

Decoding Diplomacy: GPT in Historical Archives, AI neutrality, Linguistic Biases and Multilingual Challenges (Tunisia Czechoslovakia, 1956–1992)*

Amani Mansouri^{1,†}

¹ University in Prague, Technická 2, CZ-166 27 Prague 6, Czech Republic.

Abstract

This study evaluates the versatility, capabilities, and limitations of GPT-4o and GPT-5 Thinking in the analysis of complex, multilingual historical archival documents. The reference corpus concerns diplomatic relations between Tunisia and Czechoslovakia (1956–1992), composed of sources in French and Czech digitized through OCR from scans of variable quality, often affected by significant visual noise. The experiments were intentionally conducted through the standard ChatGPT interface (no parameter tuning), to mirror early-stage archival workflows used by researchers and archivists. The objective is to assess how far these models can support intermediate phases of large-scale digitization, surface recurring biases, and foster critical reflection on the value and limitations of AI for historical research. Future replications will use the API to systematically modulate temperature, verbosity, and output length for full procedural reproducibility.

Keywords

History of technology, Manual digitization, OCR documents, GPT, Tunisia, Czechoslovakia ¹

1. Introduction

The digital transition of historical documentary archives has altered the relationship between researchers, archivists, and the volume of available data. Over time, the need for new automated techniques for the initial analysis, interpretation, and taxonomization of archival sources on a large scale has indeed increased. The need for study analysis has also emerged to understand the accuracy of new digital tools and their capacity to deeply comprehend an archival source and relate it to other sources, particularly regarding the transparency of the data output and the trust in tools that are in a constant process of update or still in an experimental phase[1]. The adoption of artificial intelligence tools and also Generative AI in the domain of archival studies is growing rapidly[2] to scale documentary workflow processes but also to experiment with entirely innovative methods for describing and cataloging sources[3]. From an overview of recent studies, the historical archival field has already embraced artificial intelligence tools for the description, indexing, and management of archival-documentary holdings with applications and software ranging from image recognition, the decoding of handwritten script, to NLP (Natural Language Processing Model)[4].

Experimentations with large language models show significant and growing potential for generating metadata compatible with existing archival standards, also through the use of AI agents[5]. Rapidly advancing technologies such as optical and handwritten character recognition (OCR/HTR) make it possible to convert previously scanned text or text with a lot of noise and render it into text with satisfactory accuracy, likewise, NLP techniques allow for the identification of named entities and themes within previously unstructured texts, fostering the exploration, connection, and accessibility of the historical archival heritage[6]. The potential of these

*VIPERC '25: 4th International Conference on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding, 3–4 December 2025

The paper is submitted to the Special Session Etech-CH

✉ mansoama@fel.cvut.cz (M. Amani)

ORCID 0009-0001-1599-6547 (M. Amani)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

methodologies, which fall within the scope of Digital Humanities, highlights how the scale, i.e., the size and heterogeneity of an archival holding, and the speed of interpretation and classification of sources are no longer limitations. Also using accessible tools, they show how the combined use of HTR, for example with Transkribus, and generative model interfaces like GPT-4o, have enabled us to make historical textual and epistolary collections usable, accelerating transcription and the archival digitization workflow[7]. Despite new technologies emerging to enhance archival digitization, the sector faces numerous obstacles. Many archives have incomplete and delayed digitization processes with discontinuities in availability to the public and researchers[8]. Furthermore, gaps emerge in data processing through digital tools, in a sector that heavily depends on public funding and in which the digital literacy of staff is often insufficient, with inadequate human and financial resources. Since these technologies are in continuous development and update, stable professional collaboration and consolidated trust in tools like AI and HTR are critical, the lack of transparency in outputs can undermine the recognition of work already done and hinder reliable relationships between sources and interpretative methods[1]. Despite promising new developments, it becomes necessary to test and therefore prioritize the development of shared guidelines, validation practices, and continuous training programs so that the integration of AI into archives happens in an ethical, transparent manner and is truly complementary to human work.

1.1. Exploring AI Capabilities in Multilingual Diplomatic sources

The use of artificial intelligence tools in the study of diplomatic decisions requires greater caution, especially in the presence of sensitive scenarios where national preferences can alter the outputs[9]. The CFPD Benchmark[10] highlighted that models are not neutral and that outputs in foreign policy contexts vary significantly based on the national context. Implicit biases thus emerge from the interpretative results. The study therefore suggests caution and human supervision in diplomatic contexts[2], these risks could increase in historical-diplomatic and multilingual contexts where sources may report lexical formalisms, text structures that respond to unclassified sectoral logics, ambiguous or not fully known historical references, unexplained acronyms[11]. Similarly, UNBench[12] uses official UN documents such as resolutions, speeches, and votes to evaluate and test the ability of LLMs to simulate diplomatic decisions[13]. Although the results are promising, the study also reports critical issues regarding text comprehension and the narrative coherence adopted[12]. Relating these risk peculiarities to historical sources, it is necessary to clarify that the risk of an epistemic nature remains substantially plausible and that technical supervision is required[11]. A study focused on transcription starting from German diplomatic sources through the maintenance of layout and formal elements demonstrated an improvement in character error rates (CER) even with the use of relatively small datasets[14]. Research emphasizes that despite the real potential for integrating AI tools in the diplomatic domain, it remains essential to guarantee transparency, human supervision, and possibly standardized procedures to avoid profound interpretative drifts due to biases and errors generated in the procedural phases.

