Multi-Modal Large Language Models for Historical Handwritten Text Recognition (HTR) and Data Augmentation

Yuanhao Zou[1,2] Merve Tekgürler[1,3]
[1]Department of History [2]Department of Computer Science [3]Symbolic Systems Program
Stanford University
{zouyh, mtekgurl}@stanford.edu

Abstract

This paper investigates the capabilities of Gemini 2.0, a multimodal large language model, for the vision task of handwritten text recognition (HTR) in transcribing historical Ottoman Turkish manuscripts. We evaluate Gemini's performance across zero-shot, text-only, and image+text input settings using a newly constructed dataset of 18th-century manuscript images aligned with scholarly transcriptions and transliterations. While the model exhibits limitations—including script mixing and occasional prompt noncompliance—our results demonstrate that multi-modal prompting significantly improves transcription quality in paleographically challenging contexts. We show that Gemini can leverage both visual and linguistic inputs to generate high-fidelity outputs, enabling a scalable method for producing Arabic-script transcription data from widely available Latin-script editions. Our approach introduces a novel, resource-efficient direction for building HTR datasets and systems for historical, non-Latin-script documents.

1. Introduction

Handwritten Text Recognition (HTR) is a vision task that is of great value to historians working with historical manuscripts. When applied to Ottoman Turkish manuscripts, the task presents a unique set of challenges. These Ottoman historical documents are typically handwritten in Arabic script, but the majority of modern scholarly work follows the convention of modern Turkish and relies on transliterations of these manuscripts into the Turkish Latin script. While a huge number of scanned images of manuscripts and their corresponding Latin-script transliterations exist, there is a significant paucity in the amount of transcriptions in the original Arabic script. However, the transcriptions in the original script of the language are essential for building datasets to develop robust HTR models that can turn Ottoman manuscripts into computer-readable forms.

This project aims to address this gap by exploring the use of large multimodal models, specifically Gemini 2.0 Flash, for transcribing Ottoman Turkish manuscripts in its original Perso-Arabic script and augment existing data for training purposes. Our work proceeded in two stages. First, we evaluated Gemini's transcription accuracy using both zero-shot and one-shot prompting strategies on a manually aligned dataset of manuscript images and Arabic-script transcriptions in order to establish a baseline for model performance. Then, we leverage the power of the LLM to perform data augmentation on Ottoman manuscript - in particular, we explored best strategies for generating accurate Arabic-script transcriptions from manuscript images and their corresponding Latin-script transliterations. Once scaled, this can enable the reconstruction of Arabic text from widely available Latin editions and thereby, significantly expand access to digitized Ottoman sources in their original script. This goal is of great significance for the fields of history, historical linguistics, and digital humanities, as well as for advancing the technological scope and linguistic versatility of HTR models more broadly.

While HTR models for historical texts have profoundly reshaped research on historical documents in North America and Western Europe, their application to manuscripts written in non-Latin scripts remains limited and underdeveloped. Ottoman Turkish, the official language of the Ottoman State (1299–1922), was written in a modified Perso-Arabic script and represents a critical historical stage in the evolution of Modern Turkish. However, the 1928 alphabet reform in Turkey led to a sharp decline in the use and accessibility of Ottoman Turkish, rendering it a digitally marginalized language. Today, Modern Turkish is written in Latin script, which means that even native speakers of Turkish cannot access Ottoman Turkish documents without specialized education. HTR technologies can help mitigate the costs-time, labor, as well as financial-associated with reading historical artefacts in the post-Ottoman world.

Despite the vast number of Ottoman Turkish manuscripts

housed in libraries and archives worldwide, these resources remain largely inaccessible to computational analysis. By developing tools to convert handwritten Ottoman manuscripts into machine-readable formats, our project not only unlocks previously inaccessible historical materials but also contributes to a broader effort to diversify digital access to non-Latin-script historical sources.

