Generative LLMs and history research

Limitations for languages, periods, and tasks

Fernanda Olival, Helena Freire Cameron, António Branco, and Renata Vieira

1. Introduction

This chapter aims to examine limitations and challenges for the use of generative large language models (GLLMs) in historical research, as encountered by the authors in their particular experience. The discussion is organized into three issues: the status of Portuguese as a pluricentric language exhibiting linguistic variation, the characteristics of the eighteenth-century Portuguese language, which pose constraints to computational processing due to their specificities, and the natural language task of named entities recognition (NER).

GLLMs are artificial neural network models designed to predict the next word in a sequence. Through large-scale exposure to text, they learn probabilistic patterns of language use—such as the likelihood that *"taking into"* is followed by *"account"*—and thus acquire the ability to generate coherent and contextually relevant text. When integrated into conversational systems, or chatbots, GLLMs can answer questions, summarize information, and perform a wide variety of linguistic tasks.

Despite their growing applications across many domains, their use in historical research remains limited, as also discussed by Büttner (2026). History is a discipline fundamentally shaped by discourse and dominantly reliant on textual sources, though other types of material evidence may also be considered. The use of GLLMs as a tool for historical research remains in its early stages. There are bibliography reporting experiments that attempt to recreate the past using GLLMs fine-tuned on the COHA (*Corpus* of Historical American English) or on books published between 1880 and 1914, but with rather unreliable results, as presented by Underwood et al (2025). Also, see Wolf (2026), about the early modern multilingual project VERITRACE. A critical engagement with both practical and conceptual issues is essential to fostering meaningful progress in the field. Both technical and epistemological issues must be carefully considered in order to make improvements.

The discussion that follows is based on our experience with the transcription, normalization, and semantic annotation of an eighteenth-century Portuguese *corpus* – the

Memórias Paroquiais (*Parish Memories*). This source was produced by multiple authors between 1758 and 1761 as part of a large-scale survey conducted in the aftermath of the devastating earthquake of November 1st, 1755. In 1758, the Portuguese Crown distributed a detailed questionnaire to all parish priests in the country, structured around three main thematic areas (the land, rivers, and mountains), seeking information about the territory, its inhabitants, and the consequences of the catastrophe, with the broader goal of compiling a comprehensive *Geographical Dictionary* of Portugal. In the 19th century, the handwritten responses were compiled into 41 volumes, later supplemented by two additional volumes and an index volume. Today, the collection is preserved in the National Archive of Torre do Tombo, in Lisbon. Transcribed and digitized versions of the *Memórias Paroquiais* pertaining to southern Portugal are available through the data aggregator **CIDEHUSdigital**[1], under a Creative Commons Attribution 4.0 license. This collection has been further enriched with semantic annotations identifying Named Entities (NE), providing a valuable resource for computational and historical research alike.

The next sections are structured around three challenges faced by GLLMs, contextualized on the work developed towards the *Parish Memories*. GLLMs, being designed for general-purpose use and based on contemporary texts, are still limited to dealing with specific languages, earlier linguistic stages of languages, and tasks requiring formally strictly structured representations.

2. Challenges due to specific languages: the case of Portuguese

The Portuguese language has undergone a long and complex process of evolution, shaped by multiple cultural and linguistic influences. While Portuguese was still closely related to Galician in the 12th century, it became a separate linguistic system. From the 15th century, the Portuguese maritime expansion brought the language to Africa, America and Asia. Portuguese served as a tool of commerce, diplomacy, and colonial governance in these areas, assimilating lexical and structural elements from various contact languages. The result of this process is a diverse linguistic legacy (see Mateus and Bacelar do Nascimento, 2005).

Today, Portuguese is spoken by approximately 260 million people (Instituto Camões, 2022), making it the fourth most spoken native language worldwide, after Mandarin, English, and Spanish. It is the official or national language of countries across several continents, including Europe, South America, Africa, and Asia. It is a "pluricentric language", comprising multiple standardized varieties that coexist within a single linguistic system. Two main patterns – European Portuguese and Brazilian Portuguese – dominate globally, while additional national standards are emerging in African contexts. Despite a high orthographic convergence achieved through reform, the linguistic and cultural trajectories of Portugal and Brazil have diverged considerably. Differences in vocabulary, syntax, and usage reflect broader social and cultural distinctions between the two societies. This internal diversity represents both an asset and a challenge for computational linguistics and language technology.

