Building Historical Corpora with Multimodal LLMs: Epistemic Gaps and Misreadings in 18th-Century Russian Books

Maria Levchenko[1]

[1] Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Abstract

Large language models offer transformative potential for digitizing historical texts, but their application to humanities research raises critical questions about temporal bias and historical representation. We present the first systematic evaluation of multimodal LLMs for historical optical character recognition (OCR), testing 11 leading models on 1,030 pages of 18th-century Russian texts printed in Civil font. Using a contamination-free dataset from the National Library of Russia, we demonstrate that while LLMs substantially outperform traditional OCR systems (achieving 3.36% vs. 21.55–45.96% character error rates), they exhibit systematic temporal biases that fundamentally compromise historical authenticity.

Our analysis reveals two distinct forms of distortion: a "modernization trap" where models automatically "correct" historical orthography to contemporary standards, and paradoxical "over-historicization" where models insert anachronistic medieval Slavonic characters into 18th-century texts. These errors reflect what we term the absence of "historical linguistic competence" — models treat historical language not as a continuum of specific periods but as an undifferentiated space labeled "old". Different model families exhibit distinct error signatures, exposing how architectural choices and training data composition shape temporal bias.

These findings reveal that "epistemic anachronism" in AI systems goes beyond inherited editorial biases. While training data explains modernization, the concurrent archaization demonstrates a fundamental architectural limitation: without temporal metadata as a training signal, models cannot develop "historical linguistic competence" even when explicitly provided with dates. Our work shows how these systems create temporal chimeras that appear historical while actively corrupting the historical record.

Keywords: Large Language Models, historical OCR, temporal bias, 18th-century Russian, Digital humanities, epistemic anachronism, Civil font, historical orthography

1 Introduction. The Historical Linguistic Competence Problem

Large language models offer transformative potential for digitizing and analysing historical texts on a large scale. However, their application to humanities research requires an understanding of how these systems model the historical or cultural past, and whether they perpetuate biases embedded in their predominantly modern training data. Even in ostensibly straightforward tasks like optical character recognition, these biases profoundly shape outcomes.

This study provides quantifiable evidence of temporal bias through systematic evaluation of 11 leading multimodal LLMs on 1,030 pages of 18th-century Russian texts. Our analysis reveals two distinct forms of distortion. First, as expected, models inherit and amplify the modernizing

Maria Levchenko. "Building Historical Corpora with Multimodal LLMs: Epistemic Gaps and Misreadings in 18th-Century Russian Books." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 725–737. https://doi.org/10.63744/SKoZVUHQbtE7.

tendencies of their training data — automatically 'correcting' historical orthography to contemporary standards. More surprisingly, LLMs demonstrate no genuine chronological understanding: they treat historical language not as a continuum of specific periods but as an undifferentiated space labeled 'old,' for example, anachronistically inserting medieval Slavonic characters into 18th-century Civil font texts. This finding supports Zaagsma's (2023) call for 'critical examination of the methodological and epistemological consequences of digital tools' [18] with constructive empirical evidence: understanding precisely how LLMs misrepresent linguistic history — conflating periods and neighboring traditions while lacking temporal reasoning — is essential for informed and responsible deployment of these tools in historical research.

Digitization has never been a neutral act of copying. Just as physical archives embody conscious and unconscious acts of selection and curation [14], LLMs inherit the biases of their training data. In the case of historical Russian texts, a century-long editorial tradition of modernizing pre-revolutionary orthography for accessibility has created a training environment that renders LLMs fundamentally unreliable for historical work. When an LLM "corrects" *дѣло* to *дело* (ѣ→e, *yat* to *e*) and *миръ* to *мир* (removal of the terminal hard sign, ъ), it performs not transcription but interpretation — erasing historical evidence and replacing it with modern construction. This single character substitution or omission, repeated across millions of instances, exemplifies how models enforce a specific, modernized archival paradigm rather than preserving diplomatic authenticity.

Our documented patterns of OCR failure serve as a diagnostic tool, demonstrating not just technical and architectural limitations but the underlying cognitive and data-driven bias: absence of historical linguistic competence in current AI systems. By providing micro-level analysis of how diachronic change impacts NLP performance, we demonstrate that orthographic evolution poses a systematic challenge to AI development. Our findings thus bridge critical debates about digital representation with practical insights into how temporal biases become computationally encoded, offering both theoretical understanding and empirical groundwork for future development of truly historically-aware AI systems.

2 Cultural and Temporal Biases in LLMs: Knowledge Gaps in Historical Texts

The rapid development of LLMs and their applications across diverse fields has prompted extensive investigation into their biases. This research track has undergone significant conceptual evolution, shifting from viewing bias as a purely technical problem to understanding it as a complex socio-technical phenomenon. This shift is driven by the growing recognition that a truly "objective" or "bias-free" model is an illusion [12]. Because any norm against which bias is measured is itself culturally and ideologically relative, the goal is moving away from bias elimination and toward principled management, transparency, and context-specific regulation of a model's inherent biases.