2. Research objectives

The study is configured as an exploratory and comparative investigation with the aim of testing the use of Generative AI interfaces and large language models for the analysis of multilingual historical diplomatic sources through non-specific prompts. The research aims to preliminarily evaluate the potential and limitations of LLMs, the GPT-4o model and GPT-5 Thinking as the latest update from OpenAI. According to the manufacturer's statements, GPT-5 Thinking has improved reliability and reduced factual errors, i.e., fanciful answers or those generated without any real basis. Furthermore, the chain of operations in a single interface context is improved and the procedural outcome is more stable[15]. According to encord.com, GPT-5 Thinking is described as a unified system that internally decides when to use a model to improve efficiency and is a more reflective model for complex tasks. indeed, not only are hallucinations reduced, but the processing

of longer texts, complex conversations, and extensive history is supported more effectively. Therefore, the research aims to analyze a corpus of scanned historical-diplomatic documents, multilingual and often presenting visual scanning noise. We wanted to test this corpus for a specific reason: a good portion of the archival sources currently in archives with an initial digitization process are scanned in this way without a text recognition transcription step. The intrinsic historical complexity and the scanning state of the source allowed us to develop this pilot study to understand how language can influence the analysis of the source and test the models' ability to understand and contextualize historical events, but also to identify recurring internal biases in the output such as hallucinations, omissions, temporal shifts, and finally, to reflect on the historical and epistemological implications of using these tools in historical and diplomatic research and archival science.

3. Materials and Methods

The main corpus comprises 45 documents preserved at the National Archive of Prague, fonds Ministerstvo školství ČSR (1970-1990). These are diplomatic sources in French and Czech, relating to cultural and economic relations between Czechoslovakia and Tunisia. The historical documents were digitized and subjected to OCR. The control group ($\approx 25\%$ of cases, 11 units) consists of texts not relevant to the historical-institutional domain in French or Czech, used to evaluate the models' performance on out-of-domain content. Sampling and randomization were performed with stratification by language and OCR noise level.

3.1. Corpus Description

The experimental corpus consists of 45 digitized diplomatic documents preserved at the National Archive of Prague, fonds Ministerstvo školství ČSR (1970-1990). The collection includes correspondence, reports, and memoranda concerning cultural, educational, and economic relations between Czechoslovakia and Tunisia from 1956 to 1992.

Out of the total corpus, 34 documents are written in Czech and 11 in French. To ensure a balanced evaluation of the models' multilingual performance, a control corpus representing approximately 25 % of the total sample (11 documents) was also defined. This control corpus contains 6 Czech and 5 French texts unrelated to the historical-institutional domain, used to test model behaviour on out-of-domain content. The remaining 34 documents form the core diplomatic corpus on which the main comparative analysis was conducted.

All materials were digitized from heterogeneous scans and subjected to Optical Character Recognition (OCR) processing. The OCR stage followed open-source methodologies aligned with current frameworks for digitizing historical archives, such as the anyOCR system [16], which enables high-accuracy text recognition and layout analysis for complex archival materials. Each document was evaluated and assigned an OCR-quality level from 1 (negligible impairment) to 5 (total impairment) according to the readability and error-density criteria described below.

3.2. Corpus Structure and File Naming

Each document in the corpus was anonymized and renamed according to a systematic convention that encodes the language, OCR quality level, and corpus membership (main or control). The file names follow the structure: "n-XX-N#C"

where:

n = sequential document number (1–45).

XX = language code (CZ for Czech, FR for French).

N# = OCR quality level, ranging from N1 (negligible impairment) to N5 (total impairment), corresponding to the five-level scale described below.

C (final capital letter) = corpus type identifier, indicating inclusion in the control corpus. The absence of “C” denotes membership in the main (diplomatic) corpus.

For example, 1-CZ-N4C denotes the first Czech document, OCR quality level 4, belonging to the control corpus, while 2-FR-N3 identifies the second French document, OCR level 3, from the main corpus.

In total, the dataset contains 45 documents, of which 34 are in Czech and 11 in French. The main diplomatic corpus includes 34 documents (28 Czech and 6 French), and the control corpus includes 11 documents (6 Czech and 5 French).

3.3. OCR noise distribution

To make the dataset composition fully transparent, Table 1 reports the cross-tabulation of documents by OCR level (N1-N5), language (CZ/FR), and corpus (main/control).

Table 1

Frequency of Special Characters.

	CZ-Main (n)	FR-Main (n)	CZ- Control (n)	FR- Contr ol (n)	Row total
N1	5	1	4	4	14
N2	15	0	0	1	16
N3	6	3	1	0	10
N4	2	2	1	0	5
N5	0	0	0	0	0
TOTAL	28	6	6	5	45

Data Examples

To illustrate the variability of OCR quality and the visual noise within the analyzed corpus, short excerpts are presented below. These fragments reproduce the text exactly as extracted from the OCR output, including typographical artifacts, missing characters, and segmentation errors. They exemplify common distortions such as the confusion between letters like “e” and “o,” irregular spacing, or the substitution and loss of accented characters. Such issues are typical of digitized historical materials produced with aged typefaces or faint ink traces.

The majority of the documents were originally typewritten, while handwritten elements appear only in the form of signatures, numbering, or brief marginal comments. These additions, although limited, can still complicate automatic recognition. The following textual fragments illustrate the degree of degradation and heterogeneity encountered across the corpus:

Example (Czech, Level 4 noise): “Ministerstvo školství ČSR ... spolupráce Tunisie ... podpis dohody o výmìni studentù a odborníkù ...”

Example (French, Level 3 noise): “Le Ministère de l’Éducation tchécoslovaque et le Gouvernement tunisien conviennent de renforcer la coopérati0n culturelle ... échanges d’étudiants et enseignants.”

These excerpts demonstrate the range of OCR distortions present in the dataset and highlight the challenges of text recognition and interpretation at higher noise levels (Levels 4-5), where characters are often partially illegible or inconsistently reconstructed.

OCR quality was estimated with a double metric, i.e., a five-level qualitative scale based on readability and error density through the use and textual extraction on txt files:

- Level 1: Negligible (text intact, minimal errors do not hinder understanding).
- Level 2: Slight impairment (sporadic errors with little impact on fluency).
- Level 3: Moderate impairment (frequent errors or distortions requiring contextual inferences).
- Level 4: Severe but recoverable impairment (presence of systematic errors or recurring interruptions).
- Level 5: Total impairment (illegible or severely distorted text). The classification was performed by a single evaluator based on replicable criteria, made transparent through annotations and archiving of classified file.

