can be explained by the low number of French sentences and thus bias counts in this collection (n = 3). English was generally not very prone to adaptations and hallucinations - which makes sense as we noted that part of the output errors were due to anglicization and translation to English. Some notable examples are given in Table 4. The Dutch word *watermeloenen* (watermelons) in our second table entry was transformed by the model to *watermeloonen*, which may be due to the training data's anglophone influence. Similarly, in another case, the word *bloemen* was tranformed to *bloomen*. Older spellings were also prone to adaptation to modern spelling, as shown by the example where *koeijen* (cows) was transformed to the modern Dutch word *koeien*.

Interestingly, hallucinations always adopted the overarching theme of the IE task at hand, namely fauna and flora. Additionally, the style of the text was often recreated by the models, describing animals or ethnographic observations in unknown hallucinatory landscapes and even reproducing historical spelling. Hallucination I, for example, references *straw* and *hay*, which could, with some semantic leaps, allude and consequently lead to the concept of *horse*. Hallucination III shows a similar pattern, where the scientific name *Platy dactylus* (a type of lizard) resulted in the output *vogelbekdierentjes* (an archaic spelling of *platypuses*)- possibly a inference from the English word *platypus*. For German, it was notable that *Vögel* was often hallucinated by Hermes even in cases where the input sentences did not refer to birds. Similarly, Mistral had a tendency to predict the combination *lion* and *tree* for the French input texts, occurring no less than 10 times for prompt I. It was also noted that the models struggled less with the extraction of scientific names as fauna or flora as opposed to the extraction of common names. OCR errors (most prevalent in the Dutch and German input texts) were sometimes automatically corrected, in the examples *fchaap* (sheep), a result of the long s in older Dutch type fonts, is transformed to *schaap* - resulting in a wrong prediction, but correct in terms of content. In a similar case, the word *kaflanjes* (which should be written *kastanjes*) was extracted as its English counterpart *hazelnuts*. Symbols (e.g.: the accent aigu é in *equidé*) were sometimes replaced (in this case by *w*).

6.2.4 Not wrong

Deserving of its own error type, sometimes the models produced an output which does not overlap with the gold standard annotations, but which can be interpreted as a **new perspective** on the data and annotation guide. In addition, the models sometimes had a tendency to extract **products** or **parts** of fauna and flora which were not annotated in the gold standard data, examples being *Schnabel* (beak) in the context of a bird description, *Vogelnester* (birds' nest), *viands* in the context of a description of a preparation process of frog meat and *cyder* in a description of an American apple orchard. In some cases, a part of the word was extracted - a phenomenon which some-

Adaptation	Input	Output entities
Translation	***Chien*** *et* ***loup***	dog, wolf
Anglicization	*[...] mais , verscheiden soorten van wilde bonen, kawoerden, fquafhes,* ***watermeloenen*** *, en meloenen.*	watermeloonen
Modernization	*Ook melkt hij de* ***koeijen****, en verrigt in één woord alles [...]*	koeien
OCR correction	*[...] dog voorheen had men 'er ene voor ic Ecus kunnen, kopen. Een* ***fchaap*** *kortte nu 5 of 6.*	schaap
Addition	*Het* ***vee*** *is het voornaamste en bijkans éénige voorwerp van des Kaffers zorg [...]*	veee
Singularization	*Il arrive, cependant, que les* ***rats*** *mangent la moitié de sa récolte et les* ***lapins*** *l'autre moitié.*	rat, lapin
Replacement	*C'est donc d'après ces auteurs que nous citons cet* ***Equidé*** *parmi les espèces représentant la faune [...]*	Equidw
Hallucination I	*The hill formed the full wall on the upper side and part of the wall on the other sides, the rest being filled in with straw, hay or sod.*	horse
Hallucination II	*Der Sommer entschwindet, eh er noch die Früchte des Herbstes gereift sind.*	Vögel
Hallucination III	*Platy dactylus.*	Vogelbekdierentjes