Current research has identified three primary sources of bias in LLMs, each contributing to the model's final behavior [5]. *Data-induced biases* originate from the training corpora, including representational imbalances and modernization practices in digitized texts. *Model-induced biases* emerge from architectural design choices, such as transformer attention mechanisms that create systematic processing preferences. *Alignment-induced biases* paradoxically arise from safety interventions designed to make models more aligned with human values. These sources are interconnected, creating cascade effects where biases introduced at one stage are propagated and amplified in subsequent stages.

These biases manifest across multiple dimensions, with cultural and temporal biases being particularly relevant to historical text processing. Cultural bias, a form of representational harm, manifests through the over-representation of Western-centric content in training corpora. This results in models that prioritize Western cultural contexts while misrepresenting or erasing non-Western cultures, norms, and values [13]. Underwood (2025) observes that models trained on contemporary internet data 'mirror the languages and nationalities best represented there' [15],

while Boelaert et al. (2025) argue that LLMs exhibit 'machine biases' that 'flatten out human population diversity' [1]. Recent research suggest that attempts to encode cultural knowledge often result in models learning 'stereotypes' rather than genuine understanding [2; 13]. In historical settings, this tendency resonates with Coeckelbergh's concept of *epistemic anachronism*, where contemporary norms are imported into past contexts [3].

Temporal bias in LLMs can be effectively investigated through historical document recognition tasks, which provide not only concrete, easily quantifiable metrics but also clear error analysis revealing how different model architectures and alignment approaches distinctly affect historical text processing. Temporal bias — where models lack historical linguistic competence and impose contemporary norms on historical texts — exemplifies all three bias sources: data-induced bias through modernized training corpora without temporal contextual information, model-induced bias through architectural limitations in temporal reasoning, and alignment-induced bias through superficial historical awareness that masks deeper temporal incompetence. The systematic errors in historical OCR thus reflect not isolated technical problems but fundamental limitations in how current LLMs model temporal and cultural variation, which may stem from the dominant "Probabilistic Language Modeling-based Paradigm" that primarily encodes statistical relationships from data rather than a causal understanding of the world.

3 Peter the Great's Civil Script Reform and Its Digital Legacy

The entire canon of Russian literature exists in a temporal orthographic capsule, bounded by two autocratic reforms. Between Peter the Great's 1708–1710 reform and the Bolsheviks' 1918 reform lies the complete textual heritage of Imperial Russia — Lomonosov, Karamzin, Pushkin, Gogol, Turgenev, Dostoevsky, Tolstoy, Chekhov — all conceived, written, and printed in a specific orthographic form that no longer circulates. This 210-year period produced Russian literature's Golden and Silver Ages, yet we have systematically erased its authentic textual form, replacing it with post-1918 modernizations. What readers encounter today as "Tolstoy" or "Pushkin" are orthographic translations, not the texts these authors actually wrote.

Peter fundamentally reshaped Russian textual culture by creating a permanent schism between sacred and secular scripts. Prior to 1708, Russian texts used Church Slavonic script with approximately 45 letters and complex diacritical marks. Peter's new Civil font (*Grazhdansky Shrift*) reduced this to 38 characters, eliminating redundant Greek letters (ѧ, ѯ, ѿ), introducing new forms (Я, Э), and adopting Arabic numerals and Western punctuation.

Throughout the 18th century, the St. Petersburg Academy of Sciences systematically refined the Civil font. The 19th century marked the final stabilization of Cyrillic letterforms — the forms in which classic Russian literature was printed. Then came the second boundary: the 1918 reform that would retroactively transform this entire textual legacy. Peter's autocratic reform method—imposing change against traditionalist opposition—set a precedent the Bolsheviks would echo in their 1918 reform. They completed Peter's rationalization, eliminating four letters (ѣ, ѳ , i, ѵ) and terminal hard signs (ъ). Table 1 shows the key transformations that created distinct temporal boundaries in Russian orthography. Despite its academic origins and pragmatic goals (the preparation of this simplification was initiated a decade before the October Revolution), the reform's implementation by the Bolsheviks made it intensely political. For the large community of Russian émigrés who had fled the revolution, the old orthography became a powerful symbol of their identity and their opposition to the new regime. White Army propaganda and émigré publications across the globe defiantly continued to use the pre-revolutionary spelling, with its "superfluous" letters, as a tangible link to the culture of Imperial Russia. The Nobel laureate Ivan Bunin swore he would "never accept the Bolshevik orthography". The choice of spelling was no longer a matter of correctness, but a political shibboleth.

The 1918 reform didn't just change future writing — it reached backward to rewrite the past.

Historical Form	Name	Modern Equivalent	Example
Ѣ, ѣ	yat	Е, е	повѣсть → повесть
І, і	decimal i	И, и	лінія → линия
Ѳ, ѳ	fita	Ф, ф	ор ѳографія → орфография
Ѵ, ѵ	izhitsa	И, и	сѵнодъ → синод
Ъ, ъ	hard sign	(omitted)	столъ → стол
з-/с-	prefix alternation	з-/с-	безполезный → бесполезный
-аго/-яго	adj. ending	-ого/-его	святаго → святого
-ыя/-ія	adj. ending	-ые/-ие	новыя книги → новые книги

Table 1: The 1918 orthographic boundary: Imperial-era characters and forms with their modern equivalents

Throughout the 20th century, standard editorial practice systematically "translated" all pre-1918 texts into contemporary orthography: canonical literature, documents, and scholarly works. Digital corpora inherited this convention: the Russian National Corpus systematically normalizes historical texts to modern orthography, prioritizing computational functionality over textual fidelity.

