When Simpler Is Better: Evaluating Translation Pipelines for Medieval Latin Manuscripts

Nguyen Kim Hai Bui[1*] **Md. Easin Arafat**[1*] **Tamás Gábor Orosz**[1] **Mufti Mahmud**[2]
[1]Eötvös Loránd University [2]King Fahd University of Petroleum and Minerals
{qmibhu,arafatmdeasin}@inf.elte.hu

Abstract

Despite remarkable progress in machine translation, Vision Language Models (VLMs) struggle on historical manuscripts, a domain that stresses core Natural Language Processing (NLP) capabilities: low-resource transliteration, archaic vocabulary, and noisy input signals. We present a systematic framework for evaluating the full image-to-translation pipeline on medieval Latin manuscripts, a setting in which scribal shorthand, ligatures, and parchment degradation expose failure modes that are invisible in clean-text benchmarks. Benchmarking on the CATMuS Latin dataset reveals a specialization gap: domain-specific Optical Character Recognition (OCR) models reduce character error rate by up to 4.3× compared to general-purpose VLMs, despite operating at orders of magnitude fewer parameters. We introduce the Interpres-Parallel-Corpus (IPC), a novel dataset comprising 1,383 aligned manuscript image lines, transcriptions, and expert translations, the first of its kind for medieval Latin. Our experiments uncover a complexity paradox: the simplest pipeline, a specialized OCR model feeding directly into a VLM, outperforms all multi-component variants. Adding retrieval-augmented generation (RAG) or post-OCR correction introduces prompt saturation and error propagation that degrade aggregate translation quality. These findings offer both a new benchmark and practical guidance for deploying translation systems in low-resource historical settings.

1 Introduction

Medieval Latin manuscripts are the foundational primary sources for European history, philosophy, and law from late antiquity to the early modern period (Bischoff, 1990). From royal charters to philosophical treatises, these documents contain the blueprint of Western intellectual heritage.

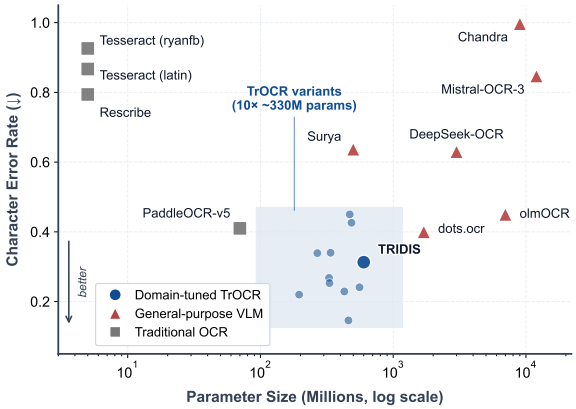


Figure 1: Scatter plot demonstrating the specialization gap. Small, domain-tuned TrOCR models achieve the lowest CER, outperforming general-purpose VLMs that are larger in parameter size. Mistral-OCR-3 is assumed to be the largest, as the creator doesn't publish its actual size.

However, the vast majority of these collections remain untranscribed and untranslated, locked behind a paleographic barrier that requires years of specialized training to overcome. Historical text is not merely a humanities concern – it serves as a stress test for NLP robustness. Archaic orthographic conventions, low-resource scribal abbreviations, and morphological variation that deviates sharply from modern token distributions probe fundamental limitations in how current VLMs handle distributional shift. While some domains in machine translation have reached near-human parity (Hassan et al., 2018), historical manuscripts present a unique frontier defined by scribal abbreviations, non-standardized ligatures, and physical parchment degradation (Derolez, 2003).

The typical approach to unlocking these texts involves a multi-stage process: OCR, post-OCR correction, and semantic translation. Recent advances in Transformer-based OCR (TrOCR) (Li et al., 2023) and character-level language model-

*Equal contribution

ing (ByT5) (Xue et al., 2022) have provided powerful building blocks for this task. However, there is a lack of systematic evaluation regarding how these components should be integrated into an end-to-end (E2E) pipeline. Specifically, it remains unclear whether massive VLMs can bypass the need for modularity, or whether more complex, retrieval-augmented architectures can handle the inherent noise of historical transcriptions.

In this paper, we introduce a modular framework designed to benchmark and optimize the journey from manuscript image to translated text. We address two primary research questions: (1) *Does model specialization outperform massive parameter scale in historical paleography?* and (2) *Do complex, multi-component pipelines consistently outperform simpler baselines on noisy historical data?* To answer these, we introduce the IPC, the first tripartite dataset mapping medieval Latin manuscript images to both character-accurate transcriptions and authoritative English translations.