2. Related Work

As outlined above, Ottoman Turkish was written in the Perso-Arabic script. Thus, HTR projects in Arabic-script languages offer invaluable insights into our research. Two most recent efforts on this topic are Muharaf [10] and Qalam [2] projects. Muharaf project was developed by North Carolina State University Khayrallah Center for Lebanese Diaspora Studies. Within the scope of this project, the researchers publicly released the Muharaf dataset. It consists of 1,644 scanned manuscript images, transcribed and annotated by experts in archival Arabic. The dataset also includes spatial polygonal coordinates for text lines and various page elements. In the same paper, they experiment with a CNN-based SFR (Start, Follow, Read) model, first applied to historical HTR by Wigington et al. [11]. Additionally, the same team published another paper, HATFormer [5], detailing their experiments with a Transformer-based Optical Charactor Recognition (TrOCR) [9] model, albeit without releasing the code. HATFormer is a transformer-based model for historical Arabic handwritten text recognition (HTR), addressing the challenges of cursive writing, context-sensitive character shapes, and diacritics. Leveraging an adapted TrOCR architecture, HATFormer integrates a vision transformer (ViT) encoder and a RoBERTa text transformer decoder, achieving state-of-the-art performance on historical and modern Arabic handwritten datasets.[3].

Our research also broadly connects to emerging efforts on developing natural language processing pipelines for Ottoman Turkish and other Turkic languages. Most noteworthy among these efforts was the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024) [1] held on August 15, 2024, in conjunction with Annual Meeting of the Association for Computational Linguistics.

For the last two years one of the authors, Tekgürler, has been collaborating with Peter Broadwell and Umar Patel on developing an HTR model for historical Arabic script languages. The first draft of their paper "Multilingual handwritten text recognition (HTR) models for large-scale processing of archival documents in low-resourced Arabic-script languages" (2025) addresses the challenge of multilingual HTR for archival documents and manuscripts in low- resourced Arabic-script languages and provides a pipeline to process these documents. This pipeline includes layout recognition, strategic pre-processing, transcription using a fine-tuned TrOCR model, and Ottoman Turkish transliteration.

3. Data

For this project, we created the **Osman Agha dataset**, a manually aligned collection of historical manuscript images and transcriptions in the Perso-Arabic script as well as transliterations in Latin script. The dataset consists of 243 page-level image-text pairs from an eighteenth-century Ottoman Turkish manuscript written in Arabic script. The manuscript contains the memoirs of Osman Agha of Timişoara, an Ottoman soldier who was captured and held as a prisoner of war in Austria between 1688 and 1699. The memoir was completed on May 18, 1724 and survives as a single authorial copy, currently housed at the British Library (MS. Or. 3213).

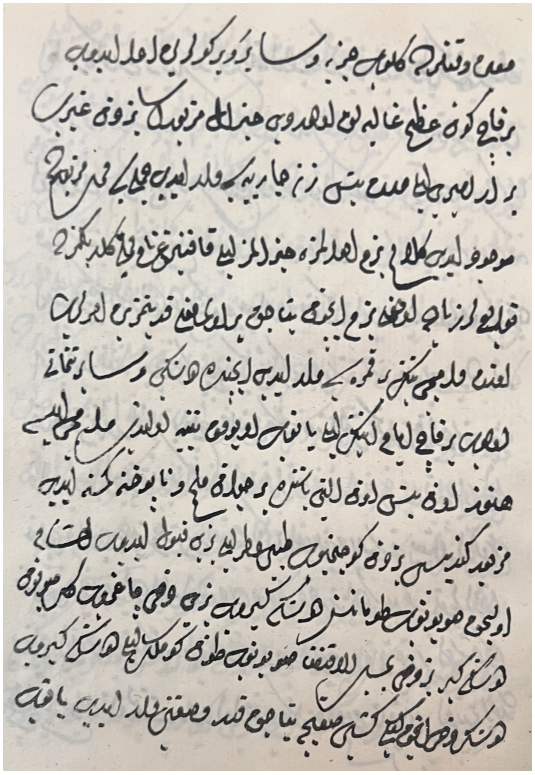

Figure 1: An example page from our manuscript *Esâretnâme*.

The manuscript is commonly known in English as *Prisoner of the Infidels: The Memoirs of Osman Agha of Timişoara*, the title of its 2021 English translation1[4]. A full transcription was previously published in 1980 [7]. Our dataset combines high-resolution images of the original manuscript—photographed by Julia Fine in August 2024—with OCR-generated text based on the 1980 tran-

scription. In order to turn the manuscript into trainable data for the purpose of this project, the authors of this article used the Gemini 2.0 Flash model to produce an initial OCR output of the transcription, which we then manually aligned with the corresponding manuscript page images.