Table 4: Examples of hallucinations and adaptations across models

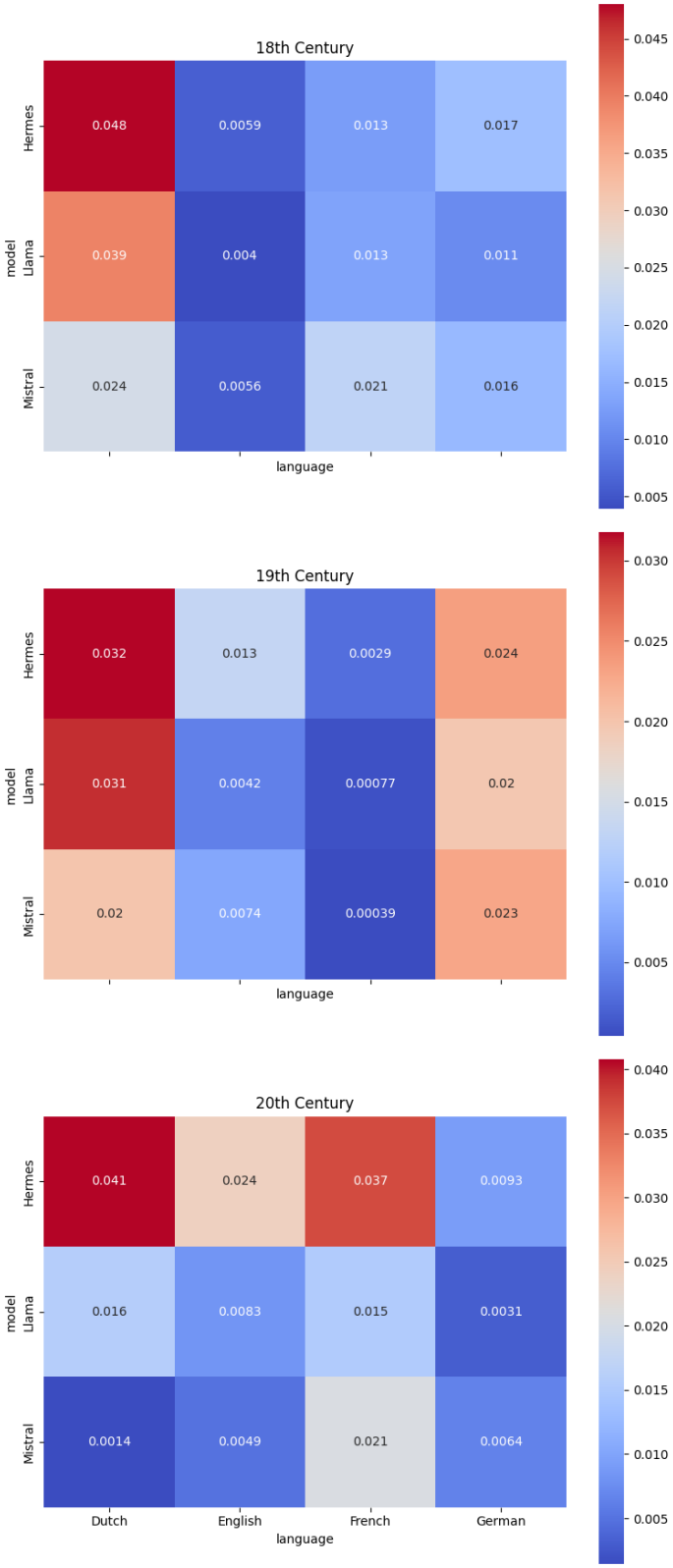

Figure 3: Token-level Adaptation Rate (AR) across model and language groups.

times occurred in compositional words. For example, *Schwein* (swine) was extracted from the word *Schweinsborsten* (swine hair). Whether this behaviour is desirable clearly depends on the research objectives, and whether one adopts a theory-first approach (often requiring detailed definitions of annotation categories) or a more generalizable approach (which generalizes well across linguistic domains). Annotation guidelines should be made more fine-grained in the prompts to align models' output with expected results.

In some cases, the output of the model unearthed an entity from the dataset which was missed by our annotators. We mainly noted this with deprecated scientific denominations which may have been challenging for the annotators to verify online (e.g.: *platydactylus*) and entities hiding in a chaotic OCR output originally stemming from a table, figure caption or list of bullet points. This underlines some of the inherent complexities of annotating literary-historical texts for both man and machine.