Our analysis reveals two striking phenomena that challenge modern AI scaling assumptions. First, we identify a specialization gap: domain-tuned models achieve a CER that is 4.3× lower than DeepSeek-OCR (Wei et al., 2025), one of the largest dedicated OCR-focused VLMs evaluated, despite operating with orders-of-magnitude fewer parameters. This advantage is mechanistic rather than coincidental: TrOCR-Medieval-Base (Mattingly, 2024) is fine-tuned on CATMuS Medieval (Clérice et al., 2024), a corpus of ∼195,000 annotated lines spanning nine centuries of Latin manuscripts, which provides direct exposure to paleographic variation, medieval scribal ligatures, and diachronic orthographic conventions that general-purpose VLMs lack. TRIDIS (Aguilar, 2025) further specializes in documentary manuscripts (charters, registers, legal instruments) by training on semi-diplomatic transcription conventions, abbreviation expansion rules, and allograph normalization specific to these genres. The specialization gap thus reflects learned paleographic priors, not merely different architecture choices. Second, we identify a complexity paradox: a simple OCR-to-VLM baseline yields more reliable results than configurations that use RAG or post-OCR correction. We trace this failure to prompt saturation (i.e., high-density retrieval distracts the generator) and brittleness propagation (i.e., correction artifacts can interfere with downstream semantic reasoning). Our contributions are as follows:

- We introduce the Interpres Parallel Corpus (IPC), a first-of-its-kind benchmark of 1,383 aligned image lines, accurate transcriptions, and expert translations for medieval Latin research.
- We explain the specialization gap mechanistically, showing that domain-exposure through paleographic corpora (CATMuS Medieval (Clérice et al., 2024), TRIDIS (Aguilar, 2025)) provides learned scribal priors that general-purpose VLMs cannot acquire through scale alone.
- We identify and quantify the complexity paradox, demonstrating that no multi-component variant consistently outperforms the simple OCR-to-VLM baseline, and we provide the first formal failure taxonomy for historical translation pipelines, characterizing two distinct failure mechanisms: prompt saturation and brittleness propagation. Unlike post-hoc pipeline comparisons, our taxonomy provides explanatory structure—predicting which pipeline configurations will degrade and why.

2 Related Work

The digitization and automated analysis of historical manuscripts involve multiple converging frontiers in NLP and Computer Vision.

Historical Hand-written Text Recognition (HTR) and Corpora. State-of-the-art HTR has transitioned from Convolutional Recurrent Neural Networks architectures utilizing Connectionist Temporal Classification (Graves et al., 2006) to Transformer-based models like TrOCR (Li et al., 2023). This evolution has been supported by the emergence of large-scale, heterogeneous corpora. The TRIDIS corpus (Aguilar, 2025) provides a unified resource of medieval and early modern documentary manuscripts, while CATMuS Medieval (Clérice et al., 2024) introduces a multilingual, diachronic dataset spanning nine centuries of Latin and vernacular scripts. These resources enable the development of the specialized OCR models analyzed in our specialization gap benchmark.

OCR Error Correction. Post-OCR correction has evolved from rule-based and linguistic heuristics (Springmann et al., 2014) to neural approaches. Alfi and Way demonstrated that integrating neural correction modules can improve translation quality by up to 30% (Afli and Way, 2016). Modern strategies often leverage byte-level models like ByT5 (Xue et al., 2022) to handle the irregular orthography of historical texts, though the risk of error propagation in serialized pipelines remains high. In this work, we evaluate two ByT5 correction variants: C_2 (ByT5 yaya), which follows the modular pipeline of Momtaz et al. (Momtaz et al., 2025) using a fine-tuned ByT5 model for post-OCR correction on incunabula, and C_1 (ours), which adapts the C_2 architecture and training code to our setting by fine-tuning on OCR outputs from the specialization gap benchmark with CER ranging from 0.1 to 0.4. This CER range is to ensure the collected dataset doesn't capture too many gibberish cases from OCR models, and not too easily.