Model query protocols

Two LLMs (GPT-4o and GPT-5 Thinking) were tested on the same inputs. To ensure replicability, the documents were renamed and randomized. The five questions, formulated directly and presented in a fixed order without prompt design techniques, aim to simulate a preliminary analysis of an unknown source according to this logic:

- Document typology.
- Extraction of main and secondary information (≤ 8 lines).
- Taxonomic list of entities (textual categories, people, places, institutions, dates).
- Historical context (≤ 8 lines).
- Drafting of a synthetic descriptive archival note (≤ 8 lines).

The structure of these questions was developed to replicate the practical logic followed by archivists when approaching an unknown document: identifying its typology, extracting key information, recognizing entities, situating it historically, and finally producing a short descriptive record. This standardized sequence enables the observation of how models handle increasingly complex interpretative tasks under consistent conditions.

Importantly, the design of the prompts was neutral and exploratory. The study did not aim to confirm any predetermined hypothesis about model performance. Instead, it sought to observe spontaneous model behavior in realistic use conditions and to identify recurring patterns of strength and error. This approach minimizes confirmation bias and ensures that the analysis reflects the models' actual generative tendencies rather than any prior expectations.

Prompt Protocols

All runs were conducted using the standard ChatGPT interface rather than the API. This methodological decision was intentional and aimed at reproducing the real working environment of historians and archivists, who typically interact with generative AI through the ChatGPT platform rather than programmatic endpoints. Using the interface ensures ecological validity, reflecting spontaneous, user-facing conditions without parameter engineering, while maintaining consistent default settings across all sessions. The absence of manual parameter control (e.g., temperature, verbosity) was therefore not a limitation but part of the experimental design. Future replications will adopt the API to complement this exploratory phase, allowing for systematic parameter modulation and full procedural reproducibility.

To ensure full transparency and reproducibility of the experimental setup, the exact prompts submitted to both models were standardized and identical across all trials. Each document was

processed in an independent chat session, where the following five questions were presented in a fixed order and without any additional contextual instruction or system message. The prompts were intentionally simple and non-engineered, designed to simulate the spontaneous use of generative AI tools in historical-archival research settings.

Exact prompt questions

1. What type of document is this?
2. Write down the central and secondary information (maximum 8 lines).
3. Identify and report the relevant textual categories, persons, places, institutions, and dates.
4. Clarify the historical context in which the document is situated (maximum 8 lines).
5. Draft a brief archival and descriptive note of the source and the document (maximum 8 lines).

This standardized questioning protocol was applied identically to both GPT-4o and GPT-5 Thinking, allowing for direct comparison of their interpretative behavior. Although the API would offer greater experimental control, such as explicit adjustment of parameters like temperature, verbosity, and response length, the present study deliberately adopted the ChatGPT interface. This approach ensures ecological validity, reproducing the practical working conditions typically encountered by researchers and archivists during early phases of archival digitization. Future extensions of this work will include API-based experiments to enable parameter transparency and procedural reproducibility while maintaining methodological continuity with the present design.

During the analysis of the results produced by the generative artificial intelligence models, it was deemed necessary to develop a taxonomy of epistemic errors to systematically classify the most recurrent distortions. This taxonomy is not merely descriptive but serves as a methodological framework for evaluating the quality, relevance, and reliability of the outputs in relation to historical documentary sources.

For each document, the number of occurrences of every error type (E1-E11) was manually recorded. The values reported in Figures 2-6 represent the total counts of each error category aggregated across all 45 documents, allowing for direct quantitative comparison between GPT-4o and GPT-5 Thinking.

The outputs were evaluated by a single annotator, therefore, inter-annotator reliability indices such as Cohen's κ were not applicable. However, internal consistency was ensured through the explicit use of a predefined and replicable error taxonomy, accompanied by systematic annotations and transparent coding criteria. This approach guarantees methodological rigor and reproducibility despite the single-evaluator setting.

1. E1: Hallucination. This involves the insertion of content that finds no correspondence in the analyzed source or in the reference historical context. The AI produces invented data, names, dates, events, or quotations, generating an artifact that compromises documentary validity.
2. E2: Omission. This is the absence, in the output, of information clearly present in the source. It can concern relevant details (actors, places, decisions) whose exclusion reduces the completeness and interpretative fidelity of the text.
3. E3: Semantic Fusion. This consists of the tendency to merge distinct concepts or entities into a single representation. This error alters the distinction between separate categories, reducing analytical precision and generating spurious interpretations.
4. E4: Translation or Meaning Error. The AI attributes an improper meaning to a term or passage, generating a semantic distortion that can alter historical or contextual understanding. This type of error manifests particularly in multilingual corpora.

5. E5: Evasive or Generic Response. The output does not respond precisely to the prompt but produces a stereotypical or non-specific text. This behavior reflects a deficit in relevance, where the form of the discourse prevails over the informational substance.
6. E6: Erroneous Attribution. This concerns the assignment of information to a wrong author, entity, or subject. This error affects the reconstruction of authority and responsibility relationships, compromising the historical reliability of the output.
7. E7: Temporal Shift. The AI places events or processes in a chronological period different from the documented one, producing anachronisms or chronological distortions that alter the historical sequence.
8. E8: Overinterpretation. This occurs when the AI draws excessive inferences, attributing intentions, meanings, or consequences that are not grounded in the source. This error improperly amplifies the content, creating hypothetical narratives masked as facts.
9. E9: Internal Contradiction. Within the same output, incompatible or inconsistent statements are found. Such contradictions reduce the logical consistency of the text and compromise its epistemic reliability.
10. E10: Distortion of Hierarchy or Weighting. The AI alters the relevance relationship between pieces of information, emphasizing marginal elements and minimizing central ones. This error compromises not so much the truthfulness of the content, but its balanced representation.
11. E11: Other (including systemic biases). Residual category that includes errors not classifiable elsewhere, such as the projection of implicit cultural, linguistic, or ideological biases. Such distortions often manifest in the form of Eurocentric, androcentric perspectives or through the unconscious adoption of anachronistic conceptual categories.