1 https://www.cidehusdigital.uevora.pt

In the digital era, the distribution of Portuguese language resources mirrors existing global inequalities. Brazil's demographic scale and technological infrastructure ensure a significantly greater online presence, while European and African Portuguese varieties remain underrepresented in most digital corpora. This digital asymmetry has critical implications for computational models, which learn linguistic patterns from large-scale online text data. As a result, most existing models tend to reflect the lexical, syntactic, and cultural characteristics of Brazilian Portuguese, marginalizing other varieties. This imbalance affects not only language representation but also the inclusivity and fairness of AI applications. Addressing these disparities is essential to ensure that future language technologies reflect the full linguistic, and cultural diversity of the Lusophone world.

Currently, we see a profusion of efforts to develop GLLMs, both worldwide and within countries.

There are multilingual models, and others are specialised for a single language. The same proliferation can be seen within models specialized for Portuguese. The Gervásio models are available with 1 and 8 billion parameters.[2] GlórIA[3] is another generative model with 1 billion parameters. There are also Portuguese models developed in Brazil, such as Sabiá[4], AmazoniaIA[5], and SoberanIA[6].

If we consider the full context large language model (FLLMs) (see Simons et al., 2026), there are other options. The Portuguese Albertina family[7] includes full context models of different sizes, with 100 and 900 billion parameters. Developed in Brazil, we have Bertimbau[8], and even domain specialized models such as LegalBERTpt[9].

While GLLMs are good in text generation tasks, such as question answering, and machine translation, where they, based on an input sentence, continue by generating a sequence of words, FLLMs, which are lighter, simpler models, perform best in classification tasks, such as named entity recognition (NER), sentiment analysis, and others.

These Portuguese LLMs (generative and full-context) can be downloaded from the national repository of PORTULAN CLARIN[10], the Portuguese National Research Infrastructure for Science and Technology of Language. The international repository, Hugging Face, is another platform where models for the Portuguese language can be found.

The recent development of GLLMs for Portuguese extends beyond technical concerns; it also embodies cultural and epistemological implications. Language models encode not only grammar and vocabulary but also the social and cultural contexts embedded in their training data. This observation underscores the need for inclusive approaches to data collection and model training. Ensuring that African, Asian, diasporic, and past varieties of Portuguese are represented in computational corpora is essential for achieving a truly global and equitable digital presence.

---

2 https://huggingface.co/PORTULAN/gervasio-8b-portuguese-ptpt-decoder
3 https://nova-lincs.di.fct.unl.pt/gloria-the-new-portuguese-european-large-language-model/
4 https://www.maritaca.ai/
5 https://amazoniaia.com.br/
6 https://sia.pi.gov.br/projetos/soberania/
7 https://portulanclarin.net/models/
8 https://github.com/neuralmind-ai/portuguese-bert
9 https://huggingface.co/raquelsilveira/legalbertpt_fp
10 https://portulanclarin.net/

In the rapidly developing field of artificial intelligence, it is crucial that the entire diversity of the Portuguese language be visible, usable, and appropriately represented, even though these methods are better at expressing synchronic rather than diachronic issues. This corroborates Simons et al. (2026) who indicate the need of LLMs literacy for the humanities researchers.

3. Challenges for specific periods: the case of the 18th century

Languages also have variation across time. Crossing different historical linguistic periods, the 18th-century stage of the language, although modern, is still distinct from the contemporary stage. Textual sources are not always in digital support, especially in languages with great textual memory, such as Portuguese.

Strict transcription guidelines are required for transcriptions. These guidelines are not always the same. This is true even when paleographers decide either to transcribe with absolute fidelity to the original or to update the language to a contemporary pattern. What to do with old words still in use, erratic use of semi-capital and capital letters, letters out of Roman pattern used as numerals[11], change of geographical names in time, abbreviations, the long 's', and other constraints? Also, transcriptions made by several human transcribers may have different outputs, bringing inconsistency to the *corpus* collected.

In the 18th century, regarding the linguistic stage of the language, most of the phonetic, phonological, syntactic, and other changes were already concluded or in consolidation, as discussed by (Cardeira, 2006, Castro, 2006, and others equally pertinent to this question). However, spelling was out of pattern till the beginning of the 20th century. We find many pseudo-etymological digraphs (e.g., -th-), non-etymological double consonants (e.g., -ll-), use of "u" with consonantal value, "i"/"j" as allographs, or "y" as a semi-vowel, as presented by Cameron et al (2023) regarding the *Parish Memories corpus*. In the same text or the same *corpus*, the same word can be written in several ways. For example, in the cited *corpus*, the word "Assunção" [Assumption] has 20 variants.