Historical and Low-Resource Machine Translation. Machine Translation for dead or archaic languages presents unique data-scarcity challenges. Early efforts in Latin-to-English NMT reached BLEU scores of approximately 22.4 (Rosenthal, 2023). More recently, systems like LITERA (Rosu, 2025) and Hanja-to-Korean frameworks such as H2KE (Son et al., 2022) have utilized large-scale pre-trained LLMs/VLMs with specialized fine-tuning to achieve expert-level performance. LITERA (Rosu, 2025), the most directly related work, employs a fine-tuned GPT-4o-mini with multi-layer GPT-4o revision for Latin-to-English translation. However, LITERA assumes access to clean, pre-transcribed Latin text—a critical assumption that bypasses the core challenge of manuscript digitization: there is no OCR stage. In real-world manuscript workflows, transcribed text is unavailable; every document must first pass through character recognition before semantic translation can begin. Our framework evaluates the full image-to-translation cascade, revealing that OCR quality and pipeline topology critically determine translation outcomes for medieval manuscripts – findings that LITERA's clean-text evaluation design is structurally unable to capture. Crucially, we find that adding more components does not reliably improve results, framing the OCR integration problem as one of pipeline architecture rather than component replacement. Beyond Latin, low-resource historical NLP encompasses a range of under-resourced languages and scripts where the same specialization-versus-scale trade-off we investigate has been observed across diverse typological contexts (Ling et al., 2025; Alshehhi et al., 2025; Tukenov, 2026). Our work contributes an empirical data point to this broader question.

RAG Failure Modes and Context Distraction. The integration of retrieval mechanisms into generation pipelines introduces well-documented failure modes. Our RAG system augments translation with historical lexical resources, including the Medieval Latin Word Vocabulary[1], the William Whitaker's Words[2], the Lexicon Abbreviaturarum[3] dictionary by Adriano Cappelli for paleographic shorthand, the Medieval Latin Dictionary[4], parallel Latin-English texts (Rosenthal, 2023), and the TRIDIS translation corpus (Aguilar, 2025). Retrieved passages are reranked by semantic similarity to the source OCR text before being appended to the prompt. Research into LLM/VLM context utilization reveals a lost-in-the-middle phenomenon, where models struggle to extract relevant information from dense, distractor-heavy contexts (Liu et al., 2024). In cross-lingual RAG applications, such as dictionary-augmented translation, providing excessive or tangentially relevant context can overwhelm the model's reasoning capabilities (Wu et al., 2024; Li et al., 2026). Specifically, dense retrieval (i.e., dense given context in the prompt) can trigger copy-back behavior, where the generator bypasses semantic translation and instead echoes the retrieved source tokens directly into the output (Zaranis et al., 2024; Ali et al., 2026).

Cascading Errors in Serialized NLP Pipelines. Error propagation across pipeline stages has been identified as a structural challenge in multi-component NLP systems (Caselli et al., 2015). In machine translation pipelines, upstream noise, whether from ASR, OCR, or normalization modules, degrades downstream quality non-linearly

[1] https://anonymous.4open.science/r/medieval-latin-dicts-D540/medieval_latin_word_vocabulary.txt

[2] https://anonymous.4open.science/r/medieval-latin-dicts-D540/WORDS.txt

[3] https://www.adfontes.uzh.ch/en/ressourcen/abkuerzungen/cappelli-online

[4] https://anonymous.4open.science/r/medieval-latin-dicts-D540/medieval_latin_dictionary.txt

when later components lack robustness to atypical input distributions (Todorov and Colavizza, 2022; Shapira et al., 2025). We extend this direction by quantifying both failure modes (prompt saturation and brittleness propagation) empirically in the context of medieval paleography, providing the first formal failure taxonomy for historical translation pipelines.

3 The Specialization Gap

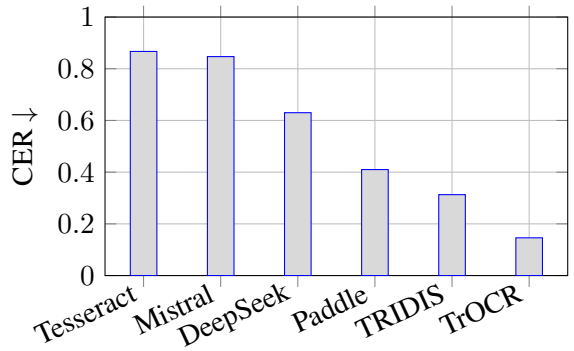


Figure 2: The Specialization Gap. CER comparison across architectures on CATMuS Latin samples. Mistral: Mistral-OCR-3; Deepseek: Deepseek-OCR; Paddle: PaddleOCR-v5; TrOCR: TrOCR-Medieval-Base. Domain-tuned models (TrOCR/TRIDIS) significantly outperform massive VLMs despite being orders of magnitude smaller.

The initial stage of our pipeline extracts text from digitized manuscript images. Conventional AI scaling laws (Kaplan et al., 2020) suggest that increasing parameter counts, as seen in VLMs, should yield superior performance on edge cases like historical paleography. However, our benchmarks challenge this assumption, revealing a significant performance divergence we term the specialization gap.