Similarly, we identified a scholarly edition of this work in Latin script [6]. We extracted texts from this edition and manually aligned it at page-level with the manuscript and the transcription.

The resulting dataset enables page-level OCR benchmarking and evaluation for historical Ottoman Turkish documents written in Arabic script, a domain that has long been underserved in digitization and computational transcription efforts.

4. Methods

We employ‘ zero-shot and one-shot prompting strategies to evaluate the transcription capabilities of the Gemini 2.0 model on a historical manuscript dataset.

In this section, we describe the various prompting strategies and input modalities used to evaluate and improve automatic transcription of Ottoman Turkish manuscripts. We first establish baseline performance using image-only inputs under zero-shot and one-shot prompting conditions. We then explore enhanced transcription generation by incorporating transliterations alongside manuscript images. Finally, we assess the effect of using transliteration alone as input, to test its potential as a supplementary or standalone modality.

4.1. Baseline Prompt Engineering

To establish a baseline, we experiment with two prompting strategies using image-only inputs: zero-shot and one-shot prompting. These approaches allow us to assess the model's capacity to generate transcriptions directly from manuscript images without relying on auxiliary information at every step.

4.1.1 Zero-shot Prompting

We begin by constructing a test set comprising approximately 10% of the manuscript—roughly 25 pages—selected to be evenly distributed across the entire document to ensure coverage of stylistic variation and handwriting evolution. For the zero-shot condition, we prompt Gemini 2.0 with a general instruction to transcribe each manuscript page into Ottoman Turkish using the original Perso-Arabic script. No examples, in-context demonstrations, or specific guidance beyond the task description are provided. This setting allows us to evaluate the model's transcription abilities based purely on its pretrained capabilities and general knowledge of paleographic conventions.

4.1.2 One-shot Prompting

In the one-shot setup, we augment the prompt with a single in-context example. We randomly select one manuscript page (excluded from the test set) and supply the model with both the image of that page and its accurate ground-truth transcription. This example establishes a pattern that informs the model of the expected transcription style, formatting, and output structure. During inference, the same prompt structure is reused for each page in the test set: we retain the in-context example while replacing the test image with the page to be transcribed. This configuration tests whether minimal in-context learning can lead to more stable and accurate transcriptions.

4.2. Multi-modal Input: Image + Text

In transcribing Ottoman Turkish manuscripts written in Perso-Arabic script, a major challenge lies in the irregularity of handwriting, orthographic ambiguity, and the inherent limitations of OCR systems trained on modern printed texts. To mitigate these challenges, we propose a novel form of data augmentation that leverages scholarly Latin-script transliterations as auxiliary supervision.

Beyond image-only transcription, we evaluate the effectiveness of multi-modal prompting by supplying both the manuscript image and its corresponding Latin-script transliteration as input. We refer to this method as *image+text*. The transliterations are scholarly romanizations of the manuscript content, typically prepared by experts or semi-automated pipelines. They serve as auxiliary scaffolding to help the model disambiguate visually ambiguous characters and orthographic conventions, particularly in contexts where the handwriting is unclear or stylized. The system prompt is carefully engineered to instruct the model to prioritize the manuscript image while using the transliteration only when needed. This strategy helps mitigate overreliance on modern spelling conventions present in the Latin script, and leverages the strengths of both modalities to generate historically grounded transcriptions in Perso-Arabic script.

4.3. Transliteration-Only Input: Text Only

For comparative purposes, we also run experiments where the model is given only the transliteration of each page, without access to the manuscript image. We refer to this setup as *text only*. In this configuration, the model is prompted to convert Latin-script Ottoman Turkish into the original Perso-Arabic script. The goal is to assess how well the model can recover historically accurate orthographic forms when given romanized input alone. This setting also serves as a reference point for evaluating the contributions of visual information from the manuscript in the *image+text* condition. In some cases, the model's output in this mode

