Analysing LLMs for spelling normalization of 18th century Portuguese

Helena Freire Cameron[1,2], **Aline Paes**[3], **Fernanda Olival**[2], **Renata Vieira**[2]
[1]Portalegre Polytechnic University, Portugal
[2]University of Évora, CIDEHUS
[3]Universidade Federal Fluminense, Instituto de Computação, Brasil
helenac@ipportalegre.pt, alinepaes@ic.uff.br,
renatav@uevora.pt, mfo@uevora.pt

Abstract

This paper presents an evaluation of large language models (LLMs) applied to the task of normalizing eighteenth-century written texts. Several LLMs were employed to process texts in pre-contemporary spellings and update them according to contemporary Portuguese orthography. Their outputs were rigorously compared against a curated reference *corpus*. The findings indicate marked disparities in model performance, with the Portuguese-specialized model Sabiá demonstrating a statistically significant advantage over multilingual alternatives.

1 Introduction

Although being the object of discussions among scholars throughout the ages - see (Gonçalves, 2003), (Kemmler, 2001) (Silvestre, 2019), and others equally pertinent - , the Portuguese language lacked a standardized orthography until the beginning of the second decade of the twentieth century (Banza, 2019). Consequently, archival collections from earlier periods contain texts in which a single word may appear in multiple orthographic forms (Mateus and Cardeira, 2007), (Martins and Carrilho, 2016). This situation constitutes a significant obstacle to scholarly research, making its resolution a matter of fundamental importance. The absence of standardization hinders not only research activities, but also the accurate reading and interpretation of texts. Even in the case of eighteenth-century materials, reading Portuguese texts can be challenging for non-linguists and non-historians, since spelling conventions differ markedly from those of contemporary Portuguese and the lexicon displays a wide range of orthographic variants. To make historical texts accessible to a broader audience, normalized versions play an essential role. However, producing such versions requires a substantial investment of time, as well as advanced linguistic and historical expertise.

Computational efforts in text analysis have been explored using various methodologies (Jurafsky and Martin, 2000), including machine learning approaches (Goldberg, 2018), as discussed in the next section. However, as far as we know, results are still in the incipient stage for the Portuguese language. Using generative LLMs to better normalize spelling in 18th-century Portuguese texts is the starting hypothesis of this work. This is not a linear process, as old texts also serve as a mirror of past realities. In the language, we have different spellings, obsolete vocabulary, different unit measures for distance, weight, and others. The society is necessarily different from the contemporary one, more hierarchical, with different social categories, titles and organizations. These factors raise an essential research question: to what extent are generative models capable of producing accurate and contextually appropriate normalization in pre-contemporary texts? This question necessarily requires a clearer understanding of both the historical status of Portuguese in the eighteenth century and the nature of the normalization process itself, which are addressed in the following section.

2 Normalization process

The Portuguese language has a long historical trajectory and is spoken worldwide as a mother tongue (Reto et al., 2016). Its global presence is largely the result of Portuguese expansion during the fifteenth and sixteenth centuries, as well as subsequent diasporas and interactions (Mateus and Nascimento, 2005). In the 18th century, the Portuguese language already exhibited a much higher degree of syntactic stabilization. However, at this stage, there were many spelling variants (Cardeira, 2006). This linguistic stage makes eighteenth-century Portuguese a particularly relevant and methodologically sound starting point for normalization studies.

Texts produced several centuries ago continue to

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 498–506

attract the interest of diverse scholarly and cultural communities, and reading transcriptions without standardized orthography still poses a significant cognitive and linguistic challenge for non-native speakers or non-specialists in the past. To make these texts available to the broad public, it is necessary to have standardized versions alongside the original ones.

Regarding automatic spelling normalization, several efforts have been made, with different approaches: rule-based (Bollman et al., 2011), statistical machine-translation (Sanchez-Martinez et al., 2013), (Pettersson et al., 2013), (Scherrer and Erjavec, 2013), machine-translation (Samardžić et al., 2015), (Amoia and Martinez, 2013), neural machine-translation (Bollmann and Søgaard, 2016) and (Tang et al., 2018), neural and statistical machine-translation (Gabay and Barrault, 2020), among others equally relevant, addressing one or several languages, and one or several historical periods at the same time. We must also refer to other relevant previous works: (Baron and Rayson, 2008),(Baron and Rayson, 2011), (Reynaert et al., 2012). More recently, LLMs have been applied to answer questions in historical texts, particularly in the context of machine translation of place and person names (Fischer and Volk, 2025). Regarding spelling normalization, the work presented in (Domingo and Casacuberta, 2025) performed two approaches, one unsupervised training and the other trained for machine translation. In contrast to these methodologies, the present study evaluates the behaviour of a series of alternative LLMs, when faced with a prompt designed with the goal of acquiring a conservative normalized version for the given original texts.