This modernization tradition creates profound consequences for AI systems. LLMs trained on vast digital corpora have almost exclusively ingested modernized versions of Russian historical texts, lacking exposure to authentic pre-reform orthography. Without temporal metadata to signal linguistic shifts, these models exhibit a fundamental lack of historical linguistic competence. They treat Russian as a monolithic, synchronic entity conforming to post-1918 rules rather than understanding it as a language that evolved over time. Our research reveals that LLMs do not recognize ѣ in 18th-century texts as historically authentic — instead treating it as an error to be "corrected" to modern e, anachronistically enforcing 20th-century Bolshevik policy on Petrine-era texts.

4 Dataset description and OCR LLM results

Dataset. We created a novel evaluation corpus of 1,030 scanned pages from 428 unique 18th-century Russian books (1750–1800) printed in Civil font, sourced from the National Library of Russia's limited-access collection. The dataset was specifically designed to prevent training data contamination using materials never previously transcribed or published online during known LLM pretraining periods. Images with resolution below 150ppi were excluded following evidence of poor LLM OCR performance at low resolution [7]. The corpus was stratified by publication period, text density, decorative elements, and subject matter (fiction, science, religion) to ensure diversity.

Ground truth was established through a multi-stage process: automated layout analysis using fine-tuned YOLOv8, initial OCR with a TrOCR model fine-tuned on 15,914 lines (575 pages) from our dataset[1] (achieving 1.83% CER on held-out test pages), and 100% manual correction following diplomatic transcription principles (i.e., preserving historical orthography, punctuation, and lineation, with no silent modernization). The 100-page sample of the resulting corpus containing 2,954 lines and 15,970 words is published online alongside the LLM-based OCR results.[2]

Baseline Performance. To quantify the challenges facing traditional OCR systems, we tested Tesseract, Surya OCR (BT5), and a specialized Transkribus model on the same 100-page sample. Results showed severe performance degradation: Surya achieved 45.96% CER, Transkribus 26.93% CER, and Tesseract 21.55% CER, compared to our fine-tuned TrOCR's 1.83% CER — demonstrating the need for specialized approaches.

[1] Available at https://huggingface.co/taiga75/ru-trocr-1700s.

[2] Viewer https://mary-lev.github.io/historical-ocr-analysis/ contains ground truth transcriptions and model outputs.

Experimental Design. We systematically evaluated 11 leading multimodal LLMs across three recognition modes: single-line processing (minimal context), full-page processing (maximum context), and sliding-window processing (3-line context windows). Models included GPT-4 variants, Claude 3.5/3.7, Gemini 2.0/2.5 series, Qwen 2.5-VL, and Llama-4 variants. Performance was assessed using standard OCR metrics (CER, WER), case-insensitive character error (CI-CER), and historical-character error (HChE; incorrect handling of period-specific glyphs).

We conducted controlled prompt engineering experiments comparing basic English prompts, context-enhanced English prompts with historical metadata, and context-enhanced Russian prompts. Model stability was evaluated through repeated testing over seven consecutive days, and sensitivity analysis quantified performance degradation across 17 document features including layout complexity, line density, and historical character frequency. Detailed performance analysis is available in [8].

Key Results. LLMs substantially outperformed traditional OCR systems, with the best-performing model (Gemini 2.5-Pro) achieving 3.36% CER and 4.69% WER compared to 21.55–45.96% CER for traditional systems. Full-page processing generally yielded optimal results (see Table 2).

It is important to note that our fine-tuned TrOCR model achieved superior performance at 1.83% CER, demonstrating that specialized models with access to 575 pages of manually corrected training data can outperform generalist LLMs. However, creating such training data requires significant time and resources for annotation and correction. When such investment is not feasible, multimodal LLMs provide a practical alternative for historical OCR, achieving reasonable accuracy without any task-specific training — though at the cost of the temporal biases we analyze below.

Model	CER (%)	WER (%)	CI-CER (%)	HChE (%)
Gemini-2.5-Pro	3.36 (0.14–20.95)	4.69 (0.08–31.43)	3.19	9.83
Gemini-2.5-Flash	4.94 (0.75–22.11)	6.70 (0.41–22.82)	4.81	12.86
Qwen-2.5-VL	5.81 (0.81–86.86)	7.48 (0.99–90.14)	5.54	16.40
Gemini-2.0	6.14 (1.51–22.16)	10.33 (1.58–30.55)	5.66	32.00
Claude-3.5	6.79 (0.70–53.96)	8.46 (0.00–51.09)	5.75	15.24
OpenAI-o4-mini	6.87 (2.18–57.54)	9.07 (2.37–58.85)	6.76	18.38
Claude-3.7	7.32 (0.61–51.40)	9.47 (0.21–53.93)	6.21	15.29
GPT-4.1	7.90 (1.20–31.14)	9.76 (1.96–31.85)	7.80	16.94
Llama-4-Maverick	8.29 (1.34–72.57)	11.87 (1.68–69.81)	7.77	22.33
GPT-4o	9.23 (1.89–40.08)	13.66 (1.30–48.87)	9.07	20.70
Llama-4-Scout	15.94 (2.09–97.00)	20.51 (1.70–99.18)	14.95	42.23

Table 2: Full page mode results for all models. CER = Character Error Rate, WER = Word Error Rate, CI-CER = Case-Insensitive CER, HChE = Historical Character Error Rate. Numbers in parentheses show (min–max) ranges across all evaluated pages.