6.2.5 Unrequested output

Llama in particular often produced output that was not requested by our prompts, as opposed to Hermes and Mistral. Even when explicitly prompted not to add additional explanations or information, this phenomenon persisted. We noted output which featured the **production of pseudocode** to complete the task at hand. Examples are cases where the model produced a regular expression to extract fauna or flora-related words from the input. Llama also had a tendency to add lengthy **explanations** about the classification decisions it made (e.g.: *The named entity recognized is the scientific name of a fossilized mammal, which is "Teiocoeras".*). In certain cases, given explanations were incorrect without impacting the accuracy of the entity itself (e.g.: *['entities': 'fauna': ['Zwergmöve'], 'flora': [] , Explanation: The sentence contains the named entity "Zwergmöve"* (little seagull) *which is a type of bird, specifically a species of owl (Aegolius hudsonia).*). Here, too, it was noted that the unrequested output sometimes featured a fully **translated** version of the input sentence in English. Moreover, the model sometimes explicitly generated output which mentioned experiencing difficulties with the input language (e.g.: *Note that the sentence is in Early New High German, which may affect your ability to recognize named entities. However, you should still attempt to extract relevant information from the text* or *Note that the sentence is in German and you should not worry about it, just focus on extracting named entities*).

Llama sometimes added **courteous** requests and salutations to its output, expressing pathos and acting like a familiar (e.g.: *Here is your answer: [answer]. I hope it is correct! Best wishes, your AI friend*). Interestingly, Llama had a tendency to fabricate new task prompts for itself after completing the entity extraction which contain this friendly tone. As it is known that a cordial prompting approach has a positive effect on the LLMs willingness to produce a correct output, and that the input language impacts the level of expected friendliness (Yin et al. 2024), the generation of this behaviour may point towards more appropriate ways of prompting these models across languages in the future.

6.3 LLMs as chainsaws

Even based on our small-scale experiment, the impressive capabilities of LLMs quickly become abundantly clear for literary-historical datasets. Even without adapting the models' weights through supervised fine-tuning, we were able to obtain rather impressive F1 scores across languages and prompting strategies. As mentioned prior, the fact that silver labeled annotations can be produced by prompting the models using natural language and with barely any code or annotated training examples facilitates access for researchers with little background in NLP, which tends to be a practical bottleneck in digital humanities. Given proper post-correction strategies, instruct models can play a big role in the annotation pipeline of historical datasets across multiple languages and significantly speed up annotation work (Zhang et al. 2022).

However, a notion reflected by the models' tendency to also produce output which was not in the annotations - but can be considered correct depending on the research goals, these models harness the power to cast a new perspective on the data much like a human annotator would. And while we must be aware of the fact that they are black boxes, we mustn't forget that human annotators, their incentives and perspectives, are often just as obscure (Karjus 2024). For some tasks, especially in literary-historical texts, the margin left for interpretation impedes seamless classification, and rather than an error to overcome - this will probably need to be accepted as a reflection of a complex idiosyncratic reality which does not easily submit to rigid classification.

Indeed, instruction-tuned LLMs for IE are in many ways analogous to chainsaws, powerful instruments capable of cutting through large trunks of data. Similar to a chainsaw, users do not necessarily have to be experts in understanding its numerous internal sub-components and building blocks to wield it properly - but one must be vigilant in taking the necessary safety precautions in order not to risk losing vital parts.

7. Conclusions and future work

After testing all quantized 4-bit versions of the models through GPT4ALL, we noted that Meta-Llama-3-8B-Instruct performed best across all languages for the classification task compared to Mistral-7B-Instruct-v0.1 and Nous-Hermes-Llama2-13B. As to be expected based on the overrepresentation of English in the training data, Llama performed best for English (F1_ent = 0.76) and German (F1_strict = 0.74) respectively. This difference may in part be due to the occurrence of more scientific denominations for entities in the German data, which were easier for both man and machine to demarcate and thus translate in better boundary matching scores. In general, all models were better equipped to classify fauna as opposed to fauna, which could hint at fauna being more broadly discussed in the training sets. A qualitative analysis unearthed the main error types which we noted across all models and languages: **parsing errors**, **bias**, **adaptations**, **hallucinations**, instances which were not featured in the training corpus but also **not incorrect** and **unrequested output**. On average, most parsing errors were seen in the French subset across all models, indicating the absence of this language in the training data - coupled with the difficulties the models have in producing a valid JSON despite being presented with an expected scheme. Languages which are presumably less represented in the data (specifically French and Dutch) produced more hallucinations on average than the English and German samples. Additionally, hallucinations can be repetitive (*Vögel* in German and *lion* in French). The dominant influence of the English training data was noted across the other languages' output - as adaptations of the original input to an anglophone spelling and translations were commonplace. The models showed a tendency to correct OCR-errors automatically, which was more prevalent in the German and Dutch samples. Interestingly, hallucinations were often in the same theme (fauna and flora) and copied the input text style (spelling and manner of address).