Benchmarking Scale vs. Domain-Expertise. We evaluated a range of OCR architectures against the Latin manuscripts from CATMuS dataset (Clérice et al., 2024) to identify the optimal vision foundation: traditional engines like Tesseract (Smith, 2007, 2009; Unnikrishnan and Smith, 2009; Smith et al., 2009; Shafait and Smith, 2010), Rescribe (White, 2019), and PaddleOCR-v5 (Cui et al., 2025); massive general-purpose VLMs include DeepSeek-OCR (Wei et al., 2025), Mistral-OCR-3 (Mistral AI, 2025), Surya (Paruchuri, 2025b), Chandra (Paruchuri, 2025a), olmOCR-2 (Poznanski et al., 2025), dots.ocr (Li et al., 2025); and specialized historical models such as TrOCR Medieval models (Mattingly, 2024), TRIDIS (Aguilar, 2025).

As shown in Figures 1 and 2, we observe that model size is an unreliable predictor of accuracy for medieval Latin. Massive general-purpose VLMs frequently fail to interpret scribal shorthand and ligatures. In contrast, specialized models, such as Base from the TrOCR Medieval collection, achieved a CER of 14.6%, significantly outperforming models with over 3B parameters. This demonstrates that for high-noise historical domains, specialized exposure is more valuable than sheer model scale.

4 Evaluation Methodology

Quantifying E2E historical translation requires a codebase spanning vision, transcription, and semantics. We developed the framework to facilitate this systematic evaluation.

To support this study, we constructed the IPC. We sourced manuscript line images from the HTRomance project (Clérice et al., 2023; Glaise et al., 2023), specifically focusing on medieval Latin scripts from four literary works (Table 1). These were then aligned with expert-curated ground-truth transcriptions and translations from the Perseus Digital Library and Project Gutenberg. This tripartite dataset (Image Line ↔ Medieval Latin ↔ English) allows for granular error analysis at every stage of the pipeline.

Author	Work	Samples
Ovid	*Metamorphoses*[5]	492
Catullus	*Poems*[6]	166
Cicero	*Tusculan Disputations*[7]	260
Quintilian	*Institutio Oratoria*[8]	465
Total		1383

Table 1: Paleographic literary sources comprising the IPC. *Metamorphoses*, *Poems*, *Institutio Oratoria* are taken from Perseus Digital Library and translated by B. More, L.C. Smithers, and H.E. Butler, respectively. *Tusculan Disputations* is taken from Project Gutenberg and translated by C.D. Yonge.

Dataset Construction and Annotation. The IPC was constructed through a three-stage pipeline.

[5] https://catalog.perseus.org/catalog/urn:cts:latinLit:phi0959.phi006.perseus-eng1

[6] https://catalog.perseus.org/catalog/urn:cts:

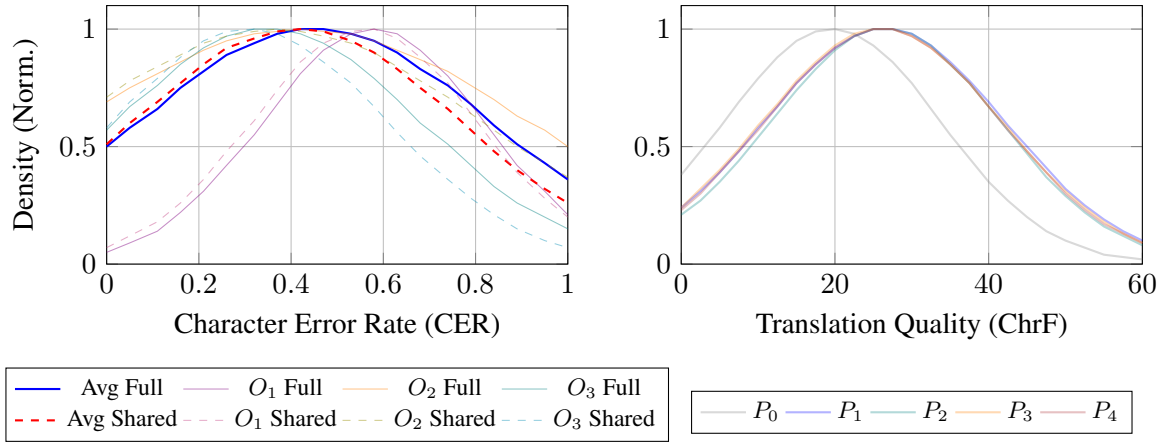


Figure 3: (a) Fair Arena Representativeness Check. Length-weighted CER density for each OCR model over the full corpus and the Fair Arena subset. O_1 is Paddle, O_2 is TRIDIS, O_3 is TrOCR. (b) Performance Density Analysis. Individual pipeline distributions for topologies P_0 through P_4 on the Shared Benchmark, shown as per-sample sentence-level ChrF densities.