5 What LLMs Do Wrong: Statistics and Error Types

Our analysis of 1,030 recognized pages shows that LLMs fundamentally misunderstand historical texts, which results in systematic error patterns. These errors are not random; rather, they form coherent patterns that can be used to investigate the temporal biases embedded in model architectures and training data.

5.1 The Modernisation Trap

The most pervasive and anticipated error pattern is the systematic "correction" of historical orthography to modern standards — a phenomenon we term the *modernization trap*. This manifests as a frequency bias where models overwhelmingly prefer modern character forms over their historical counterparts, effectively enforcing post-1918 orthographic norms on 18th-century texts.

1. Hard Sign (ъ) Elimination. The hard sign, though retained in modern Russian, appears far less frequently today than in the 18th century, when it marked every word ending in a hard consonant. Models systematically either delete it entirely or replace it with soft sign (ъ → ь).

2. Historical Character Replacement / Elimination. Characters eliminated in the 1918 reform are consistently modernized (ѣ (yat) → e), i (decimal i) → и, ѳ (fita) → ф) or totally skipped. Notably, the treatment of ѣ varies by model architecture: some recognize it visually and attempt substitution with phonetically or visually similar characters (e, ъ), while others simply omit it, suggesting different models process historical characters through different pathways — visual recognition versus linguistic mapping.

3. The Lost ї Phenomenon. The most revealing pattern involves the letter ї (iota with diaeresis), inherited from the Greek alphabet. This character, standard in pre-Petrine texts, underwent gradual replacement during the 18th century as part of the broader latinization of Russian typography. Unlike the abrupt eliminations of Peter's 1708 reform, the ї → i transition represents a century-long orthographic evolution that has escaped scholarly documentation.

Our corpus captures this transition in progress: while ї appears consistently in texts from the 1750s–1770s (8.15 occurrences per document on average), it gradually yields to the Latin-influenced i by century's end. Character-level analysis reveals systematic failure across all evaluated models: preservation rates range from merely 2.1% (Qwen-2.5-VL, o4-mini) to at best 30.3% (Gemini-2.5-Flash), with 60–84% of instances systematically substituted to modern i. This is not confusion between similar forms—models recognize the character position and visual structure, yet automatically "correct" it to contemporary orthography. The substitution pattern (rather than mere deletion) demonstrates that models *see* ї but treat it as erroneous, lacking any representation of its historical legitimacy.

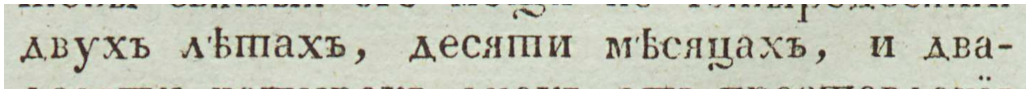
Model-specific patterns expose deeper issues: GPT-4.1 substitutes ї with *Latin* i (not Cyrillic і) in 19.5% of cases, revealing visual encoder confusion between scripts that would be impossible with genuine historical linguistic competence. Even the best-performing model (Gemini-2.5-Flash) fails to preserve ї in 70% of cases—a failure rate that would be unacceptable for any standard Cyrillic character, yet passes unnoticed because ї occupies a blind spot created by its gradual (rather than decreed) obsolescence, the lack of scholarly attention to this transitional period, and universal modernization in all subsequent editions.

The ї case exemplifies how the modernization trap operates at multiple levels: models inherit not just the editorial biases of 20th-century publishing, but also perpetuate gaps in historical scholarship itself. When even the most advanced models achieve at best 30% preservation of a character that appears 8 times per document in authentic 18th-century texts, they demonstrate that training data absence creates not just statistical bias but *epistemic impossibility*—they cannot preserve what they have never learned existed.

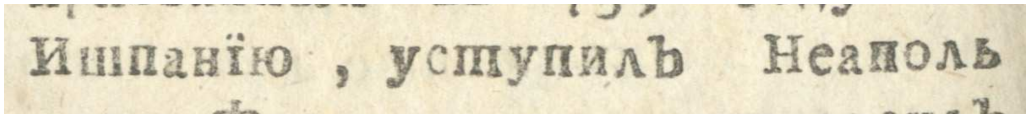
5.2 Over-historization: The Medieval Default

Paradoxically, when multimodal LLMs are prompted to recognize 18th-century book pages as 'historical' Russian text, they frequently insert anachronistic characters drawn from much earlier periods — a phenomenon we term *over-historization*. This pattern suggests that current models lack awareness of periodization.