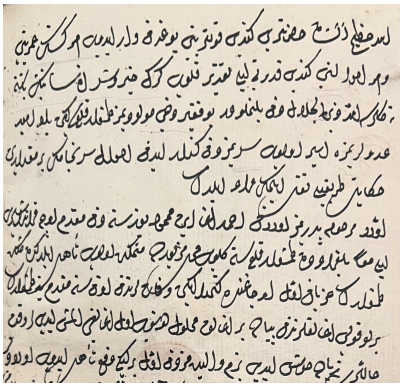

Figure 2: Original Manuscript

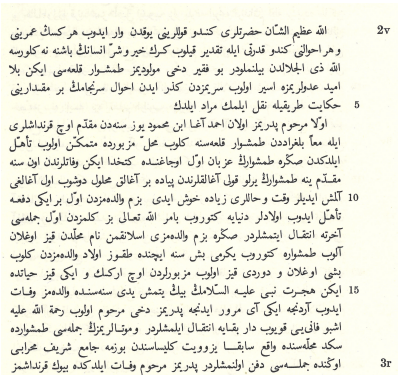

Figure 3: Transcription

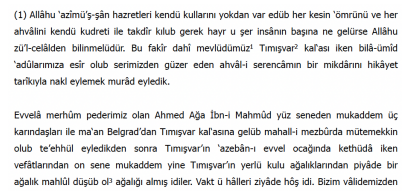

Figure 4: Transliteration

may reflect regularization effects or modern orthographic bias, providing insight into its internal representation of Ottoman morphology and vocabulary.

It should be noted that generating accurate Ottoman Turkish transcription in Perso-Arabic script from Latin-script transliteration is a challenging task. The Latin transliteration, while systematic, does not encode all the orthographic and phonological nuances of the original script. Ottoman Turkish employs the Arabic script with adaptations from Persian, featuring multiple letters for the same phoneme (e.g., "sin" vs. "sad" for /s/), silent letters, and etymologically motivated spellings that are not recoverable from pronunciation-based transliterations. Moreover, the lack of a one-to-one correspondence between the two alphabets means that restoring the original script requires not only phonological inference but also historical, morphological, and lexical reasoning. The model must therefore reconstruct script forms that were often standardized according to etymology or scribal convention rather than pronunciation alone. These all render this a non-trivial transcription task.

4.4. Prompting

Given the sensitivity of large language models to prompt phrasing and task framing, we experiment with a range of prompting strategies to optimize the accuracy and consistency of transcription outputs. Our core task involves generating Ottoman Turkish transcriptions in Perso-Arabic script based on inputs that may include Latin-script transliterations, manuscript images, or both. We design system prompts that explicitly define the model's role as a specialist in Ottoman paleography and orthography, and carefully instruct it to adhere to historical spelling norms, omit short vowel markings, and preserve line breaks. To ensure stable and high-fidelity outputs across varying manuscript styles and input lengths, we test both minimal and elaborate prompt formulations. Minimal prompts provide essential task instructions only, while more elaborate variants include contextual information about Ottoman script conventions and common transcription challenges. We also experiment with prompt chaining and in-context examples to improve performance on ambiguous or degraded passages. By systematically evaluating these different strategies, we identify prompt formats that consistently yield more accurate, structurally coherent transcriptions with fewer orthographic errors and higher similarity to ground-truth data. This prompting methodology is central to the model's reliability and plays a key role in stabilizing output quality across the dataset.

Method	**Input**	**Description**
Zero-shot	Manuscript image only	No examples or prior context; model transcribes directly from image using a general prompt.
One-shot	Manuscript image + 1 in-context example	Single example page with transcription used to guide output format and style.
Image+Text	Manuscript image + Latin transliteration	Model instructed to prioritize image and use transliteration as auxiliary input.
Text Only	Latin transliteration only	Model transcribes into Perso-Arabic script without access to manuscript image.

Table 1: Summary of prompting and input configurations used for Ottoman Turkish transcription.

4.5. Evaluation Metrics

We use **Levenshtein similarity** as our primary evaluation metric to assess the quality of transcriptions generated by the Gemini model.[8] Levenshtein similarity is a normalized measure derived from Levenshtein distance, which calculates the minimum number of single-character insertions, deletions, or substitutions required to transform one string into another. By normalizing this distance with respect to the length of the longer string, we obtain a similarity score ranging from 0 to 1, where higher values indicate greater textual overlap and accuracy. The metric is defined as:

$$\text{Similarity}(a, b) = 1 - \frac{D(a, b)}{\max(|a|, |b|)}$$

where $D(a, b)$ is the Levenshtein distance between strings a and b, and $|a|$, $|b|$ denote their respective lengths. This metric is particularly well-suited for historical manuscript OCR tasks, where minor deviations in character-level output can significantly affect readability and fidelity to the original source.