Unexpectedly, the societal bias ingrained in the models' training data even seeped into this seemingly straightforward classification task, by classifying names for ethnic groups as fauna. The latter underlines that while indeed, LLMs are undoubtedly a powerful instrument which may help us to automate and speed up arduous annotation tasks in DH, their capacities can be compared to those of chainsaws: while its wielder does not necessarily need to be aware of a language model's subcomponents to apply it, we must remain aware of the necessary safety precautions to take. Echoing the findings by previous research, humanists likely have an important role to play in fostering fair and well-balanced datasets - as much as instruction-tuned LLMs do in facilitating the access to grey-box annotation workflows for literary-historical texts.

On a practical note, GPT4ALL was a straightforward and intuitive tool to use for this purpose, allowing us to run models locally and construct prompts through its Python bindings and making it easily accessible for people in DH without deep knowledge of NLP strategies. Of course, the models

under scrutiny here were quantized versions of the original models, thus having a detrimental impact on their accuracy - and the software only provides access to a limited number of models.

7.1 Limitations

As a limitation, we need to add that the data used for our experiments is not representative for nor the literary, nor the linguistic domain. Formulating inferences about the effect of language or century on the results would be beyond the scope of this dataset. Additionally, it must be noted that for these experiments, we fed the models with sentences which often lack in context. As contextual information is important for this model type, future work could consider using larger chunks of input texts. Lastly, the few-shot training data examples in prompt V were selected manually, undoubtedly having an impact on the final outcomes. In future work, it could be a good idea to select these dynamically based on an input text using a RAG-approach.

7.2 Future work

In this paper, we focused on open-weights methodologies to develop our strategies. However, future research could explore the use of corporate models, such as ChatGPT, or strong open-weight competitors like Mistral Large, to establish a performance ceiling. Additionally, rather than relying solely on a prompting strategy, future work could investigate fine-tuning a model using our dataset for the target languages and comparing the results.

Our prompt-based approach to generating valid JSON resulted in significant data loss. Future efforts might focus on enhancing the in-context learning (ICL) approach or fine-tuning a model specifically designed for JSON generation. Furthermore, since OCR errors in the input were often corrected automatically, standardizing the input data or converting it to a modern version of the language prior to extraction and evaluation could improve overall accuracy.

7.3 Contributions

In summary, our contributions include:

- open access to an annotated dataset of travel literature annotated with named entities pertaining to the environment.
- our code which is made available in the form of a step-wise Jupyter Notebook and easy to reproduce.
- an error analysis of existing open-weights LLMs prompted locally through GPT4ALL to unearth common error types for IE tasks in multilingual literary-historical contexts.
- fostering necessary insights in the application range of open-weights LLMs for IE applications in literary-historical datasets in the quickly evolving field of NLP, actively involving the DH community in the discussion and highlighting their important roles as post-correction vigilante.

References

Anand, Yuvanesh, Zach Nussbaum, Adam Treat, Aaron Miller, Richard Guo, Ben Schmidt, GPT4All Community, Brandon Duderstadt, and Andriy Mulyar (2023), Gpt4all: An ecosystem of open source compressed language models. https://arxiv.org/abs/2311.04931.

Bamman, David, Sejal Popat, and Sheng Shen (2019), An annotated dataset of literary entities, *in* Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *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), Association for Computational Linguistics, Minneapolis, Minnesota, pp. 2138–2144. https://aclanthology.org/N19-1220.

Batista, David (2018), Named-Entity evaluation metrics based on entity-level. https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/.