Regardless of whether it is a human-developed or a machine-approach, when dealing with historical texts, spelling normalization is a complex task, requiring a process with several layers of intervention (see (Cameron et al., 2023)). In the first instance, normalization primarily involves the regularization of orthography (see selected examples in Table 1); however, in more limited cases, it may also entail the modernization of specific lexical forms (e.g., altar-maior → altar-mor; habitadores → habitantes). It may also involve punctuation, aligning it with contemporary standards.

Type	Original	Reference
'u' as consonant	priuilegio	privilégio
nasal diphtongs	sejão	sejam
Clitic pronouns	tomandose	tomando-se
double consonants	ditta	dita
'h'	contheudo	conteúdo
'h'	huma	uma
's' / 'z'	copiozaz	copiosas
'c' / 's'	cittio	sítio

Table 1: Normalization examples

3 Methodology

The research team designed a conservative normalization process of texts whereas the changes to be performed would only be of a spelling nature and minimal updating of punctuation, to maintain the lexical characteristics of the text. Additional changes in the process must also be implemented like correction of errors of the writers or updating disused place names. For that, the aimed automatic normalization is then much more tailored, to maintain the linguistic and discourse variation, assumed by the team, which represents an additional challenge to the computational process.

To evaluate the capacity of contemporary LLMs to perform historical natural language processing and orthographic normalization of eighteenth-century Portuguese, we employed a standardizing prompt specifically designed for this task, exhibited in Figure 1. The prompt defines strict constraints on output format, permissible linguistic transformations, orthographic rules, and JSON-only responses, to facilitate post-processing. It also provides explicit instructions regarding preservation of lexical items, handling of clitic pronouns, removal of noise markers, and normative orthographic adaptation. All models were queried with exactly the same prompt to ensure experimental comparability.

3.1 Prompt

The prompt was originally formulated by the research team, grounded in their domain expertise and close familiarity with the *corpus*. It was later refined with the aid of ChatGPT-5.1. Throughout this refinement process, the core rationale and methodological essence of the prompt were fully preserved; however, ChatGPT contributed by making the model's roles more explicit and by clearly defining output formatting rules, particularly with respect to the structure and consistency of the generated JSON.

The prompt was designed specifically for the

reality of 18th-century Portuguese texts, and the spelling normalization was designed to maintain all historical characteristics of the texts. For that, the research team decided not to update some words, although disused, but still in use currently in some contexts: 'mui' (for 'muito' [a lot]), 'cousa' (for 'coisa' [thing]), 'el-Rei' (ancient form for 'o' [the King]), 'Sua Senhoria' [His lordship/Her ladyship], 'Vossa-mercê' (old title) (see (1) in the Prompt). The word 'connosco' was highlited here as it is the only word with double 'n' in European pattern, and, therefore, it must not be normalized.

Regarding Spanish and Latin words, they occur frequently in some texts, inside the discourse, and their translation was not wanted (2). Another issue to be normalized by the models is the transcription mark [sic], inserted in the text by the transcriptor, to show that the word may not be correct, but its transcription is validated (3). In the human normalizing process, we have removed all occurrences, and it is aimed that they must also be removed in the automatic normalization. Concerning the fourth command of the prompt, it is related to the placement of clitic pronouns in Portuguese (European). In pre-contemporary texts, they are often juxtaposed to the verb, whereas currently they are separated by a hyphen.

Instructions (5), (6) and (8) are related to expected changes, where the model can change more than a word per sentence (5), more than a character per word (6) and the mandatory insertion of text in the JSON even when no changes have been performed. Instructions related to graphical accentuation, capitalization and hyphenation are included in (7), where the model is asked to update words to the contemporary pattern.