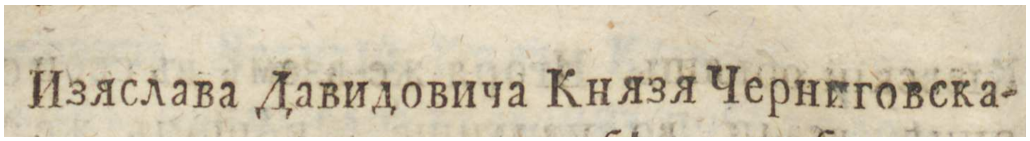
The most striking examples appear in GPT-4 variants, which often introduce characters that



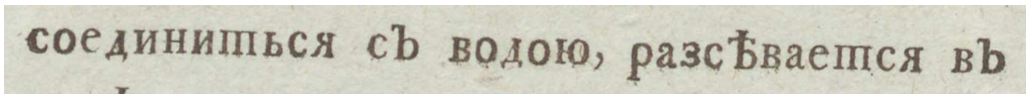
Ground truth: двухъ лѣтахъ, десяти мѣсяцахъ, и два-
gpt-4o: двухъ лѣтахъ, дес ѧ ти мѣсяцахъ, и два-



Ground truth: Ишпанїю, уступилъ Неаполь
gpt-4.1: Іспані ьж , уступилъ Неаполь



Ground truth: Изяслава Давидовича Князя Черниговска-
gemini-2.5-flash: Из ѧ слава Давидовича Кн ѧ з ѧ Черниговска-



Ground truth: соединиться съ водою, разсѣвается въ
claude-sonnet-4: соединишься съ водою, разсѣваетс ѧ въ

Figure 1: Over-historization examples: model outputs compared to ground truth.

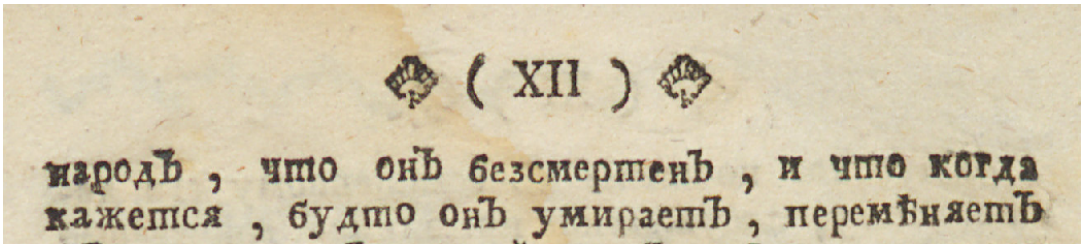
were eliminated in Peter the Great's 1708 reform. See Figure 1 for representative cases.

In these cases, the model substitutes contemporary or period-appropriate characters with obsolete forms, such as ѧ , ѫ or even ьа — characters already obsolete by the early 18th century. Notably, these symbols are inserted precisely in the orthographic positions where they would have appeared in Old Church Slavonic texts but not in 18th-century printed books, suggesting the model recognizes the patterns, but applies them with the temporal gap.

Gemini-2.5-Flash exhibits the most aggressive over-historicization, inserting archaic features including not only pre-Petrine letters (ѧ , ѵ, ѯ) but also Old Church Slavonic diacritical marks. GPT-4.1 and o4-mini show moderate over-historicization, while Claude models remain nearly archaic-free. Russian-language prompts describing the Civil font period reduced insertions by 45% (0.23 → 0.13 per document, $p = 0.200$), though this effect was weaker and non-significant compared to the highly significant improvements in legitimate character preservation ($p < 0.001$).

This reveals models access not merely isolated archaic glyphs but attempt to reconstruct complete historical orthographic systems. The asymmetry between preservation (strong prompt effect) and suppression (weak prompt effect) suggests models struggle with precise temporal boundaries when *generating* absent features, even when successfully *retaining* period-appropriate ones.

This behavior suggests that, when prompted to consider the time period of the text, the models rely on an undifferentiated notion of "Old Russian," conflating the full temporal range from medieval manuscripts to pre-revolutionary print. Rather than demonstrating true historical or linguistic awareness, the models default to the most distinctive and archaic features available in their training data, revealing a lack of temporal and contextual sensitivity.



Book page excerpt (reference)

❖ (XII) ❖

OpenAI o4-mini

✧ (XII) ✧

Gemini 2.5 Flash

⁕ (XII) ⁕

Gemini 2.5 Pro

Figure 2: "Ornamentization" examples. Top: a historical book page with printed ornaments. Bottom: Model outputs, each inserting Unicode symbols in place of decorative elements.

5.3 Visual Similarity Errors and Lost Context

A third category of errors reveals how multimodal models' reliance on visual features can override even basic linguistic knowledge. In 18th-century Russian Civil font, certain letterforms — such as т and ш — share visual similarities absent from modern typography (see the visualization of the word "прошествии" in Figure 3).

Models frequently confuse these characters, producing substitutions like т ↔ ш even when the surrounding word or sentence context would disambiguate the correct reading. This indicates that the models, when faced with degraded or ambiguous glyphs, default to a "visual guess" rather than employing linguistic or contextual analysis. Notably, this error pattern is not universal across models: for example, Qwen and Llama-4-Maverick more often produce т → п substitutions, suggesting that visual similarity errors are model-dependent and may reflect differences in visual encoder architecture or training data composition.