So far, we have manually transcribed and normalized the *corpus* under study, aiming to diminish the effects of these issues and increase accuracy in automatic tasks. For that, we developed a conservative approach, trying to perform the minimal changes to maintain fidelity to the original. We maintained all linguistic variation, old words still in use, punctuation, but we updated spelling to the contemporary pattern. All this path, with constraints at several levels (transcription, variation, normalization) necessary for textual sources preparation, still can not be directly solved with available GLLMs, as far as we know.

GLLMs are mostly based on contemporary versions of the language. Although they do present an adaptability capacity, in our research, we observed some difficulties. GLLMs need to be trained on large volumes of texts from different historical periods; however, such texts are not available in sufficient quantity or in a condition suitable for large-scale

11 E.g. 'R' equal to 40.

processing. For that, all kinds of materials should be considered, including administrative, legal, and everyday documents, to represent the full social spectrum rather than privileging elite voices, avoiding bias as much as possible.

4. Challenges considering specific tasks: the case of Named Entity Recognition

Named entity recognition (NER) is a Natural Language Processing (NLP) task that helps historians to answer the questions, Who? Where? What? When? The main categories usually taken into account for this task are PERSON, PLACE, ORGANIZATION, and TIME. For the purpose of historical research, these categories might need proper specialization. In our previous work, the categories were subdivided into several subcategories. We must always have in mind that we are dealing with a past society, necessarily different from the contemporary one: it was a hierarchical society, ruled by privileges, where inequality is assumed in the law. For that, for example, the PERSON category was subdivided into Saints, Social Category, Name, Occupation, Author, and others (Vieira, 2025), trying to better describe the reality in its complexity. In addition, annotation was carried out with the broader context in mind, considering the larger expression to assist with homonymy and enable disambiguation.

Annotated corpora are useful resources. The manual annotation process, however, is quite demanding, so this has become an important NLP task. NER is mainly based on machine learning models. The first approaches were based on predefined features. As the learning models evolved, less and less feature pruning was required. The models were developed in a way that the features could be captured entirely from observed examples. For the development of these systems, the existence of human made datasets are required to be able to train the models (see Zang and Colavizza, 2025).

The new developments in GLLMs, however, did not prove to bring improvements in this particular task. Following the fast evolution of the language models themselves, we have been testing available models, both generative and full-context, aiming for the improvement of the NER task performance (Nunes et al., 2025). Our last best result was achieved with Albertina, a full-context BERT-like model specifically trained on European Portuguese, and not by a generative model.

The limitations of GLLMs here may be connected to the fact that NER is a sequence labeling task. It performs a classification of tokens based, among other things, on their position in a sentence. Each token is identified as pertaining to a named entity or not, and a category for those identified as named entities has to be pointed out. In the sequence, we present an example of the required output for the NER task, where each token in a sentence is classified according to predefined categories; 'B' indicates the beginning of an entity, 'I' is the continuation of the identified expression (inside), 'O' indicates the token is not referring to a named entity (outside). In the following example in Table 1, the general category is place (PLC), specialized for facility (FAC).

Table 1: Example NER output

se	O	Terceiros	I-PLC_FAC
vai	O	de	I-PLC_FAC
seguindo	O	São	I-PLC_FAC
até	O	Francisco	I-PLC_FAC
o	O	,	O
convento	B-PLC_FAC	sempre	O
de	I-PLC_FAC	subindo	O
Santa	I-PLC_FAC	acima	O
Clara	I-PLC_FAC	até	O
e	O	o	O
igreja	B-PLC_FAC	castelo	O
dos	I-PLC_FAC		

GLLMs are language generators. They present very well-written sequences which, however, may describe situations that do not correspond to known facts. This has been called hallucinations, and this is due to the "creativity" abilities of these models. This presents specific challenges for historical analysis, which is grounded in empirically verified data, statistically supported patterns, or qualitatively established dominances – never in made-up information. Consequently, the risk that these models may prove unreliable for historians is significant. In the cases where the expected result has to follow a strict structure, such as the one required for the NER task, this creativity capacity of the model may interfere in a negative way. Another relevant difference, here, between the two types of models is that full-context models have bidirectional information about each word in the sequence. They work by seeing both left and right contexts of each word, which especially affords tasks such as NER.

Whereas GLLMs may present excellent performance for question-answering, dialogue, language translation, and reinterpretation, some specific tasks may not yet be best attended by these powerful models. Nevertheless, we are aware that this technology rapidly evolves, and it might be the case that the next models may overcome current limitations.