3.2 Models

We conducted experiments using four LLMs: Sabia-3 (Abonizio et al., 2024), Qwen-3-32B (Yang et al., 2025), Llama-70B-Versatile (Dubey et al., 2024), and GPT-OSS-120B (Agarwal et al., 2025). The Sabia-3 model was accessed through its native API, while the remaining three models were executed via the Groq inference platform (Inc., 2024), which offers low-latency, high-throughput execution for state-of-the-art language models. We selected these models because they are available on Groq and because they have demonstrated strong performance in Portuguese-language tasks (especially Sabiá-3 and Qwen-3 as shown by recent Portuguese benchmarks (Almeida et al., 2025)). For each model, we collected the JSON outputs and assessed their adherence to the prompt constraints, orthographic accuracy, and consistency with the expected structure. No additional post-processing was applied, except for verifying JSON syntactic validity as required by the prompt. `Temperature` was set to 0.0 and `max_tokens` to 16,384, except for Sabia-3, given its limit is 12,000 tokens.

3.3 Dataset

We selected 5 texts from the eighteenth-century *Parish Memories corpus*, which comprises responses to a sixty-question survey administered to parish priests in 1758 (Olival et al., 2023). This *corpus* brings together testimonies produced by the same type of respondents - parish clergy - from across Portugal. It encompasses a wide diversity of authors and topics, ranging from demographic data to descriptions of river fish. Each text selected originates from one of the principal towns in the Alentejo region, in southern Portugal. For each memory, a manually transcribed version of the original manuscript and a corresponding manually normalized version, produced by the authors, were available. The transcribed version was given to four different LLMs, which were asked to produce a normalized version in response to the prompt presented in Figure 1. Consequently, the resulting dataset comprises six alternative versions for each of the five selected memories.

Examples of the original, reference and model output follows.

Original transcribed version *Excelentissimo e Reverendissimo Senhor Ordename Vossa Excelencia Reverendissima que responda ao que nos interrogatorios por Vossa Excelência enviados se pergunta e satisfazendo ao presc[r]ito [?] com a obediencia de humilde subdito respondo o seguinte. A freguezia de são Iullião, (da qual sou indigno Parocho) está cituada no Termo da Villa de Maruão Bispado, e Comarca da Cidade de Portalegre tem noventa e tres fogos Com duzentas e Vinte pessoas só a El Rey nosso senhor reconhece por senhorio Tem a sua Cituasão parte Em Valles, e parte em montes = as serra que por todas as partes a cercão inpedem que della se devizem terras algumas = confina a freguezia Com a a Arraya de Castella, e com os termos de Portalegre, e Alegrette terras do Reyno de Portugal, e com os termos de Vallença de Alcantara, São Vicente, e Codiceyra terras do*

Prompt

You are a highly accurate model for historical natural language processing and orthographic normalization. Your role is to convert eighteenth-century Portuguese texts to contemporary orthography, without altering style, vocabulary, or syntactic structure.
-
You must respond exclusively in valid JSON format, with no comments, explanations, or additional text. Before sending the final output, internally validate that the JSON is well-formed (with no syntax breaks, missing braces, or incorrect quotation marks). If any formatting issue is found, correct it automatically before outputting the valid JSON.
-
Main rules: no explanations outside JSON; no introductory text; no extra, unrequested fields; guarantee JSON parsing compatibility in any programming language. You will receive one or more text excerpts in eighteenth-century Portuguese.
-
These texts contain pre-contemporary spellings that must be updated according to contemporary Portuguese orthography.
-
Instructions:
(1) Do not alter the following words or expressions: mui, cousa, el-Rei, Sua Senhoria, Vossa mercê, connosco.
(2) Do not alter words in Spanish or Latin.
(3) Remove all occurrences of [sic].
(4) In verb forms with clitic pronouns (e.g., "disselhe," "acostumamonos"), add a hyphen between the verb and the pronoun.
(5) Each sentence may contain several words to correct.
(6) Each word may require multiple letter changes.
(7) Respect contemporary norms for accentuation, capitalization, and hyphenation.
(8) If a text excerpt contains no words to correct, it must still be included in the JSON.
-
Required output format (a single JSON): "textos": ["título": "Título 1", "texto_corrigido": "Texto 1 corrigido." , "título": "Título 2", "texto_corrigido": "Texto 2 corrigido."]

Title:
`{Original text: }`

Figure 1: Prompt provided to the LLMs. The original prompt is written in Portuguese.

Reyno de castella: ...