4. Results

4.1. OCR noise distribution

In the first half of the experimental sample, clear differences emerged between the two versions of GPT, namely ChatGPT-4o and GPT-5.0 Thinking.

- GPT-5 Thinking uses a richer and more technical vocabulary, although it can sometimes lack specialization. It is particularly strict in its self-evaluations, often flagging minor issues. In two instances, it did not utilize the text present in the documents to infer contextual information.
- ChatGPT-4o provides more general responses, which are less precise but overall correct, adopting a more narrative and explanatory style.

In summary, GPT-5 Thinking reasons in a more detailed and structured manner, while GPT-4o prioritizes speed and generality. However, the overall accuracy in detecting and classifying elements remains comparable.

4.2. Final Comparative Results: ChatGPT-4o vs GPT-5 Thinking

Building on these preliminary observations, the final comparative analysis highlights systematic stylistic and structural differences between the two models (see Table 2).

Table 2

Stylistic and Structural Differences

Building on these preliminary observations, the final comparative analysis highlights systematic stylistic and structural differences between the two models (see Table 2).

Aspect	GPT-5 Thinking	Chat GPT-4o
Response Time	Longer (≈ 15 sec), with explicit mention of the delay ("Thinking for 15 seconds...").	Immediate, with no time indication.
Style	Brief responses, often in list form or independent sentences, technical tone	Long, narrative responses, structured in paragraphs (introduction, body, conclusion).
Organisation	Direct answers, without initial rephrasing.	Contextualized answers, often rephrasing the question before addressing it.

As shown in Table 2, the table summarizes the main stylistic and structural differences observed between the two models. GPT-5 Thinking adopts a concise and analytical tone, whereas ChatGPT-4o tends toward narrative elaboration and contextual expansion, reflecting two distinct generative strategies.

General GPT-5 e GPT-4

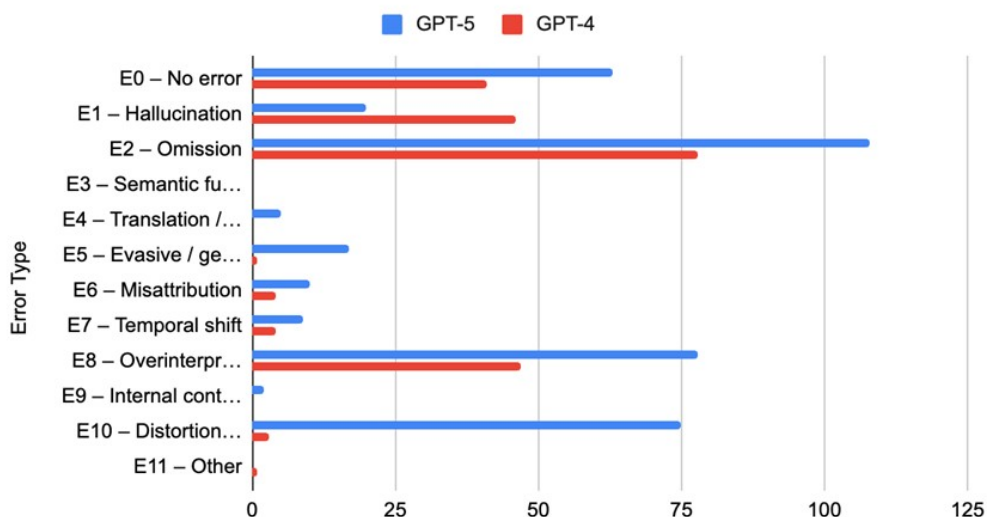


Figure 1: General comparison between ChatGPT-5 thinking and ChatGPT-4o (by error type)

The figure presents a comprehensive comparison of epistemic error frequencies between GPT-5 Thinking and ChatGPT-4o across categories E0-E11. Overall, GPT-5 displays a higher proportion of error-free outputs (E0), indicating improved factual accuracy and reliability. It shows a slight decrease in Hallucinations (E1) compared to GPT-4, but a notable increase in Omissions (E2), Overinterpretations (E8), and Distortions of Hierarchy or Weighting (E10). This pattern suggests

that GPT-5 tends to favor analytical caution and structural precision, yet sometimes sacrifices balance between accuracy and completeness. In contrast, GPT-4 produces fewer Omissions (E2) and Distortions of Hierarchy or Weighting (E10) but exhibits more Hallucinations (E1) and minor Semantic Fusion (E3) errors, reflecting a more narrative and context-expansive style that occasionally undermines factual rigor. Minor error types, such as misAttribution (E6) and Temporal Shift (E7), remain rare in both models. Overall, the comparison indicates that GPT-5 Thinking adopts a more analytical and structurally disciplined approach, whereas ChatGPT-4o favors narrative flexibility and contextual elaboration, highlighting two distinct generative strategies.

Per-Question Analysis

Figure 3, Document Typology presents a comparison of epistemic error types for the first evaluation question (Q1). The figure shows that GPT-5 Thinking achieved a higher number of correct classifications (E0) than GPT-4o, confirming its superior accuracy in identifying document typology. GPT-5 Thinking also produced fewer Hallucinations (E1) and Semantic Fusion (E3) errors, indicating greater factual consistency.

In contrast, GPT-4o generated a much higher number of Overinterpretations (E8), revealing its tendency toward elaborated and sometimes speculative analyses. However, GPT-5 Thinking occasionally introduced Distortion of Hierarchy (E10) errors, a phenomenon largely absent in GPT-4o’s responses. Minor categories such as Omission (E2) and Temporal Shift (E7) remained negligible for both models.

Overall, the figure highlights a clear trade-off: GPT-5 Thinking offers more precise and reliable classifications, while GPT-4o tends to produce longer, interpretive outputs more prone to semantic exaggeration.