Additional, less frequent error types further illustrate the challenges of multimodal processing: advanced multimodal models sometimes interpret period printing ornaments — such as historical asterisks or flourishes — as textual characters, inserting Unicode symbols into their outputs (see Figure 2). This behavior suggests that models may overfit to visual salience, treating any prominent element as "text," and highlights persistent difficulties in distinguishing content from decorative formatting or noise.

5.4 Model Family "Signatures" and Error Taxonomy

Distinct "error signatures" emerge for each model family, reflecting their underlying training data and architectural biases.

OpenAI models (GPT-4, o4-mini) exhibit systematic over-historicization, frequently inserting pre-Petrine archaic characters such as ѧ , ѫ (little yus and big yus), and monograph uk, despite these being obsolete by the period of our corpus. This "medieval default" pattern appears when the model is uncertain about period features. Paradoxically, OpenAI models also insert modern characters such as ё, which only appeared after 1797. Notably, GPT-4.1 demonstrates cross-script confusion, substituting Cyrillic ї with Latin i in 19.5% of cases—a unique error suggesting visual encoder confusion between similar glyphs across writing systems. While these models preserve ъ at moderate rates (81.9%), their ї preservation remains very low (3.8%), and they exhibit the highest ѣ → е modernization rates (2.3–6.1%), indicating conflicting temporal signals: medieval archaization coexisting with orthographic modernization.

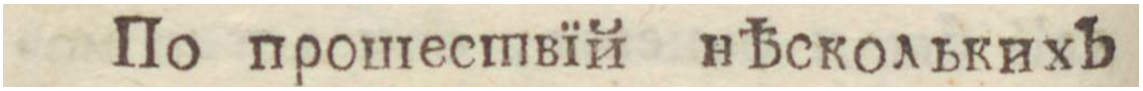


Figure 3: Above-the-line positioning of the character Ѣ in 18th-century Russian print.

Google Models (Gemini series) achieve the highest historical character preservation rates among evaluated systems: Gemini-2.5-Pro preserves 93.1% of ъ and 91.7% of ѣ, while Gemini-2.5-Flash attains 30.3% ї preservation—substantially outperforming all competitors. However, these models still fail ї recognition in 70% of cases and occasionally introduce characters from other Cyrillic traditions (e.g., Ђ from Serbian, є from Ukrainian), indicating "pan-Slavic" training data contamination. Their reliance on visual similarity manifests through frequent т/ш confusions and systematic ї→i substitution (62–72%), suggesting robust visual recognition paired with incomplete historical-linguistic knowledge. Unlike models that delete uncertain characters, Gemini architectures attempt preservation but default to modern equivalents when historical forms lack sufficient training data representation.

Anthropic Models (Claude series) display systematic case inversion errors (ъ → Ъ: 16.5–18.4%, ѣ → Ѣ: 2.2–4.2%) resulting from misinterpretation of typographic positioning. As Figure 3 illustrates, ѣ and ъ frequently appear raised above the baseline in 18th-century typography—a convention Claude's visual encoder interprets as capitalization. The model achieves moderate historical character preservation (ъ: 68–73%, ѣ: 82–86%) but remains vulnerable to this position-based confusion. Occasional insertions of non-Cyrillic characters (e.g., Latin ї) suggest additional cross-script confusion.

Table 3: Historical error signatures in multimodal LLMs. Archaic insertions include pre-1708 characters (ѡ , ѧ , ѵ) and Old Church Slavonic diacritics, eliminated in Peter's reform.

Model	Archaic Insertions (per document avg.)	Most Common Errors
Gemini-2.5-Flash	1.06 (ѧ ,◌̑, ѵ)	ї→i, т→ш
o4-mini	0.40 (ѧ , ѥ , ҂)	ї→i, ъ→ь
GPT-4.1	0.32 (ѧ , ѡ ,◌̓)	ї→i, ї→i Latin
Gemini-2.5-Pro	0.02 (҂)	ї→i, т→ш
Claude-3.5	0.02 (ѵ, ѳ)	ї→i, ъ→Ъ
Claude-3.7	0	ї→i, ъ→Ъ
Qwen-2.5-VL	0.03 (ѧ)	ї→i, т→п

These family-specific error patterns reveal a fundamental paradox in how LLMs process historical text. Models simultaneously modernize authentic 18th-century features (ѣ→e, ї→i) while inserting anachronistic medieval characters (ѧ , ѫ,◌̓) that were already obsolete by 1708. This is not simple confusion but temporal blindness — models possess knowledge of various orthographic stages but cannot map these features to their correct historical periods.

This paradox is compounded by a visual-linguistic disconnect in multimodal processing. When confronting ambiguous historical letterforms (like the similar т/ш in 18th-century font), models default to visual pattern matching rather than leveraging linguistic context that would resolve the ambiguity. This reveals that multimodal integration remains superficial — visual and linguistic processing operate in parallel rather than synergistically.

This goes far beyond expected editorial bias. We anticipated modernization from training data hegemony, but the simultaneous archaization reveals an architectural impossibility: without temporal metadata as a training signal, models cannot develop historical linguistic competence — the capacity to correctly place orthographic features within their appropriate temporal contexts. While

models possess fragmented knowledge of historical forms, they cannot organize this knowledge temporally — transforming apparent digitization success into an invisible mechanism of historical distortion that makes these tools particularly dangerous for historical research.