Reference *Excelentíssimo e Reverendíssimo Senhor Ordena-me Vossa Excelência Reverendíssima que responda ao que nos interrogatórios por Vossa Excelência enviados se pergunta e satisfazendo ao prescrito com a obediência de humilde súbdito respondo o seguinte. A freguesia de São Julião, (da qual sou indigno pároco) está situada no termo da vila de Marvão bispado, e comarca da cidade de Portalegre tem noventa e três fogos com duzentas e vinte pessoas só a el-Rei Nosso Senhor reconhece por senhorio Tem a sua situação parte em vales, e parte em montes = as serras que por todas as partes a cercam impedem que dela se divisem terras algumas = confina a freguesia com a Raia de Castela, e com os termos de Portalegre, e Alegrete terras do reino de Portugal, e com os termos de Valença de Alcântara, São Vicente, e Codiceira terras do Reino de Castela: ...*

LLM version (Sabia) *Excelentíssimo e Reverendíssimo Senhor, Ordename Vossa Excelência Reverendíssima que responda ao que nos interrogatórios por Vossa Excelência enviados se pergunta e satisfazendo ao prescrito [?] com a obediência de humilde subdito respondo o seguinte. A freguesia de São Julião, (da qual sou indigno Páraco) está situada no Termo da Vila de Marvão, Bispado e Comarca da Cidade de Portalegre. Tem noventa e três fogos com duzentas e vinte pessoas, só a El-Rei nosso Senhor reconhece por senhorio. Tem a sua situação parte em vales e parte em montes; as serras que por todas as partes a cercam impedem que dela se devolvam terras algumas. Confina a freguesia com a Arraia de Castela e com os termos de Portalegre e Alegrete, terras do Reino de Portugal, e com os termos de Valência de Alcântara, São Vicente e Codiceira, terras do Reino de Castela; ...*

3.4 Difference tables

A diff Python script was applied to produce difference tables of the LLMs' results in the JSON output, original and reference. Tables 2 and 3 show,

as an example, tables for the original and LLM, and reference and LLM, respectively. There are replaces and inserts. Note that the table lines vary in the number of tokens. Note that differences in the reference involve the insertion of punctuation and changes in capitalization.

3.5 Evaluation

In the evaluation process, three tables were generated to illustrate the differences between the original texts and both the manually and automatically normalized versions. The differences are presented in blocks of tokens, not allowing a precise alignment between the three tables. Therefore, we cannot provide measures such as precision and recall. Instead, the evaluation was based on the number of changes performed by the models to transform the original 18th century into normalized texts, and the number of errors observed in comparison with the reference. We counted 1) the number of changes made from the original in the manual reference (Reference Changes) 2) the differences from the original for each LLM (LLM name Changes), and 3) the number of errors by analysing the differences between each LLM output and the reference (Errors). Note that errors are those cases in which the normalized variants produced by the model were wrong and also those cases in which the model didn't perform the change it should have done. Differences in punctuations, capital letters, numerals, and alternative accents (acute and grave when both are correct) were not considered as errors. The reason is that the manual reference does not add any punctuation to the original, the texts have some numbering, and accents vary according to European or Brazilian Portuguese. We didn't take into account these differences in error counting. See examples in Tables 4 and 5.

4 Results

Table 6 presents the results. It shows for each source the number of words, the changes made from the original in the reference and in each LLM, and the number of error lines according to diff tables of reference and models output.

As mentioned earlier, the differences in groups of tokens for each diff tables do not allow us to infer measures such as precision and recall correctly; that is the reason for presenting the number of lines in which errors were observed.

The table helps us to see that Sabia performs better, followed by Llama, GPT, and Qwen. Qwen is the model with the smallest number of changes. Its error rate is high since it counts all the changes that were not made. This explains why the number of errors exceeds the number of changes. Sabia and Llama are the ones with the greatest number of changes; however, Sabia has produced fewer errors than Llama. Qwen exhibits different behaviors across the texts, considering that each text was a single run of the prompt. All the other models show a much more stable behaviour.

5 Error analysis

Spelling normalization in historical texts must be performed carefully to avoid corrupting historical reality and linguistic complexity, regardless of the historical period. For that, the changes performed in the normalization process are only of spelling nature, maintaining all the variation inherent of the language, maintaining, for example, variants like 'ouro'/'oiro' [gold], 'terremoto/'terramoto' [earthquake], and so on. Also, words that are not in current dictionaries but may be part of the language were not normalized e.g. 'asperosos' [rough] can be a portuguese word although 'áspero' is the word in dictionaries, or 'valeroso' where 'it can be also 'valoroso' [valorous].