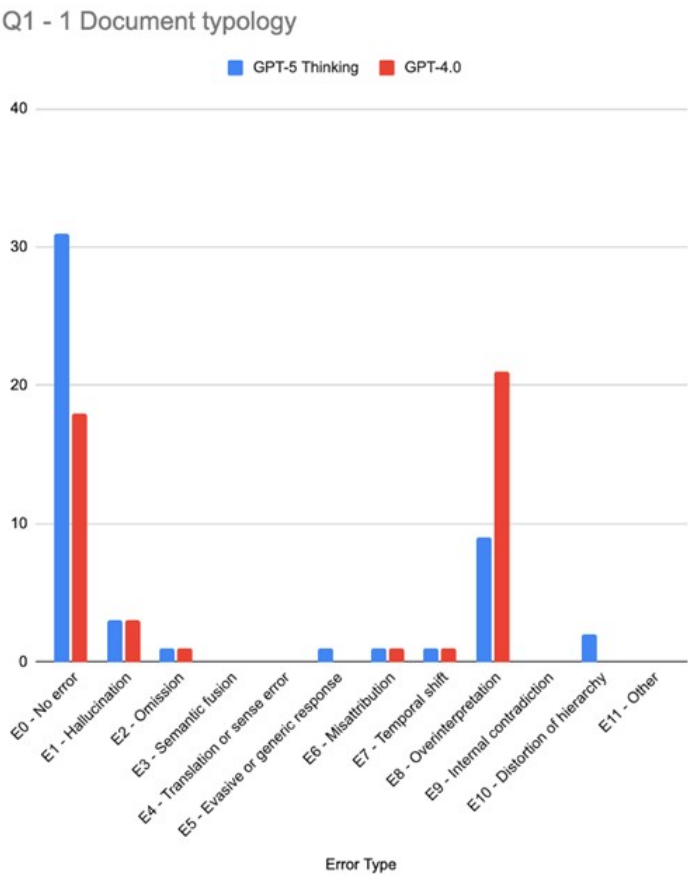


Figure 2: Document Typology

Figure 4 Enucleation of Main and Secondary Information compares the distribution of epistemic error types for the second evaluation question (Q2). The figure shows that GPT-5 Thinking achieved a higher number of correct responses (E0) than GPT-4o, confirming its superior accuracy in distinguishing between main and secondary information. GPT-5 also displayed a moderate number of Omissions (E2) and some Distortions of Hierarchy (E10), suggesting that its focus on structural precision may occasionally lead to partial loss of secondary details.

In contrast, GPT-4o exhibited more frequent Omissions (E2) as well as higher rates of Hallucinations (E1) and Overinterpretations (E8), reflecting its tendency toward more expansive and sometimes speculative elaboration. Minor categories such as Semantic Fusion (E3) and Misattribution (E6) remained infrequent for both models.

Overall, Figure 4 highlights a clear methodological divergence: GPT-5 Thinking prioritizes accuracy and structural clarity, whereas GPT-4o leans toward contextual expansion and interpretive depth, representing two distinct generative strategies for extracting primary and secondary information.

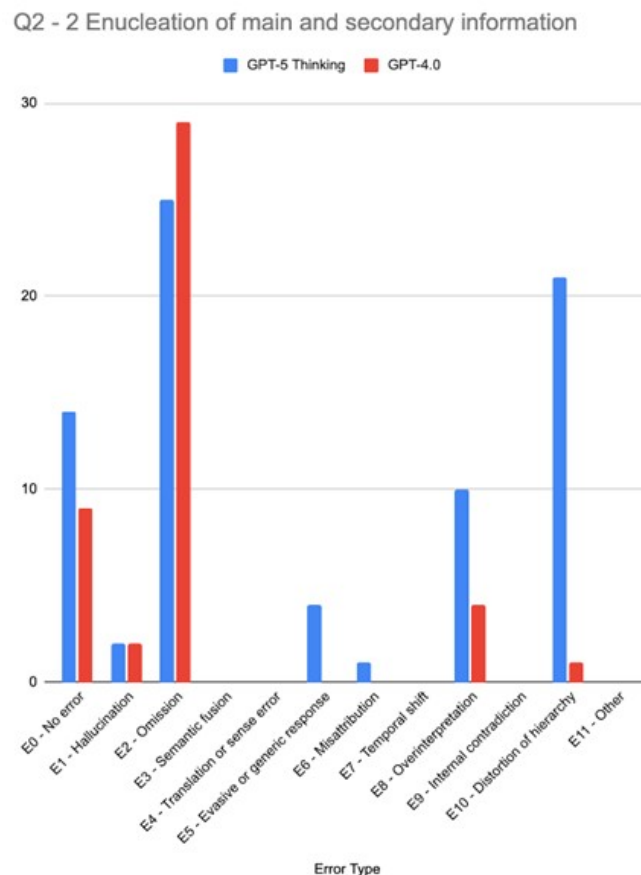


Figure 3: Enucleation of main and secondary information

Figure 5 Taxonomic List of Entities compares the distribution of epistemic error types for the third evaluation question (Q3). The figure shows that GPT-5 Thinking achieved a higher number of correct results (E0) and fewer Hallucinations (E1) than GPT-4o, confirming its stronger factual reliability. However, GPT-5 also displayed more Omissions (E2), slightly more Overinterpretations

(E8), and a significant increase in Distortions of Hierarchy (E10), indicating that its analytical caution sometimes caused incomplete or misstructured classifications.

In contrast, GPT-4o produced fewer correct results (E0) and more Hallucinations (E1), revealing a tendency toward generating additional but occasionally inaccurate content. While GPT-4o also showed some Omissions (E2), their frequency was lower than in GPT-5, and it exhibited very few E8 or E10 errors. Both models showed only minimal traces of Semantic Fusion (E3) or other secondary error types.

Overall, Figure 5 highlights the balance between accuracy and completeness: GPT-5 Thinking prioritizes factual precision and structural discipline, whereas GPT-4o favors broader coverage and generative fluency, even at the cost of factual stability.