5. Experiments and Results

5.1. One-shot Prompting

To evaluate the effectiveness of different prompting strategies for historical manuscript transcription, we conducted a comparison between zero-shot and one-shot prompting outlined above.

We computed the average normalized Levenshtein similarity between the transcriptions produced by the model and a gold-standard reference. The zero-shot setting outperformed one-shot, achieving a similarity of 0.4204 versus 0.3309. This indicates that, contrary to expectations, the inclusion of a single example did not improve transcription accuracy and may have introduced misleading context.

Prompting Method	Levenshtein Similarity
Zero-shot	0.4204
One-shot	0.3309

Table 2: OCR performance comparison between prompting strategies.

This result suggests that providing a single in-context example did not improve the model's ability to generalize; in fact, it appears to have introduced noise or misalignment in the model's behavior. One possible explanation is that the one-shot prompt may have biased the model toward the stylistic or structural features of the example page, which did not generalize well across the diverse set of test images. These findings highlight the need for careful prompt engineering and example selection when using large multimodal models such as Gemini for historical OCR tasks. Given that the zero-shot method outperforms the one-shot method we attempted, we proceeded to perform zero-shot transcription on the entire manuscript to establish a baseline for our project.

5.2. Best Prompting Technique

Across our experiments, we found that prompt design has a significant impact on the accuracy, stability, and stylistic fidelity of the model-generated transcriptions. Two prompting strategies consistently produced the most effective results: one designed for multi-modal input combining manuscript images with Latin-script transliterations, and another tailored for transliteration-only input.

The most robust and consistent results were achieved using the prompt designed for the *image+text* configuration, in which each input includes both a scanned manuscript image and a scholarly Latin-script transliteration. This prompt frames the model as an "expert in Ottoman Turkish paleography and orthography" and explicitly defines the dual-modality input. The model is instructed to rely primarily on the manuscript image while using the transliteration to resolve ambiguities—particularly in cases where handwriting is unclear or idiosyncratic. This approach emulates historical editorial practices, where experts draw on both paleographic cues and linguistic knowledge to reconstruct intended meanings. The prompt further emphasizes orthographic discipline: preserving original line breaks, omitting short vowel marks, and adhering to the spelling conventions of Persian and Arabic-origin vocabulary. It also directs the model to ignore scribal inconsistencies that can be resolved through reference to the transliteration. This combination of domain specificity, modal prioritization, and strict output constraints proved especially effective in handling degraded or ambiguous passages.

The second effective prompt is tailored for tasks involving only Latin-script transliteration as input. It casts the model as an "expert in Ottoman Turkish script and orthography" and instructs it to convert scholarly transliterations into their original Perso-Arabic script. This prompt benefits from similar structural features: it clearly defines the input format, describes the expected output, and outlines specific constraints to avoid modern Turkish spelling conventions and the addition of short vowels. By forbidding explanatory content and enforcing strict formatting rules, the prompt minimizes variation and improves consistency across outputs. While it does not benefit from image-based contextual cues, it performs reliably when the transliteration is accurate and unambiguous.

Both prompts share essential traits that make them highly effective. They explicitly define the model's role and expertise, clearly articulate the transformation task, impose strict linguistic and formatting constraints, and prohibit unnecessary or unintended output. These design elements work to-

gether to ensure not only high transcription accuracy, but also consistent structure and historical fidelity. The success of these prompting strategies underscores the importance of task contextualization, domain-aware instructions, and strict output formatting in AI-assisted transcription of historical texts.

5.3. Results and Comparison

Combining the methods and the best prompting strategy we derived, we conducted experiments on the entire dataset and reached the following results across the methods:

Approach	Average Levenshtein Similarity
Zero Shot	0.4412
Text Only	0.8536
Image and Text*	0.8224

*We removed the 44 outputs that contained Latin script text for this average. We discuss these corrupted model outputs in section 5.5

Table 3: OCR performance comparison between prompting strategies.