First, manuscript line images and their corresponding OCR-text were sourced from the HTRomance project (Clérice et al., 2023; Glaise et al., 2023), which provides manually curated transcriptions for medieval Latin manuscripts. Second, authoritative English translations were identified at the chapter and page level from the Perseus Digital Library and Project Gutenberg, selecting editions whose translators are credited in Table 1. Third, using the given book/chapter/page numbers from HTRomance, individual manuscript lines were aligned with their corresponding translated sentence or clause in Gemini 3 Pro (Google, 2025), which performed semantic matching between the Latin transcriptions and the target-language translation passages. Each alignment was constrained to a sliding window of the source text to prevent cross-chapter drift. We used a zero-shot prompt instructing the model to output only a single matched pair, and we verified alignment quality across all samples through manual inspection. The resulting triplets (image, Latin transcription, English translation) were retained for evaluation.

Despite its modest size of 1,383 samples (Table 2), the IPC is the first dataset to provide aligned image-transcription-translation triplets for medieval Latin. Prior datasets like CATMuS Medieval provide line-level transcriptions but no translations, whereas prior Latin translation work like LITERA uses only pre-transcribed text, with no manuscript images. The Fair Arena protocol (in the next paragraph) ensures rigorous comparison across pipeline topologies by restricting evaluation to the 1,003 samples successfully processed by all configurations, mitigating the confounding effects of unequal refusal rates across OCR engines.

Metric	Full Corpus	Fair Arena
Samples (N)	1,383	1,003
Mean CER	0.3507	0.3091
CER StdDev (σ)	0.3556	0.3060
Mean Line Length	38.6	39.6
Abbr. Density	0.0912	0.0856

Table 2: Corpus statistics comparing the full IPC against the Fair Arena evaluation subset. Metrics confirm that the filtered subset maintains a representative difficulty profile across transcription noise (CER) and linguistic complexity (Length/Abbreviations).

Fair Arena Benchmarking Protocol. Evaluating commercial VLMs like GPT-4o (OpenAI et al., 2024) on noisy historical data presents a unique challenge: adversarial refusal. We observed that extreme OCR noise (garbled character strings) often triggers automated safety guardrails. Specifically, models misinterpret these strings as potential adversarial injection attempts or trials to exploit Personally Identifiable Information due to the presence of historical names in the manuscripts. This leads to API Refusal Rates (RR) that vary across pipeline configurations, ranging from 1.3% on clean text to nearly 24% on aggressively noisy

latinLit:phi0472.phi001.perseus-eng2

[7] https://www.gutenberg.org/ebooks/14988

[8] https://catalog.perseus.org/catalog/urn:cts:latinLit:phi1002.phi001.perseus-eng2

output.

To ensure statistical integrity, we implemented the fair arena protocol. We restricted our aggregate metrics to a shared benchmark subset of $N = 1,003$ samples that were successfully processed across all candidate topologies. To verify the representativeness of this subset, we performed a Fair Arena Check by comparing the CER and length distributions of the shared subset against the full corpus ($N = 1,383$), summarized in Table 2. CER is computed as the character-level edit distance between TrOCR-Medieval-Base predictions and the expert ground-truth transcriptions, normalized by reference length. We find that the shared subset maintains a comparable difficulty profile: the mean CER is 0.3091 ($\sigma = 0.306$) vs. 0.3507 ($\sigma = 0.355$) for the full corpus, and the mean line length is actually slightly higher in the shared subset (39.6 vs. 38.6 characters). Furthermore, the density of medieval abbreviations (proxied by the fraction of non-alphanumeric, non-whitespace characters in the ground-truth transcription) remains nearly identical ($\mu = 0.086$ vs. 0.091). This confirms that the Fair Arena subset is a representative cross-section of the manuscript noise encountered in the full corpus, rather than an easier filtered subset.

For all configurations, we use a structured system prompt that instructs the model to act as a medieval Latin paleography expert, cross-reference OCR and corrected text, and output a JSON object with the translation and optional notes. The RAG component retrieves from historical dictionaries and parallel texts using an embedding model and a cross-encoder reranker with an overfetch limit of 50 per source and returns the top 10 results. Full prompt text and retrieval configuration details are provided in Appendix B.