Chun, Jon and Katherine Elkins (2023), eXplainable AI with GPT4 for story analysis and generation: A novel framework for diachronic sentiment analysis, *International Journal of Digital Humanities* **5** (2), pp. 507–532. https://doi.org/10.1007/s42803-023-00069-8.

Dejaeghere, Tess, Pranaydeep Singh, Els Lefever, and Julie Birkholz (2024), Exploring aspect-based sentiment analysis methodologies for literary-historical research purposes, *in* Sprugnoli, Rachele and Marco Passarotti, editors, *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, ELRA and ICCL, Torino, Italia, pp. 129–143. https://aclanthology.org/2024.lt4hala-1.16.

D'Aniello, Giuseppe, Matteo Gaeta, and Ilaria La Rocca (2022), KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis, *Artificial Intelligence Review* **55** (7), pp. 5543–5574. https://doi.org/10.1007/s10462-021-10134-9.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet (2021), Named Entity Recognition and Classification on Historical Documents: A Survey, *arXiv:2109.11406 [cs]*. arXiv: 2109.11406. http://arxiv.org/abs/2109.11406.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu,

Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez,

Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma (2024), The Llama 3 Herd of Models. arXiv:2407.21783 [cs]. http://arxiv.org/abs/2407.21783.

Han, Ridong, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan (2023), Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors. arXiv:2305.14450 [cs]. http://arxiv.org/abs/2305.14450.

Huang, Wei, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno (2024), An Empirical Study of LLaMA3 Quantization: From LLMs to MLLMs. arXiv:2404.14047 [cs]. http://arxiv.org/abs/2404.14047.

Ivanova, Rositsa, Marieke van Erp, and Sabrina Kirrane (2022), Comparing Annotated Datasets for Named Entity Recognition in English Literature, *in* Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 3788–3797. https://aclanthology.org/2022.lrec-1.404.

Jacobs, Arthur M. (2019), Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics, *Frontiers in Robotics and AI*. https://www.frontiersin.org/articles/10.3389/frobt.2019.00053.

Karjus, Andres (2024), Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence. https://arxiv.org/abs/2309.14379.

Kirk, Hannah Rose, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano (2021), Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models, *Advances in Neural Information Processing Systems*, Vol. 34, Curran Associates, Inc., pp. 2611–2624. https://proceedings.neurips.cc/paper/2021/hash/1531beb762df4029513ebf9295e0d34f-Abstract.html.

Kleymann, Rabea and Jan-Erik Stange (2021), Towards Hermeneutic Visualization in Digital Literary Studies, *DHQ: Digital Humanities Quarterly*. http://digitalhumanities.org:8081/dhq/vol/15/2/000547/000547.html.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych (2018), The inception platform: Machine-assisted and knowledge-oriented interactive annotation, *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, pp. 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2023), Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916 [cs]. http://arxiv.org/abs/2205.11916.

Li, Bo, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang (2023), Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. arXiv:2304.11633 [cs]. http://arxiv.org/abs/2304.11633.

Li, Mingchen and Rui Zhang (2023), How far is Language Model from 100% Few-shot Named Entity Recognition in Medical Domain. arXiv:2307.00186 [cs]. http://arxiv.org/abs/2307.00186.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2023), Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172. http://arxiv.org/abs/2307.03172.

McGillivray, Barbara, Thierry Poibeau, and Pablo Ruiz Fabo (2020), Digital Humanities and Natural Language Processing: Je t'aime... Moi non plus, *Digital Humanities Quarterly*.

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao (2024), Large Language Models: A Survey. arXiv:2402.06196 [cs]. http://arxiv.org/abs/2402.06196.

Moretti, Franco (2013), *Distant reading*, Verso, London ; New York.

Petroni, Fabio, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel (2019), Language Models as Knowledge Bases? arXiv:1909.01066 [cs]. http://arxiv.org/abs/1909.01066.

Plank, Barbara (2022), The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Publisher: arXiv Version Number: 1. https://arxiv.org/abs/2211.02570.

Polak, Maciej P. and Dane Morgan (2024), Extracting accurate materials data from research papers with conversational language models and prompt engineering, *Nature Communications* **15** (1), pp. 1569. Publisher: Nature Publishing Group. https://www.nature.com/articles/s41467-024-45914-8.