As the table demonstrates, both *image+text* and *text only* methods lead to significant increase in model accuracy, and that even inputing transliteration texts and assigning it an auxiliary role can greatly boost model performance. While the Gemini model at its current stage cannot yet reliably transcribe Ottoman Turkish manuscripts, it can still serve as a tool in creating more transcription data. The combination of manuscript images and the transliterations of these manuscripts can be processed jointly by Gemini to create usable transcription data to be used in HTR tasks. This method shows high accuracy, over 82% and thus offers a viable and scalable solution.

Despite that the average accuracy for *image+text* slightly behind that of *text only*, the following subsection shows several cases where the former method outperforms the latter. These cases demonstrate that the model is, in fact, leveraging visual information extracted from manuscript pages in meaningful ways.

5.4. Qualitative Analysis

This subsection looks at several interesting word-level examples and offers interpretations to these cases. these are important cases

5.4.1 *Götürmek*: Mixing Latin and Arabic

This example is a notable failure we observed: the unintended mixing of Latin and Arabic characters within the same output. The target word is the Ottoman Turkish verb *götürmek* "to take (somewhere)". The zero-shot method produces a completely incorrect output, transliterated as

Method	Output	Transliteration of the output
Zero-shot	گوریکی	göriki
Text only	گوتورمك	götűrmek
Image+text	گوتürmک	göt(ürm)ek
Ground truth	گوترمك	götürmek

Figure 5: The model output for *götürmek* with different methods. Note that the *image and text* method transcribes a segment of the word with Latin letters

göriki, which is phonetically implausible and likely the result of misinterpreting both the image and the linguistic context - it is likely that Gemini interpreted it as a Persian word instead. The *text only* method performs better, producing a transcription that corresponds to *götűrmek*, with minor vowel irregularities but overall structural correctness. Most surprising is the *image+text* output, which, when transliterated, appears as *göt(ürm)ek*, suggesting that the model inserted parts of the Latin input (*ürm*) directly into the transcription, instead of converting them into Arabic-script equivalents5. We interpret this kind of hybrid output—partially Latin, partially Arabic— as a mistake that demonstrates an instance where the model "looks" at the manuscript page, but referred to the Latin transliteration of the page when encountering a paleographically challenging part of the word.

5.4.2 *Hot gehe*: German word in Ottoman Text

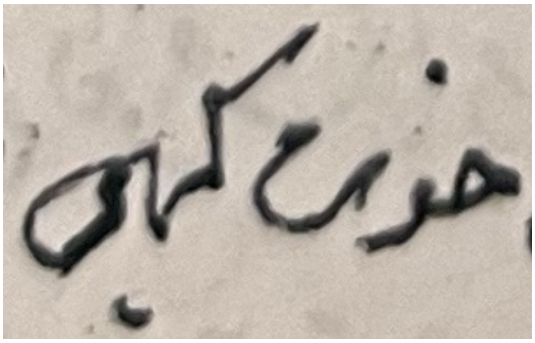

Figure 6: The German dialect phrase *hot gehe* written out in the Ottoman script in the original manuscript

This example presents a rare but highly significant case in which the model correctly identifies and reproduces a foreign-language phrase embedded in an Ottoman Turkish manuscript. The original phrase in the manuscript—written in Perso-Arabic script as *hot gehi*—represents an Ottomanized rendering of the German dialect expression *Hot! Gehe*,

Method	Output	Transliteration of the output
Zero-shot	خوات کهی	hovât gehi
Text only	Hot! Gehe	Hot! Gehe
Image+text	Hot! Gehe	Hot! Gehe
Ground truth	خوت کهی	hot gehi

Figure 7: The model output for *hot gehi* with different methods. Note that the *zero shot* method most successfully generates an output in the original Perso-Arabic script

meaning “Okay! Let’s go.”6 Despite the prompt’s instruction to output all transcriptions in Perso-Arabic script, both the *text only* and *image+text* methods return the phrase in its original German and Latin-script form7. While technically a violation of the script constraint, the output is linguistically accurate and contextually appropriate.

Interestingly, the zero-shot method—relying solely on the manuscript image—produces a partial approximation (*hovât gehi*) in Arabic script that captures some phonetic elements of the phrase but lacks the clarity and semantic precision of the other methods. This suggests that the presence of the transliteration in the prompt significantly helped the model recognize the embedded German, even if it overrode the script instruction in the process.