5. Final remarks

In historical analysis, both the time variable and context are indispensable. Besides questions of duration, time must be understood as the condition for dynamism and change, while context reflects the inherent complexity, diversity, and heterogeneity that characterise social and natural phenomena. Since history is a form of discourse fundamentally grounded in texts and deeply shaped by the temporal dimension—which inherently brings change—it is crucial to acknowledge these challenges to advance historical analy-

sis. With the arrival of GLLMs, the ability of dealing with information coming from vast amounts of text was increased, influencing the way research is conducted by historians, linguists and others.

However, these models, as seen before, still have limitations in dealing with the past dimension. For the studies performed by the authors of this chapter, the main limitations are due to specific language, time and task. Critical approaches are needed for reflecting on the present and the future—without losing sight of the past as, we believe,societies without memory cannot survive. Currently, these models, as far as we know, do not adequately address the need for diachronic variation, a crucial aspect for digital history.

Having AI tools tailored to languages and their variants has become essential. This includes the creation of specialized training corpora for earlier stages of particular languages, the development of domain-specific fine-tuning strategies, and hybrid systems that can enforce formal constraints. Only through such targeted efforts GLLMs can become genuinely useful tools for historical linguistics and digital history. Until then, their use must remain critically mediated by human expertise, ensuring that automated processes do not obscure the interpretive depth and contextual sensitivity that historical research demands. Multidisciplinary teams are the only ones who can operate this change.[12]

Acknowledgements

This work is funded by CIDEHUS-University of Évora and the Foundation for Science and Technology (Portugal), under the project UID/00057/2025 - https://doi.org/10.54499/UID/00057/2025

References

Büttner J (2026) Why pursue temporally-grounded AI for historical disciplines, and what makes it so challenging? In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-2.

Cameron HF, Olival F, Vieira R (2023) Planear a normalização automática: tipologia de variação gráfica do corpus das Memórias Paroquiais (1758), *LaborHistórico*, Rio de Janeiro, ISSN 2359–6910, vol.9, n.1. DOI: 10.24206/lh.v9i1e52234

Cameron HF, Olival F, Vieira R (2022) INCEpTION-Use Cases: Portuguese Parish Memories (1758–1761), available at: https://inception-project.github.io/use-cases/parish-memories/

Cardeira E (2006) *O essencial sobre a História do Português*. Alfragide: Editorial Caminho.

12 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Castro I (2006) *Introdução à História do Português*. Lisboa: Colibri, 2nd edition.

Gonçalves MF (2024) Contribuições para o estudo do português falado do século XVIII: o Compendio de Orthografia (1767) de Monte Carmelo". In: Merlan A and Schäfer-Prieß B, Randromania im Fokus. Gesprochenes Galicisch, Portugiesisch und Rumänisch, Lausanne – Berlin – Bruxelles – Chennai – New York – Oxford: Peter Lang, 540 S. (Reihe Romanistische Arbeiten interkulturell und interdisziplinär, Herausgegeben von Rafael Arnold, Thomas Johnen, Aurelia Merlan, Jürgen Schmidt-Radefeldt und Rudolf Windisch), 393–416.

Instituto Camões (2022) Dia Mundial da Língua Portuguesa – 5 maio 2022, available at https://www.instituto-camoes.pt/images/pdf_noticias/Dados_sobre_a_l%C3%ADngua_portuguesa_de_2022.pdf

Mateus MHM and Bacelar do Nascimento F (2005) A mudança da língua no tempo e no espaço, In: *Língua Portuguesa em Mudança*, Lisboa: Caminho, 13–30.

Nunes RO, Santos J, Spritzer A, et al (2025) Assessing European and Brazilian Portuguese LLMs for NER in Specialised Domains. In: Paes A, Verri FAN (eds) *Intelligent Systems. BRACIS 2024. Lecture Notes in Computer Science*, vol 15412, Springer, Cham, 215–230. ISBN: 978-3-031-79029-4. https://doi.org/10.1007/978-3-031-79029-4_15

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

Underwood T, Nelson LK and Wilkens M (2025) Can Language Models Represent the Past without Anachronism?, arXiv e-prints, Art. no. arXiv:2505.00030, 2025. doi:10.48550/arXiv.2505.00030.

Vieira R, Olival F, Cameron HF, et al (2025) Anotação, análise e aprendizagem de Entidades Nomeadas em textos históricos portugueses (séc. XVIII). *Linguamática*, *17*(1): 121:136. Available at: https://linguamatica.com/index.php/linguamatica/article/view/445

Wolf JC (2026) LLMs and multilingual historical corpora in a digital history project. Reflections from the Berlin workshop. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-2.

Zang S and Colavizza G (2025) Named Entity Recognition of Historical Text Via Large Language Model, preprint arXiv:2508.18090v1 [cs.DL] 25 Aug 2025