When models do normalization tasks, they perform several changes, of different types. Besides measuring the errors, we must observe what errors models do, and what changes they were supposed to perform, to eventually adjust the process, to have more accurate results.

ChatGPT changed incorrectly:

- bispado [bishopric] → bispo [bishop],
- côngrua [ecclesiastic income]→ coroa [crown],
- roins (for 'ruins' [bad])→ rochas [rocks],
- orago [patron Saint]→ oratório [oratory].

These changes are part of a vocabulary specially related to the church environment that, although still current, may not be common knowledge in today's societies. Other wrong changes in other semantic fields can be observed:

- outto [eight] → outtras [others],
- nimiamente [too much] → nativamente [natively],

Tipo	Trecho A	Trecho B	Índices (A, B)
replace	Excelentissimo e Reverendissimo Senhor	Excelentíssimo e Reverendíssimo Senhor,	28:33 / 16:20
replace	Excelencia Reverendissima	Excelência Reverendíssima	35:37 / 22:24
replace	interrogatorios	interrogatórios	42:43 / 29:30
replace	presc[r]ito	prescrito	52:53 / 39:40
replace	obediencia	obediência	56:57 / 43:44
replace	freguezia de são Iullião,	freguesia de São Julião,	64:68 / 51:55
replace	Parocho)	Páraco)	72:73 / 59:60
replace	cituada	situada	74:75 / 61:62
...	...	...	

Table 2: Original x Model difference

Tipo	Trecho A	Trecho B	Índices (A, B)
replace	Senhor Ordena-me	Senhor, Ordename	31:33 / 19:21
insert		[?]	52:52 / 40:41
replace	súbdito	subdito	57:58 / 46:47
replace	páraco)	Páraco)	70:71 / 59:60
replace	termo da vila de Marvão bispado,	Termo da Vila de Marvão, Bispado	74:87 / 63:76
replace	pessoas	pessoas,	95:96 / 84:85
replace	el-Rei Nosso	El-Rei nosso	98:100 / 87:89
replace	senhorio	senhorio.	103:104 / 92:93
replace	vales,	vales	110:111 / 99:100
replace	montes =	montes;	114:116 / 103:104
replace	divisem	devolvam	129:130 / 117:118
replace	algumas = confina	algumas. Confina	131:134 / 119:121
...	...	...	

Table 3: Reference x Model difference

Type	Reference	Model
change	arcebispado de Évora;	Arcebispado de Évora,
change	páraco	Páraco
change	ordem de Estremoz;	Ordem de Estremoz,
change	igreja paroquial	Igreja Paroquial
change	mais,	mais
change	vila do Redondo:	Vila do Redondo;

Table 4: Examples of changes made by the model, different from the reference, but not considered as errors.

Type	Reference	Model
error	Senhor	Reverendo
error	Reconhecença.	Reconheçença.
error	Freguesia de Santa Ana de Bencatel	Freguezia de santa Anna de Bencatel,
error	Estremoz, arcebispado	Estremos, Arçebispado
error	também	tambem
error	apresentação pertence	aprezentação pertençe

Table 5: Examples of changes made by the model, different from the reference, and considered as errors.

Source/ Words	Reference Changes	Changes	Sabia Errors	Changes	Llama Errors	Changes	GPT Errors	Changes	Qwen Errors
Evora/1501	236	253	10	255	25	240	41	205	76
Portalegre/2153	188	187	42	187	58	196	83	27	162
Vila Viçosa/1241	211	230	27	230	50	211	61	28	197
Elvas/1423	254	256	15	251	47	239	66	73	184
Beja/2791	235	259	27	238	53	243	73	32	120
Total/9109	1124	1185	121	1161	233	1129	324	365	739

Table 6: Number of proposed changes (Changes) from the original by reference and each model, and number of lines with errors verified in the diff tables of the manual reference x proposed by the models (Errors).

where the model did not recognize or did not perform well the required change. This model did not succeed to change 'leguas' to 'léguas' [unit of distance], or 'moyos' to 'moios' [unit of measure], among other examples.