Rather than a simple failure of prompt compliance, this case illustrates the model’s capacity to detect and preserve multilingual features within historical texts. In doing so, it mirrors the linguistic hybridity characteristic of many Ottoman documents, particularly those written in regions with strong German influence. The fact that the model preserves the German phrase verbatim suggests a degree of contextual reasoning and multilingual awareness, though it also raises important questions about when and how script constraints should be enforced. This example points to the need for nuanced evaluation metrics that account not only for transcription accuracy but also for the preservation of embedded linguistic diversity in historical sources. Future research should take this linguistic diversity in Ottoman texts into account.

5.4.3 *Arabacı*: learning from both image and text

A particularly illustrative example of the benefits of the *image+text* method is found in the transcription of the word *arabacı* (coachman) on page 10r of the manuscript, as shown in Figure 8. This case reveals how the model responds to different input configurations. In the zero-shot setting, the model misreads the word as *avrcı*, a non-existent form. This error reflects a plausible paleographic confusion when the model is only exposed to the image: the scribe’s form of the letter combination ”be” and ”he” closely resembles ”re”, a common visual ambiguity in Ottoman handwriting conventions. When the model is prompted solely with the Latin-script transliteration, it generates *arâbacı*, overcorrecting by inserting a long â in the second syllable, likely due to its exposure to modern or Persianized spelling patterns. However, when both the image and the transliteration are provided, and the prompt explicitly instructs the model to rely primarily on the manuscript while using the transliteration as supplementary reference, the model successfully outputs the correct form *arabacı*9. This example underscores not only the limitations of image-only and text-only settings, but also the unique strength of the *image+text* method: by combining visual evidence from the manuscript with phonological cues from the transliteration, the model is better able to disambiguate ambiguous letterforms and produce a historically accurate transcription for Ottoman texts where both manuscript images and transliteration pages, when standing along, can generate confusions.

The case of *arabacı* offers evidence that Gemini, when

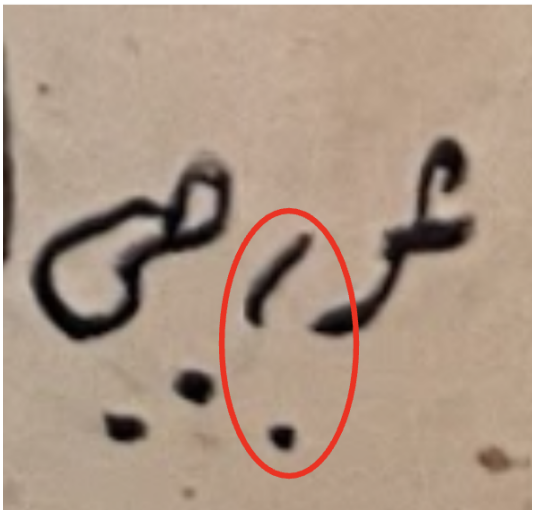

Figure 8: The Ottoman Turkish word *arabacı* on page 10r of the manuscript. The letter combination ”be” + ”he” circled out in the figure proves to be paleographically challenging for the model

prompted with the *image and text* method, is not merely defaulting to the Latin-script input, but actively incorporating visual features—such as character shape, spacing, and layout—into its output generation. This indicates that the model is performing a form of computer vision, interpreting the handwriting and integrating it with linguistic knowledge to produce more accurate or historically faithful transcriptions in complex cases. Leveraging both visual and linguistic information, This method is clearly a scalable in creating high-quality labeled datasets of transcribed texts for the many Ottoman manuscripts that were transliterated but not

Method	Output	Transliteration of the output
Zero-shot	عورجی	avrcı
Text only	عرابەجی	arâbacı
Image+text	عربەجی	arabacı
Ground truth	عربەجی	arabacı

Figure 9: The model output for *arabacı* with different methods. Note that when the model has access to both the manuscript image and the transliteration text, it generates the correct output.

transcribed.