6 Limitations

Scope and Generalizability. Our evaluation focuses specifically on 18th-century Russian texts printed in Civil font, which limits generalizability in several dimensions. First, our findings may not extend to other historical languages, scripts, or time periods, as different orthographic traditions present distinct challenges for LLMs. Second, the Civil font represents a particular typographic style; handwritten manuscripts, earlier print traditions, or non-European scripts may exhibit different error patterns. Third, our dataset spans only 1750–1800, capturing one specific period of Russian orthographic evolution. The temporal biases we identify may manifest differently for other historical transitions.

Dataset and Ground Truth Limitations. Despite rigorous manual correction, our ground truth may contain residual annotation errors, particularly for visually ambiguous or degraded source material. The decision to use diplomatic transcription principles reflects specific scholarly choices that may not align with all use cases. Our 1,030-page corpus, while substantial, represents a fraction of 18th-century Russian print culture and may not capture the full range of typographic variation, subject matter, or regional printing practices. Additionally, our stratification by text density, decorative elements, and subject matter, while systematic, cannot account for all factors that influence OCR difficulty.

Methodological Constraints. Our evaluation relied mostly on API access to commercial models, preventing analysis of internal model architecture or training data composition. This constrains our ability to definitively explain the mechanistic origins of observed error patterns. Model outputs can vary between runs due to inherent non-determinism, though our stability testing suggests this variation is relatively minor. We excluded some potentially relevant models (e.g., OpenAI o3) due to cost constraints, and our evaluation period represents a snapshot in time as these models continue to evolve rapidly.

Theoretical and Interpretive Limitations. While our concept of "epistemic anachronism" provides a useful framework for understanding LLM temporal bias, the causal connections between specific error patterns and broader claims about training data composition remain partially inferential. We cannot directly examine the historical texts present in each model's training data, limiting our ability to definitively prove contamination or editorial bias transmission. Our model family "signatures" represent observable patterns rather than confirmed architectural explanations.

Reproducibility Challenges. While we commit to releasing our evaluation dataset and scripts, the commercial models we evaluated are continuously updated, making exact replication challenging. Future researchers may observe different results as model capabilities evolve, and some of our findings may become dated as training practices improve.

7 Paths Forward

The challenges identified in our analysis require solutions that address both the fundamental data problems and the architectural limitations of current LLMs. We are considering several concrete approaches that could move toward genuine historical linguistic competence.

Current models largely ignore temporal metadata. The research agenda should propose adapting existing time-aware frameworks. For example, one could replicate the approach of Dhingra et al. [4], who modified T5's pre-training objective to be parameterized with timestamp information, or the TALM model, which transfers general language models to time-specific domains. A concrete experimental proposal would be to fine-tune a model where every document fragment is

explicitly conditioned on its publication year. This would provide the model with an explicit signal for periodization that is currently absent.

The implications extend beyond traditional humanities research. Varnum et al. (2024) propose using 'Historical Large Language Models' (HLLMs) trained on historical corpora to simulate responses from past populations [16]. However, their approach depends fundamentally on access to accurately digitized historical texts, requiring researchers to 'acquire a sizable amount of historical text from a society from a specific time period' and convert it 'to a machine-readable format.' This emerging application reveals the circular nature of the challenge: creating HLLMs requires historical corpora, but building such corpora has been historically constrained by digitization limitations. Reliable OCR systems that preserve historical authenticity with minimal human intervention could break this cycle, enabling LLMs to help construct the comprehensive historical corpora that were previously unattainable.

While research on previous generations of models has shown that time is systematically encoded in the model's weights, allowing for fine-tuning to enhance performance on text from a specific target time period [10], this is significantly more complicated for modern large language models (LLMs). This complexity arises because today's foundational models are typically trained on massive, mixed-temporal corpora all at once, which results in a more deeply integrated and less decomposable representation of time than the modular "time vectors" identified in smaller, time-specific models. Consequently, knowledge about time is not a discrete, editable component but is encoded as complex statistical patterns across the entire network, making the isolation and manipulation of specific temporal features a far greater challenge [17].

Second, a highly promising and computationally efficient approach is the use of Time Vectors, a concept introduced by Nylund et al. [9]. A time vector is created by fine-tuning a pre-trained model on data from a specific time period and then subtracting the original model's weights. The resulting vector represents a direction in the model's weight space that specializes it for that period. The research agenda could propose a fascinating experiment: create a "1750s vector" and a "1790s vector" from the paper's dataset. The structure of these vectors could then be explored, and critically, they could be interpolated to create a "virtual" 1770s model without any additional training. This presents a novel path toward creating models for periods with sparse data.

8 Conclusion

Our findings present a paradox. While recent studies show LLMs develop rich internal representations of spatiotemporal information [6], our results indicate the simultaneous modernization and archaization of historical texts. This suggests a fundamental disconnect between what the models "know" and what they generate.

We hypothesize this arises from a conflict between representational and generative components. Early transformer layers may encode temporal awareness, but the final output layers—dominated by modern text frequencies in training data — override this historical information [11]. When modern Russian orthography outnumbers 18th-century forms by orders of magnitude, statistical priors overwhelm subtler historical signals.