Llama changed incorrectly:

- a el-rei → ao Rei[to the King],
- sereijas → cereijas[cherries],
- Alemtejo → Além-Tejo [Alentejo, a Portuguese region],
- acodirão → acodem [help],
- asolavão [they devastate] → asolou [it devastate],
- prepozito → prepozo (for 'propósito' [purpose]),
- devizão → devisam (for 'divisão' [division]),
- correyo [post office] → corrente [flowing].

The model didn't succeed to change 'sego' to 'cego' [blind], 'cincoenta' to 'cinquenta' [fifty], 'delle' to 'dele' [his], and others.

Regarding the Qwen model, it changed incorrectly:

- delle → d'elle (for 'dele' [his]),
- amtiguos → antiguos (for 'antigos'[old]),
- colatraez → colatraes [collateral],
- demenutas → deminutas [few],
- cazeyros → caizeiros (for 'caseiros' [Farm stewards]),
- vila → villa (for 'vila' [village]).

Qwen model didn't succeed in changing basic words, like: 'cecenta e sinco' to 'sessenta e cinco' [seventy five], 'hum' to 'um' [one], 'freguezia' to 'freguesia' [parish], 'legoa' to 'légua' [distance mesure], 'meya' to 'meia' [half], 'çento' to 'cento' [hundred], 'distançia' to 'distância' [distance], 'dous' to 'dois' [two], 'agoa' to 'água' [water], and many other examples.

Regarding Sabiá model, it changed incorrectly some words like:

- couza → coisa [thing],
- decliuio → declínio (for 'decline'[decay]),
- sem proporção [without proportion] → desproporcionadamente [disproportionately],
- criaóse (for 'criam-se' [they create]) → Criase [it creates],
- orago [the patron saint of a Catholic church] → oráculo [oracle],
- Alentejo [Portuguese region] → Além-Tejo,
- canteiras [flower beds]→ pedreiras [quarries].

Sabiá didn't change words like 'Caya' to 'Caia' [name of town], 'Seuora' to 'Sévora' [name of a river] and 'Souza' to 'Sousa' [family name].

6 Conclusion

To the best of our knowledge, this is the first attempt of normalization of 18th century Portuguese texts using LLMs and a reference corpus. The experiment shows the performance of different models, where a Portuguese specialized model, Sabia, clearly outperforms other multilingual models. Given the nature of the problem, we consider that this path is worth trying. We now have better ideas for improvements of the process.

The experiment also served to reflect on the normalization task. It is important to have a clear

idea of the purpose of the normalized version. For instance, the manual reference was based on the assumption of minimal interference, mainly spelling adaptation to the contemporary pattern. This preserves historical and linguistic features that are very important for many studies. However, other goals for the normalized text produced, for instance, readability, may require other assumptions.

6.1 Limitations

We encountered some difficulties in evaluating the models. Although we had a reference as a basis for the evaluation, the process required manual analysis of the outputs, especially due to the difference tables. They were based on the diff computation of two texts. However, not all differences were considered errors. We didn't penalize, for instance, the inclusion of correct punctuation, or the differences in capital letters. For automating this analysis, some improvements in the process are needed. These issues had an impact on the way we could report errors, as metrics such as recall and precision could not be calculated automatically.

Another issue to be considered is that our choice of models were limited by the platform we used to run the experiments.

6.2 Future work

After this first experiment, we have many new insights about the problem, which may lead both to a new prompt design and evaluation methodology. Considering the evaluation challenges we faced, we plan to change the output formats to be able to evaluate it with less human efforts in the output verification against the reference.

As we have seen a better performance for one Portuguese LLM, there are other Portuguese LLMs, such as GlorIA[1], and Gervásio [2] that we would like to evaluate in the future. In addition, we intend to expand our exploration to encompass other prompt settings, with a particular emphasis on few-shot prompting.

Acnowledgments

This work is funded by national funds through the Foundation for Science and Technology (FCT - Portugal), under the project UID/00057/2025, partially financed by CNPq (National Council for Scientific and Technological Development), grant 307088/2023-5, FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We also thank the support of the CNPq National Institutes of Science and Technology, IAIA (grant 406417/2022-9), TILD-IAR (grant 408490/2024-1).

[1] https://huggingface.co/NOVA-vision-language/GlorIA-1.3B

[2] https://huggingface.co/PORTULAN/gervasio-7b-portuguese-ptpt-decoder

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabia-3 technical report. *arXiv preprint arXiv:2410.12049*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Thales Sales Almeida, Rodrigo Nogueira, and Hélio Pedrini. 2025. Poeta v2: Toward more robust evaluation of large language models in portuguese. *arXiv preprint arXiv:2511.17808*.