5.5. Generation Issues

While Gemini demonstrates strong capabilities in both image understanding and text generation, we observed recurring instability when the model is prompted to transcribe Ottoman Turkish from both manuscript images and accompanying transliterations. In this multi-modal setting, the model occasionally fails to perform the intended transcription task altogether—simply returning the input transliteration in Latin script rather than generating the corresponding text in Perso-Arabic script. Even when it does produce transcription output, the model sometimes violates prompt constraints, such as by adding short vowel signs, despite explicit instructions to omit them in accordance with historical manuscript conventions. These failures suggest that the model struggles with modality control and instruction adherence when managing two forms of input. Without fine-grained conditioning or a modality-aware architecture, Gemini may default to the more accessible input (i.e., the transliteration) or blend outputs inconsistently. We will experiment with more stabilizing prompting strategies in the future. For this project, we had to discard 44 outputs of *image and text* because of this issue.

6. Conclusion

This study explores the use of large multimodal language models—specifically Gemini 2.0 Flash—for transcribing Ottoman Turkish manuscripts written in Perso-Arabic script. By leveraging paired Latin-script transliterations and manuscript images, we demonstrate that Gemini can produce historically faithful transcriptions that were previously difficult to obtain at scale. Through systematic evaluation across *zero-shot*, *text-only*, and *image+text* modalities, we find that the combination of visual and textual input often outperforms unimodal approaches, especially in resolving paleographic ambiguities.

Our results highlight both the promise and current limitations of applying multimodal LLMs to historical handwritten text recognition (HTR) in non-Latin scripts. On the one hand, the model demonstrates encouraging accuracy, flexibility, and the ability to learn from both visual layout and linguistic cues. On the other hand, issues such as script mixing, prompt instability, and partial noncompliance reveal the need for improved modality control and task alignment.

Despite these challenges, we argue that multimodal prompting—particularly with aligned manuscript-transliteration pairs—offers a scalable, low-cost method for generating transcriptions of Arabic-script Ottoman texts. This approach not only facilitates the creation of much-needed HTR training data, but also paves the way for more inclusive and historically sensitive computational methods for digitally marginalized scripts. Our future work will focus on refining prompt stability, exploring fine-tuning strategies, and expanding our framework to include multilingual contexts characteristic of early modern manuscripts that we as historians work with on a daily basis.

References

[1] D. Ataman, M. O. Derin, S. Ivanova, A. Köksal, J. Sälevä, and D. Zeyrek, editors. *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, Bangkok, Thailand and Online, Aug. 2024. Association for Computational Linguistics.

[2] G. Bhatia, E. M. B. Nagoudi, F. Alwajih, and M. Abdul-Mageed. Qalam: A multimodal LLM for Arabic optical character and handwriting recognition. In N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, and K. Mrini, editors, *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 210–224, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[3] P. Broadwell, U. Patel, and M. Tekgürler. Multilingual handwritten text recognition (HTR) models for large-scale processing of archival documents in low-resourced Arabic-script languages. In *Social Science Research Network (SSRN)*, 2025.

[4] G. Casale. *Prisoner of the Infidels: The Memoirs of Osman Agha of Timişoara*. University of California Press, 2021.

[5] A. Chan, A. Mijar, M. Saeed, C.-W. Wong, and A. Khater. Hatformer: Historic handwritten arabic text recognition with transformers, 2024.

[6] U. Koç. *Bir Osmanlı Türk askerinin maceralı esirlik hikayesi: Temeşvarlı Osman Ağa'nın Esaretnâmesi'nin orijinal ve sadeleştirilmemiş Latin harfleriyle transkripsiyonu*. Unknown, Istanbul, 2020.

[7] R. F. Kreutel. *Osman Ağa: Die Gefangenschaft Osman Ağas und seine Rückkehr aus der Gefangenschaft*. St. Augustin: Klaus Schwarz Verlag, 1980.

[8] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[9] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. Trocr: transformer-based optical character recognition with pre-trained models. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[10] M. Saeed, A. Chan, A. Mijar, j. Moukarzel, G. Habchi, C. Younes, a. elias, C.-W. Wong, and A. Khater. Muharaf: Manuscripts of handwritten arabic dataset for cursive text recognition. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 58525–58538. Curran Associates, Inc., 2024.

[11] C. Wigington, C. Tensmeyer, B. Davis, W. Barrett, B. Price, and S. Cohen. Start, follow, read: End-to-end full-page handwriting recognition. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VI*, page 372–388, Berlin, Heidelberg, 2018. Springer-Verlag.