This hypothesis points toward new research directions. Mechanistic interpretability techniques, successfully used to dissect social biases in LLMs, could trace which attention heads or layers enforce modernization. The systematic errors we document are not technical glitches but symptoms of misalignment between AI development and historical scholarship. As digital methods increasingly mediate access to the past, these biases risk becoming invisible filters on history itself.

References

[1] Boelaert, J., Coavoux, S., Ollion, E., Petev, I. D., and Präg, P. “How Do Generative Language Models Answer Opinion Polls?” In: *Sociological Methods and Research 0*, no. 0 (2025). DOI: 10.1177/00491241251330582.

[2] Chandna, Bhavik, Bashir, Zubair, and Sen, Procheta. “Dissecting Bias in LLMs: A Mechanistic Interpretability Perspective”. 2025. URL: arXiv:2506.05166.

[3] Coeckelbergh, Mark. “LLMs, Truth, and Democracy: An Overview of Risks”. In: *Science and Engineering Ethics 31* (Jan. 2025). DOI: 10.1007/s11948-025-00529-0.

[4] Dhingra, Bhuwan, Cole, Jeremy R., Eisenschlos, Julian Martin, Gillick, Daniel, Eisenstein, Jacob, and Cohen, William W. “Time-Aware Language Models as Temporal Knowledge Bases”. In: *Transactions of the Association for Computational Linguistics 10* (2022), ed. by Brian Roark and Ani Nenkova, pp. 257–273. DOI: 10.1162/tacl_a_00459.

[5] Gallegos, Isabel O., Rossi, Ryan A., Barrow, Joe, Tanjim, Md Mehrab, Kim, Sungchul, Dernoncourt, Franck, Yu, Tong, Zhang, Ruiyi, and Ahmed, Nesreen K. “Bias and Fairness in Large Language Models: A Survey”. In: *Computational Linguistics 50*, no. 3 (Sept. 2024), pp. 1097–1179. ISSN: 0891-2017. DOI: 10.1162/coli_a_00524.

[6] Gurnee, Wes and Tegmark, Max. “Language Models Represent Space and Time”. 2024. DOI: arXiv:2310.02207.

[7] Inoue, Kotaro. “Context-Independent OCR with Multimodal LLMs: Effects of Image Resolution and Visual Complexity”. In: (2025). DOI: arXiv:2503.23667.

[8] Levchenko, Maria. “Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities”. In: (2025). DOI: arXiv:2510.06743.

[9] Nylund, Kai, Gururangan, Suchin, and Smith, Noah. “Time is Encoded in the Weights of Finetuned Language Models”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2571–2587. DOI: 10.18653/v1/2024.acl-long.141.

[10] Nylund, Kai, Gururangan, Suchin, and Smith, Noah A. “Time is Encoded in the Weights of Finetuned Language Models”. 2023. arXiv: 2312.13401 [cs.CL]. URL: https://arxiv.org/abs/2312.13401.

[11] Park, Yein, Yoon, Chanwoong, Park, Jungwoo, Jeong, Minbyul, and Kang, Jaewoo. “Does Time Have Its Place? Temporal Heads: Where Language Models Recall Time-specific Information”. 2025. DOI: arXiv:2502.14258.

[12] Poibeau, Thierry. “Bias, Subjectivity and Norm in Large Language Models”. In: *https://ceur-ws.org/Vol-3808/*. Saint Jacques de Compostelle, Spain, Oct. 2024. URL: https://cnrs.hal.science/hal-04838836.

[13] Rao, Abhinav, Yerukola, Akhila, Shah, Vishwa, Reinecke, Katharina, and Sap, Maarten. “NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models”. In: *North American Chapter of the Association for Computational Linguistics*. 2024. DOI: arXiv:2404.12464.

[14] Theimer, Kate. “A Distinction Worth Exploring: “Archives” and “Digital Historical Representations””. In: *Journal of Digital Humanities 3*, no. 2 (2014). Summer. URL: https://journalofdigitalhumanities.org/3-2/a-distinction-worth-exploring-archives-and-digital-historical-representations/.

[15] Underwood, Ted, Nelson, Laura K., and Wilkens, Matthew. “Can Language Models Represent the Past without Anachronism?” 2025. DOI: arXiv:2505.00030.

[16] Varnum, Michael E. W., Baumard, Nicolas, Atari, Mohammad, and Gray, Kurt. “Large Language Models based on historical text could offer informative tools for behavioral science”. In: *Proceedings of the National Academy of Sciences 121*, no. 42 (2024), e2407639121. DOI: 10.1073/pnas.2407639121.

[17] Ye, Xiaotian, Zhang, Mengqi, and Wu, Shu. “Open Problems and a Hypothetical Path Forward in LLM Knowledge Paradigms”. 2025. DOI: arXiv:2504.06823.

[18] Zaagsma, Gerben. “Digital History and the Politics of Digitization”. In: *Digital Scholarship in the Humanities 38*, no. 2 (Sept. 2022), pp. 830–851. ISSN: 2055-7671. DOI: 10.1093/llc/fqac050.