Marilisa Amoia and Jose Manuel Martinez. 2013. Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In *Proceedings of the 7th workshop on language technology for cultural heritage, social sciences, and humanities*, pages 84–89.

Ana Paula Banza. 2019. Variação linguística e ortografia: das regras renascentistas ao ao90. In *Estudos de historiografia linguística portuguesa*, pages 9–22. Faculdade de Letras da Universidade do Porto / Centro de Linguística da Universidade do Porto.

Alistair Baron and Paul Rayson. 2008. Vard 2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*. Aston University, Birmingham.

Alistar Baron and Paul Rayson. 2011. *Dealing with spelling variation in Early Modern English texts*. Lancaster University.

Marcel Bollman, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. *arXiv preprint arXiv:1610.07844*.

Helena Freire Cameron, Fernanda Olival, and Renata Vieira. 2023. Planear a normalização automática: tipologia de variação gráfica do corpus das memórias paroquiais (1758). *LaborHistórico*, 9(1):2359–6910.

Esperança Cardeira. 2006. *O essencial sobre a História do Português*. Caminho.

Miguel Domingo and Francisco Casacuberta. 2025. Two spelling normalization approaches based on large language models. *arXiv:https://arxiv.org/pdf/2506.23288v1*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Dominic P. Fischer and Martin Volk. 2025. Name consistency in llm-based machine translation of historical texts. In *Proceedings of Machine Translation Summit XX*, page 204–219.

Simon Gabay and Loïc Barrault. 2020. Traduction automatique pour la normalisation du français du xviième siècle. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole - JEP, 33e édition, Traitement Automatique des Langues Naturelles -TALN, 27e édition*.

Yoav Goldberg. 2018. Neural network methods for natural language processing. *Computational Linguistics*, 44(1):194–195.

Maria Filomena Gonçalves. 2003. *As ideias ortográficas em Portugal: de Madureira Feijó a Gonçalves Viana (1734-1911)*. Fundação Calouste Gulbenkian.

Groq Inc. 2024. Groq: Ultra-low latency llm inference platform. https://groq.com/. Used for inference of Qwen-3-32B, Llama-70B-Versatile, and GPT-OSS-120B.

Daniel Jurafsky and James H Martin. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Rolf Kemmler. 2001. Para uma história da ortografia portuguesa: o texto metaortográfico e a sua periodização do século xvi até à reforma ortográfica de 1911. *Lusorama. Zeitschrift für Lusitanistik. Revista de Estudos sobre os Países de Língua Portuguesa*, 47-48:128–319.

Ana Maria Martins and Ernestina (Eds) Carrilho. 2016. *Manual de Linguística Portuguesa*. De Gruyter.

Maria Helena Mira Mateus and Esperança Cardeira. 2007. *Norma e variação*. Editorial Caminho.

Maria Helena Mira Mateus and Fernanda Bacelar Nascimento. 2005. *A Língua Portuguesa em Mudança*. Caminho.

Fernanda Olival, Helena Freire Cameron, Fátima Farrica, and Renata Vieira. 2023. As memórias paroquiais (1758) do atual concelho de vila viçosa. *Callipole: revista de Cultura*, (29):85–128.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2013. Normalisation of historical text using context-sensitive weighted levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 163–179. Linköping University Electronic Press, Sweden.

Luís Antero Reto, Fernando Luís Machado, and José Paulo Esperança. 2016. *Novo Atlas da Língua Portuguesa*. ISCTE-IUL / Instituto Camões.

Martin Reynaert, Iris Hendrickx, and Rita Marquilhas. 2012. Historical spelling normalization. a comparison of two statistical methods: Ticcl and vard2. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), Lisbon*. Colibri.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference*, pages 294–298. University of Zurich.

Felipe Sanchez-Martinez, Isabel Martinez-Sempere, Xavier Ivars-Ribes, and Rafael C. Carrasco. 2013. An open diachronic corpus of historical spanish. *Language Resources & Evaluation*, 47:1327–1342.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical slovene words with character-based smt. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 58–62.

João Paulo Silvestre. 2019. *Escrita e Ortografia*. Centro de Linguística da Universidade de Lisboa.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Vivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. *Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, USA*, pages 1320–1331.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.