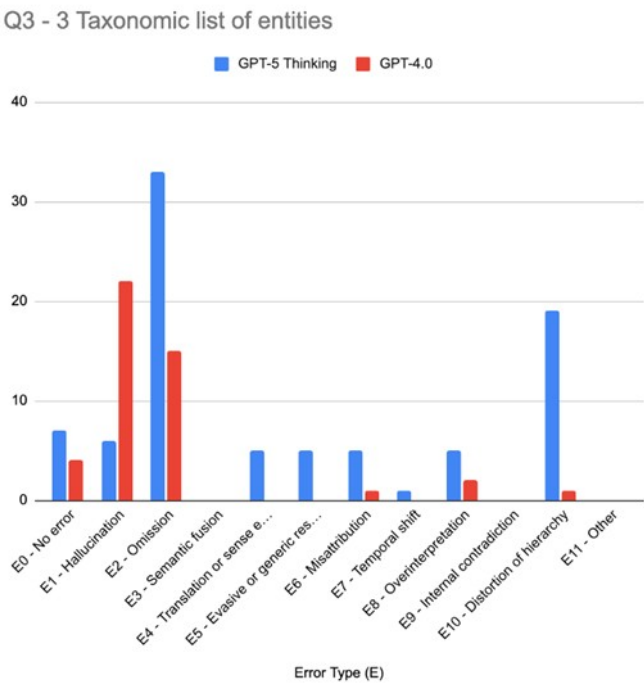


Figure 4: Taxonomic list of entities

Figure 6 Historical Context compares the distribution of epistemic error types for the fourth evaluation question (Q4). The figure shows that GPT-5 Thinking, while achieving a higher number of correct classifications (E0) than GPT-4o, also produced many Omissions (E2) and a very high rate of Overinterpretation (E8), the dominant error category for this question. This pattern indicates that GPT-5 often inferred causal or contextual links that were not explicitly supported by the documents. It also displayed some Temporal Shifts (E7) and Distortions of Hierarchy (E10), reflecting moments where its analytical precision led to speculative or disproportionate reasoning.

In contrast, GPT-4o generated fewer Omissions (E2) and less Overinterpretation (E8), but showed more Hallucinations (E1) and occasionally introduced invented or inaccurate historical elements.

Overall, Figure 6 highlights a clear trade-off between depth and factual neutrality: GPT-5 Thinking tends to favor analytical depth and causal inference, which can result in excessive interpretation, while GPT-4o remains more moderate but less systematically accurate in reconstructing historical context.

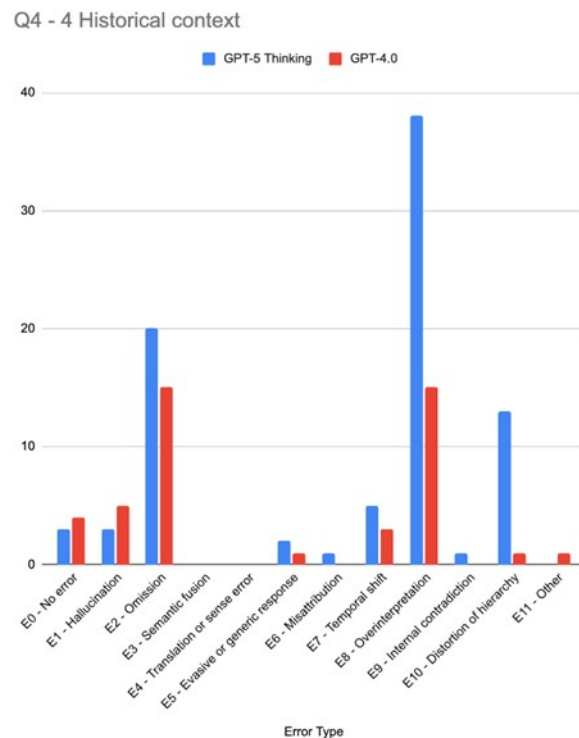


Figure 5: Historical context

Figure 7 Archival Descriptive Note compares the distribution of epistemic error types for the fifth evaluation question (Q5), which required both models to produce a concise descriptive summary of archival materials. The figure shows that GPT-5 Thinking achieved a higher number of accurate results (E0) and fewer Hallucinations (E1) than GPT-4o, confirming its stronger factual consistency. However, GPT-5 also displayed a very high number of Omissions (E2), several Evasive or Generic Responses (E5), and a notable increase in Distortion of Hierarchy (E10). It also produced slightly more Overinterpretations (E8), indicating that while GPT-5's summaries were precise and factually sound, they often lacked contextual breadth and balance.

In contrast, GPT-4o generated fewer Omissions (E2) but more Hallucinations (E1), producing broader yet less reliable descriptions. Both models showed negligible Semantic Fusion (E3) and similar levels of Misattribution (E6).

Overall, Figure 7 highlights the enduring trade-off between factual precision and completeness: GPT-5 Thinking favors accuracy, structure, and analytical control, whereas GPT-4o prioritizes narrative fluency and contextual inclusiveness, often at the cost of factual rigor.