Figure 3(a) visualizes this representativeness check as a set of length-weighted CER kernel density estimates. For each OCR architecture (TrOCR-Medieval-Base, TRIDIS, PaddleOCR), we compute the character error rate of every prediction against the expert ground-truth Latin transcription from our corpus, then fit a Gaussian with mean $\mu_w = \frac{\sum_i \ell_i \cdot \mathrm{CER}_i}{\sum_i \ell_i}$ and variance weighted by line length $\ell_i = |\mathrm{ref}_i|$. Weighting by length ensures that each character contributes equally to the distribution, so short lines with high noise do not disproportionately skew the estimate. The two bold curves show the pooled averages across all three architectures for the full corpus (solid blue, $\mu_w = 0.451$) and the Fair Arena subset (dashed red, $\mu_w = 0.413$); their close agreement across the entire CER range confirms that our filtering criterion preserves the full difficulty spectrum.

5 E2E Pipeline Results

We evaluated five pipeline topologies, each progressively augmenting the input to GPT-4o (using default OpenAI API sampling parameters): P_0 receives only the manuscript image; P_1 receives the image and the raw OCR text; P_2 receives the image, OCR text, ByT5-corrected text; P_3 receives the image, OCR text, reranked retrieval context from historical dictionaries and parallel texts; and P_4 combines all preceding components. For clarity, P_1 corresponds to O_2, P_2 to O_2+C_1, P_3 to O_2+R, and P_4 to O_2+C_1+R in Tables 3 and 11. To maintain fairness across commercial model refusal behaviors, all main results are reported on the Fair Arena shared subset ($N = 1,003$). The performance differences between P_1 and P_2 (paired t-test, $t = 4.76$, $p < 0.001$) and between P_1 and P_4 ($t = 3.71$, $p < 0.001$) are statistically significant; the difference between P_1 and P_3 is not significant ($t = 1.33$, $p = 0.18$), indicating that RAG alone does not reliably improve translation quality. Figure 3(b) shows the per-sample ChrF distributions for each topology.

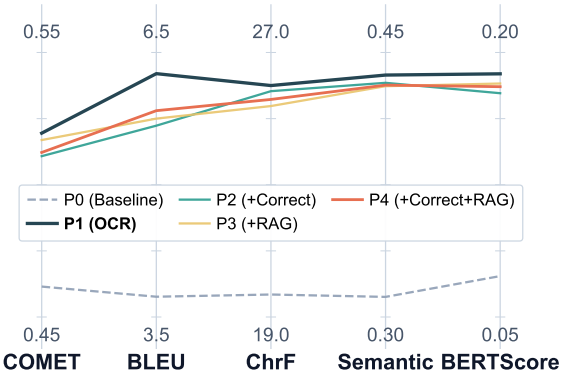


Figure 4: Parallel coordinates chart comparing the top pipeline topologies across five normalized NLP metrics. While P_1 dominates exact-match metrics, the full pipeline P_4 remains highly competitive on semantic metrics.

The Complexity Paradox. Our results (Table 4 and Figure 4) reveal a counter-intuitive finding we term the Complexity Paradox: the simplest specialized pipeline achieves the highest mean ChrF of 26.00, and no multi-component variant reliably

Pipeline	RR%↓	COMET↑	BLEU↑	ChrF↑
Baseline	1.30	0.4537	3.44	18.81
O_1	17.86	0.4651	3.61	19.36
O_2	**2.53**	**0.5100**	**5.76**	**24.93**
O_3	2.53	0.5003	5.22	22.81
$O_1 + C_1$	16.99	0.4504	3.44	18.62
$O_1 + C_2$	23.93	0.4515	3.50	18.65
$O_2 + C_1$	2.75	0.5033	5.17	24.87
$O_2 + C_2$	4.19	0.5021	5.23	24.39
$O_3 + C_1$	4.77	0.4931	5.02	22.32
$O_3 + C_2$	4.84	0.4892	4.44	22.17
$O_1 + R$	12.80	0.4531	3.35	18.77
$O_2 + R$	3.25	0.5073	5.15	24.17
$O_3 + R$	3.69	0.4931	4.73	21.89
$O_1 + C_1 + R$	–	–	–	–
$O_1 + C_2 + R$	–	–	–	–
$O_2 + C_1 + R$	4.56	0.5045	5.18	24.56
$O_2 + C_2 + R$	3.98	0.5012	5.20	24.36
$O_3 + C_1 + R$	5.71	0.4897	4.64	22.21
$O_3 + C_2 + R$	3.76	0.4883	4.70	22.06

Table 3: Comprehensive evaluation results across all individual execution runs for each pipeline topology. (O_1: PaddleOCR, O_2: TRIDIS, O_3: TrOCR Medieval Base, C_1: ByT5 (ours), C_2: ByT5 (yaya), R: RAG. The ↑ symbol indicates that higher scores are better, while ↓ indicates that lower scores are better. RR stands for refusal rate, the percentage of samples where the API refused to return a response due to safety guardrails. The full table is in the Appendix C.