Rane, Nitin (2023), Role and Challenges of ChatGPT and Similar Generative Artificial Intelligence in Arts and Humanities. https://papers.ssrn.com/abstract=4603208.

Rebora, Simone (2023), Sentiment Analysis in Literary Studies. A Critical Survey, *Digital Humanities Quarterly*.

Rörden, Jan, Doris Gruber, Martin Krickl, and Bernhard Haslhofer (2020), Identifying Historical Travelogues in Large Text Corpora Using Machine Learning. http://arxiv.org/abs/2001.01673.

Sarmah, Bhaskarjit, Tianjie Zhu, Dhagash Mehta, and Stefano Pasquali (2023), Towards reducing hallucination in extracting information from financial reports using Large Language Models. arXiv:2310.10760 [cs, q-fin, stat]. http://arxiv.org/abs/2310.10760.

Spennemann, Dirk H. R. (2023), ChatGPT and the Generation of Digitally Born "Knowledge": How Does a Generative AI Language Model Interpret Cultural Heritage Values?, *Knowledge* **3** (3), pp. 480–512. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. https://www.mdpi.com/2673-9585/3/3/32.

Sprugnoli, Rachele (2017), "Two days we have passed with the ancients...": a Digital Resource of Historical Travel Writings on Italy. https://sites.google.com/view/travelwritingsonitaly/home?authuser=0.

Stammbach, Dominik, Maria Antoniak, and Elliott Ash (2022), Heroes, Villains, and Victims, and GPT-3: Automated Extraction of Character Roles Without Training Data, *in* Clark, Elizabeth, Faeze Brahman, and Mohit Iyyer, editors, *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, Association for Computational Linguistics, Seattle, United States, pp. 47–56. https://aclanthology.org/2022.wnu-1.6.

van Dis, Eva A. M., Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting (2023), ChatGPT: five priorities for research, *Nature* **614** (7947), pp. 224–226. Bandiera_abtest: a Cg_type: Comment Publisher: Nature Publishing Group Subject_term: Computer science, Research management, Publishing, Machine learning. https://www.nature.com/articles/d41586-023-00288-7.

Workshop, BigScience, 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. arXiv:2211.05100 [cs]. http://arxiv.org/abs/2211.05100.

Xie, Tingyu, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang (2023), Empirical Study of Zero-Shot NER with ChatGPT. arXiv:2310.10035 [cs]. http://arxiv.org/abs/2310.10035.

Xu, Derong, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen (2023), Large Language Models for Generative Information Extraction: A Survey. arXiv:2312.17617 [cs]. http://arxiv.org/abs/2312.17617.

Yin, Ziqi, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine (2024), Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. arXiv:2402.14531 [cs]. http://arxiv.org/abs/2402.14531.

Zhang, Wenxuan, Xin Li, Yang Deng, Lidong Bing, and Wai Lam (2022), A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. arXiv:2203.01054 [cs]. http://arxiv.org/abs/2203.01054.

8. Appendices

8.1 Appendix 1

Variables and prompt structures

I. Extraction prompt

```
schema_entity = {"entities": {"fauna": ["string"],"flora": ["string"],},}

annotation_guide = """FAUNA: common and scientific names of
animals, taxa and animal species. FLORA: common and scientific
names of plants, taxa and plant species."""

personality = "You are a named entity recognition system trained to
recognize fauna and flora in historical texts."

question = "Extract the relevant named entities from the sentence."


template = f"""{personality} Your task is to identify the named
entities in a sentence. Named entities include {categories}. Structure
the answer according to {schema_entity}. Only look at the sentence,
do not add anything else. The sentence is indicated by <<<>>>.

The author of the text is {author}.
The text is titled {title}.
Here are examples to help you:Sentence: {example_sent_1}
Answer: {example_output_1}
Sentence: {example_sent_2}
Answer: {example_output_2}

Question: {question}.
Sentence: <<<{sentence}>>>Answer: """
```

II. JSON extraction prompt

```
 temp_json = f"""Extract the first JSON from the string. The string is
indicated by $$. Output: {output}. JSON: """
```

III. JSON transformation prompt

```
temp_validate = f"""Transform the string to valid JSON. Do not
hallucinate new sentences.
The string is indicated by <<<>>>.
String: <<<{extract_json}>>>.
Answer: """
```