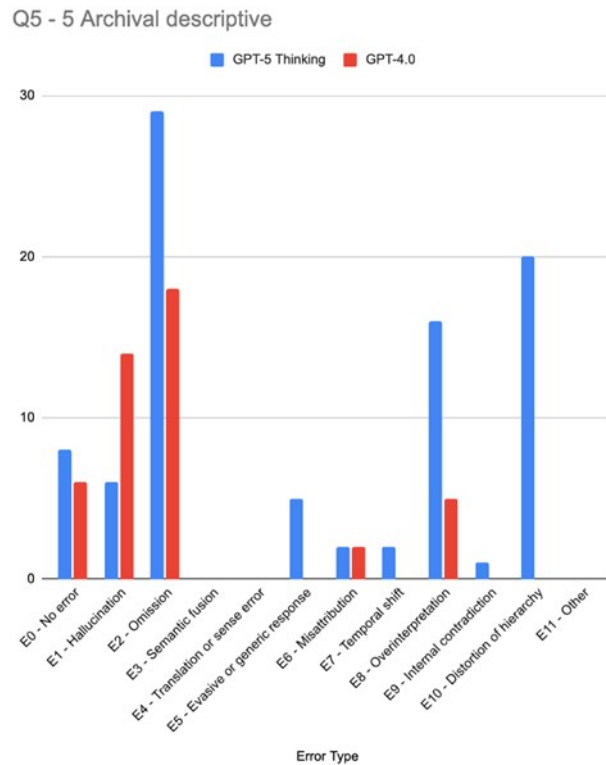


Figure 6: Archival descriptive

Error Assessment and Frequency

The quantified results by error type show distinct performance profiles for each question. In Q1 (Document Typology), GPT-5 recorded more correct answers (31 E0 vs. 18 for GPT-4) but exhibited 2 cases of hierarchy distortion (E10). GPT-4, in contrast, committed substantially more over-interpretations (21 cases of E8). For Q2 (Primary/Secondary Information), GPT-5 again achieved more correct answers (14 vs. 9) but was prone to significantly more hierarchy distortions (21 cases E10 vs. 1 for GPT-4). Omissions (E2) were a frequent issue for both models in this task.

The analysis of Q3 (Taxonomic List) reveals a telling contrast: GPT-4 hallucinated content much more often (22 E1 vs. 6 for GPT-5), but GPT-5 more frequently omitted entities altogether (33 E2 vs. 15). GPT-5 also presented more attribution and hierarchy distortion errors in its lists. Q4 (Historical Context) proved particularly challenging for GPT-5, which engaged in significant over-interpretation (38 cases E8 vs. 15 for GPT-4) and committed more temporal and hierarchy distortions. GPT-4 achieved a somewhat better, though still limited, balance for this question. Finally, for Q5 (Descriptive Note), GPT-4 hallucinated more (14 vs. 6), but GPT-5 presented a greater number of omissions (29 E2) and hierarchy distortions (20 E10).

Control Group Analysis

The control corpus (11 documents, 25 % of the dataset) was used to observe model behavior on out-of-domain material unrelated to the historical-institutional context. Both GPT-4o and GPT-5 Thinking displayed similar overall error patterns to those found in the main corpus, confirming that epistemic distortions were primarily linked to language and OCR quality rather than to topical relevance. GPT-5 Thinking maintained a higher proportion of correct classifications (E0) and fewer hallucinations (E1), even on non-diplomatic content, but continued to show frequent omissions (E2). GPT-4o, conversely, generated richer but more speculative outputs with slightly higher rates

of hallucination and over-interpretation (E8). The absence of significant divergence between in-domain and out-of-domain texts suggests that both models' performance depends more on linguistic and structural noise than on subject-matter familiarity. This finding supports the robustness of the error taxonomy across heterogeneous corpora.

5. Discussion

The results of this comparative study reveal a fundamental dilemma in integrating large language models into archival analysis: improving factual reliability often comes at the cost of interpretative completeness. GPT-5 Thinking, while more consistent and structurally rigorous than ChatGPT-4o, favors analytical control over contextual richness. Its reduction in hallucinations and formal inconsistencies is counterbalanced by increased omissions and distortions of hierarchy, showing that precision does not necessarily guarantee depth. This evolution places archivists and researchers before a crucial methodological choice: should one rely on a system that invents less but omits more, or the reverse?

This transition from GPT-4o to GPT-5 Thinking reflects two distinct generative logics. GPT-5 reasons in a segmented and deliberate manner, producing concise, technical outputs governed by internal analytical rules. GPT-4, by contrast, operates through a narrative mode of explanation more fluent and discursive, but also more prone to speculation and contextual overreach. The trade-off thus shifts from fluency to restraint, from interpretative generosity to structural discipline.

At a conceptual level, these findings challenge the presumed neutrality of large language models. GPT-5's tendency to reorganize information according to its own internal logic reveals a subtle epistemic bias: rather than simply summarizing, the model constructs a coherent, but sometimes artificial, narrative, amplifying marginal details or inferring unsupported causal links. This phenomenon suggests that LLMs are not neutral instruments but rhetorical engines capable of producing persuasive discourse that can mask factual gaps or interpretative imbalance. Consequently, critical human supervision remains essential to identify and correct such narrative reconstructions.

Furthermore, the persistence of translation and semantic errors in the multilingual, OCR-affected corpus demonstrates that technical progress has not yet resolved the challenges of noisy or heterogeneous data. GPT-5 occasionally produced more translation and meaning errors than GPT-4, revealing vulnerability when confronted with degraded or linguistically complex sources, a common feature of digitized archival collections. Future model development must therefore focus on improving robustness to such imperfections.

Ultimately, these results advocate for a measured and hybrid integration of LLMs in archival workflows. Their strengths in preliminary classification, entity extraction, and synthesis generation are undeniable and can accelerate processing. Yet the systemic errors identified, omissions, structural imbalances, and over-interpretations, make their unsupervised use in interpretative or historical analysis problematic. The sustainable future of such technologies lies in the creation of standardized validation protocols, shared methodological guidelines, and specialized training programs, ensuring that human expertise remains central to evaluation, interpretation, and accountability.

6. Conclusion and Future Perspectives

This exploratory study demonstrates that the evolution of language models, embodied by the transition from GPT-4 to GPT-5, is not unidirectional. While GPT-5 represents a tangible advance in factual reliability by reducing hallucinations, it simultaneously introduces new challenges, notably reduced comprehensiveness and instability in the hierarchical representation of information. For historical research and archival sciences, this means these tools can serve as valuable assistants in intermediate phases of sorting and description, but they cannot replace the critical acumen and contextual understanding of the archivist or historian.

Future research perspectives should be oriented along several axes. Firstly, diversifying the corpora, to include other languages, time periods, and document types, would allow for testing the robustness of these conclusions. Secondly, concrete experiments integrating LLMs into cataloging or critical edition projects, using well-defined hybrid protocols, would be necessary to assess their operational added value. Finally, the development and refinement of error taxonomies, such as the one proposed here, constitute a promising methodological path for standardizing the evaluation of LLM performance in the field of digital humanities.