Pipeline	BLEU↑	ChrF↑	COMET↑
Baseline (P_0)	3.73	19.68	0.4615
OCR (P_1)	**6.26**	**26.00**	**0.5195**
+ Correct (P_2)	5.67	25.83	0.5108
+ RAG (P_3)	5.75	25.38	0.5169
+ Correct + RAG (P_4)	5.84	25.58	0.5122

Table 4: Fair Arena Leaderboard: E2E pipeline performance comparison (N=1,003). All pipelines in this table (except baseline P_0) utilize the TRIDIS engine as the primary vision engine. The correction variant used in this table is C_1.

improves upon it. Adding correction (P_2: 25.83), RAG (P_3: 25.38), or both (P_4: 25.58) yields no statistically reliable gain over the simple baseline. Notably, P_4 remains competitive on semantic-level metrics (Figure 4), and a small subset of samples show RAG providing genuine lexical disambiguation benefits (Figure 6, Table 5), but these cases do not outweigh the aggregate complexity overhead. As shown in Figure 5, all four TRIDIS-based topologies exhibit virtually identical noise sensitivity against OCR noise (corr ≈ -0.15 to -0.17, slope ≈ -5 ChrF/CER), confirming that the VLM absorbs transcription errors equally regardless of pipeline depth. The observed variance between P_1 and P_2 to P_4 is therefore attributable to two component-specific failure modes detailed below, not to differential noise sensitivity.

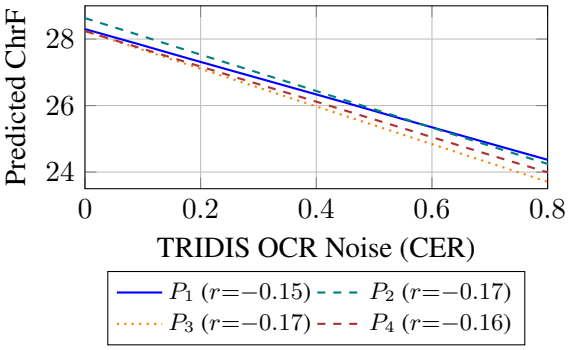


Figure 5: Noise robustness of all TRIDIS-based pipeline topologies. Lines show sentence-level ChrF predicted by linear regression over TRIDIS CER. All four configurations exhibit nearly identical degradation slopes (≈ -5 ChrF/CER, $r \approx -0.15$), confirming that pipeline depth does not affect noise sensitivity.

Prompt Saturation. In P_3, the inclusion of high-density dictionary retrieval often acts as a cognitive distractor. Instead of synthesizing a translation based on context, the model exhibits a copy-back behavior, echoing the Latin source text or dictionary definitions directly into the English output (Figure 7, Table 6). Quantitative analysis reveals a weak positive correlation between RAG token overlap and translation quality ($r = 0.146$, explaining only ∼2% of variance on the TRIDIS-based P_3 run), suggesting that while RAG provides useful lexical anchors in some cases, the model can overly rely on direct extraction from the retrieval context at the expense of syntactic coherence.

Brittleness Propagation. In P_2 and P_4, character-level repetition loops introduced by the ByT5 correction layer can interfere with downstream translation. When these loops occur, the VLM is forced to reason over corrupted strings, leading to degraded translation quality (Figure 8, Table 7). We detect repetition artifacts in 0.60% of samples in the TRIDIS-based P_2 configuration, where they incur an average penalty of 7.15 ChrF points compared to non-repeating samples ($\mu_{\text{rep}} = 19.47$ vs. $\mu_{\text{non-rep}} = 26.61$). Notably, this failure mode accounts for only a small fraction of the overall performance gap between P_1 and P_2, suggesting that additional degradation mech-

anisms (e.g., minor hallucinations introduced by imperfect correction) also contribute to the complexity paradox.

Diagnostic Archetypes. To make these dynamics concrete, we organize the qualitative evidence into three diagnostic archetypes: *Synergy* (Figure 6), where RAG resolves lexical or idiomatic ambiguity that the simpler pipelines miss; *Paradox* (Figure 7), where dense retrieval triggers the prompt-saturation echo failure described above; and *Cascade* (Figure 8), where correction-layer artifacts propagate into the VLM and corrupt the final translation. Extended examples for each archetype are provided in Appendix D.