Limitations

It is important to consider the limitations of this research. The preliminary nature of this study is based on a restricted corpus of 45 documents, which limits the general statistical scope of the results. The evaluation of the outputs was performed by a single coder and not in a double-blind process, a method that could introduce subjective biases, although mitigated using an explicit error taxonomy. Since the outputs were annotated by a single evaluator, inter-annotator reliability indices (e.g., Cohen's κ) could not be computed, which limits the statistical validation of the results. Finally, the variable quality of the Optical Character Recognition (OCR) of the source documents is a confounding factor that may have exacerbated certain model errors, particularly those related to nuanced understanding and the omission of information.

Declaration on Generative AI

During the preparation of this work, the author used Chat-GPT-5.0 in order to: Grammar and spelling check. The author take full responsibility for the publication's content.

Availability of Data and Code

The complete results table will be made available in a public repository after acceptance.

Acknowledgements

The author thanks the Historical Laboratory of (Electro)Technology, FEE, CTU in Prague, and of SVK26/25/F3, CTU in Prague, for its support of the project.

Conflicts of Interest

The author declares no conflicts of interest.

Methods – Model Specifications

We evaluated GPT-4o and GPT-5 Thinking via the ChatGPT interface (September 2025). Terminology note (models). Throughout the paper we use the names GPT-4o and GPT-5 Thinking. Some reviews and external documents refer to the same systems with alternative labels (e.g., “GPT-4”, “GPT-4 (Legacy)” for GPT-4o, and “GPT-5”, “GPT-5.0”, “GPT-5 (Thinking)” for GPT-5

Thinking). To avoid ambiguity, we fix the following correspondence: GPT-4o = “GPT-4/Legacy”, GPT-5 Thinking = “GPT-5/5.0/Thinking.” Unless otherwise specified, “GPT-5 Thinking” denotes the Thinking mode available in the ChatGPT interface at the time of our experiments (September 2025). We harmonize these names across the text, tables, and figures.

Each document was submitted in an isolated chat, the five questions (Q1-Q5) were asked in a fixed order, without system prompts or prompt-engineering. Platform defaults applied (no explicit control over temperature or length). Outputs were coded with the E0-E10 taxonomy.

The quantitative values reported in Tables and Figures correspond to the manual sum of occurrences of each epistemic error type (E1-E11) across the full corpus of 45 documents. No averaging or normalization was applied, since the aim was to compare total error frequencies between models rather than per-document means.

References

- [1] Jaillant, L., & Rees, A. (2022). Applying AI to digital archives: Trust, collaboration and shared professional ethics. Digital Scholarship in the Humanities. Advance online publication.
- [2] Alalaq, Ahmed Shaker. “The History of the Artificial Intelligence Revolution and the Nature of Generative AI Work.” *DS Journal of Artificial Intelligence and Robotics* 2, no. 4 (2024) <https://doi.org/10.59232/AIR-V2I4P101>.
- [3] Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An overview of current debates and future perspectives. *Journal on Computing and Cultural Heritage*, 15 (1), 4:1 – 4:15. <https://doi.org/10.1145/3479010>
- [4] Shinde, G., Kirstein, T., Ghosh, S., & Franks, P. C. (2024). AI in Archival Science – A Systematic Review. <https://doi.org/10.48550/arXiv.2410.09086>
- [5] Groppe, J., Marquet, A., Walz, A., & Groppe, S. (2025). Automated Archival Descriptions with Federated Intelligence of LLMs. arXiv. <https://doi.org/10.48550/arXiv.2504.05711>
- [6] Bourne, J. (2024). CLOCR-C: Context Leveraging OCR Correction with Pre-trained Language Models. <https://arxiv.org/abs/2408.17428>
- [7] Spina, S. (2023). Artificial Intelligence in archival and historical scholarship workflow: HTR and ChatGPT. <https://doi.org/10.48550/arXiv.2308.02044>
- [8] Trace, C. B. (2022). Archival infrastructure and the information backlog. *Archival Science*, 22(1), 75-93. <https://doi.org/10.1007/s10502-021-09368-x>
- [9] Bennani A. Histoire et perspectives de l’intelligence artificielle. *Med Sci (Paris)*. 2024;40(3):283-286. doi:10.1051/medsci/2024022
- [10] Jensen, B., Reynolds, I., Atalan, Y., Garcia, M., Woo, A., Chen, A., & Howarth, T. (2025). Critical Foreign Policy Decisions (CFPD)-Benchmark: Measuring Diplomatic Preferences in Large Language Models. arXiv. <https://doi.org/10.48550/arXiv.2503.06263>
- [11] Mendonça, W. P. “Intelligence artificielle et signification. À propos des limites et des possibilités des sciences cognitives.” *Philosophiques* 17.1 (1990).
- [12] Liang, Y., Yang, L., Wang, C., Xia, C., Meng, R., Xu, X., Payani, A., & Shu, K. (2025). Benchmarking LLMs for Political Science: A United Nations Perspective (UNBench). arXiv.
- [13] Buchanan, Bruce G. “A (Very) Brief History of Artificial Intelligence.” *AI Magazine* 26, no. 4 (2005): 53. <https://doi.org/10.1609/aimag.v26i4.1848>.
- [14] Mayr, M., Neumeier, K., Krenz, J., Bürcky, S., Kordon, F., Seuret, M., Zöllner, J., Wu, F., Maier, A., & Christlein, V. (2025). Data-efficient handwritten text recognition of diplomatic historical text. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-024-20545-9>
- [15] OpenAI. (2025, August 7). Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/>

- [16] ukhari, S. S., Kadi, A., Jouneh, M. A., Mir, F. M., & Dengel, A. (2017). anyOCR: An open-source OCR system for historical archives. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (pp. 118–123). IEEE. <https://doi.org/10.1109/ICDAR.2017.58>

A. Archive:

Archives nationales de Prague : Fond : Ministerstvo školství ČSR (1970-1990).