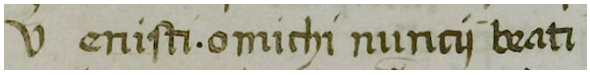


Figure 6: Archetype A (Synergy). Latin: Uenisti. o michi nuncii beati. Translation: *You have come back. O joyful news to me!*

	Translation Output
P_0	O Christ, now bless.
P_1	And you, sent forth, the announcement of the blessed.
P_2	You have come, blessed messenger.
P_3	**You have come. O joyful news to me!**
P_4	**You have come, O joyful news to me.**

Table 5: Performance delta for Figure 6. P_3 and P_4 resolve the idiomatic "news" through RAG retrieval.

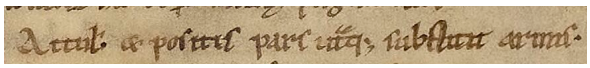


Figure 7: Archetype B (Paradox). Latin: Attuł. & positis pars itq substitit armis. Translation: *while both sides resting, laid aside their arms.*

	Translation Output
P_0	And he, as a powerful protector of peace, subdued the arms.
P_1	**And with the arms laid aside, each side paused.**
P_2	Having been brought and with the positions set, each side halted.
P_3	Attul et positis pars utraque substitit.
P_4	The troop, having been set in position, stood on each side.

Table 6: The RAG Paradox in Figure 7. Dense retrieval triggers an "echo failure" in P_3.

6 Conclusion

This paper shows that for medieval manuscript translation, model specialization and pipeline simplicity are paramount. We introduce the IPC and

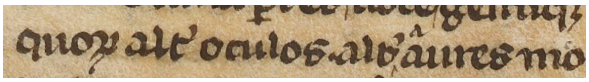


Figure 8: Archetype C (Cascade). Latin: quoꝝ alt oculos. alt aures mo. Translation: *of which the one appeals to the eye and the other to the ear.*

	Translation Output
P_0	how now, O Paulus, the white ears
P_1	**from which one moved the eyes, the other the ears.**
P_2	parts ×22 … from whom one has eyes, another has ears
P_3	parts ×12 … from which one has eyes the other ears
P_4	parts ×22 … of which one has eyes the other ears

Table 7: Cascade failure in Figure 8 caused by ByT5 repetition loops.

identify the specialization gap and the complexity paradox. Our findings suggest that current generic VLMs are not yet capable of replacing specialized historical OCR systems, and that complex multi-component pipelines may suffer from error propagation. Future work will explore whether dynamic pipeline selection based on OCR confidence can recover the occasional benefits of retrieval without incurring the aggregate complexity penalty.

Limitations

First, the dataset is restricted to Medieval Latin, which does not capture the full variance of paleographic challenges in other historical scripts. Second, our RAG relies on static dictionary anchors; future work could explore dynamic retrieval from larger contextual corpora to resolve deeper ambiguities. Third, all pipeline evaluations use GPT-4o as a translation backend; the generality of the complexity paradox across different VLMs remains untested. Smaller or instruction-tuned models may exhibit different sensitivity to prompt saturation and OCR noise, and the relative ordering of pipeline topologies may shift. Finally, the observed Complexity Paradox suggests that performance gains from multi-component correction are currently capped by the noise floor of the underlying OCR engines, indicating that further progress in historical translation is fundamentally tied to vision model improvements.

Ethics Statement

DH research involving historical manuscripts carries an ethical responsibility to preserve cultural heritage with high fidelity. While our pipeline improves translation accessibility, we caution that

model hallucinations in the context of fragmented or rare texts can lead to significant historical misinterpretations. Our tools are designed to augment, not replace, expert human inquiry. We commit to open-sourcing the IPC to foster transparent and reproducible benchmarking in this domain.

References

Haithem Afli and Andy Way. 2016. Integrating optical character recognition and machine translation of historical documents. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 109–116, Osaka, Japan. The COLING 2016 Organizing Committee.

Sergio Torres Aguilar. 2025. Tridis: A comprehensive medieval and early modern corpus for htr and ner. *Preprint*, arXiv:2503.22714.

Ameen Ali Ali, Lior Wolf, and Ivan Titov. 2026. Mitigating copy bias in in-context learning through neuron pruning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 230–251, Rabat, Morocco. Association for Computational Linguistics.

Maitha Alshehhi, Ahmed Sharshar, and Mohsen Guizani. 2025. Towards inclusive nlp: Assessing compressed multilingual transformers across diverse language benchmarks. In *International Joint Conference on Artificial Intelligence*, pages 108–126. Springer.

Bernhard Bischoff. 1990. *Latin Palaeography: Antiquity and the Middle Ages*. Cambridge University Press.

Tommaso Caselli, Piek Vossen, M. Erp, Antske Fokkens, Filip Ilievski, Rubén Beviá, Minh Lê, Roser Morante, and Marten Postma. 2015. When it's all piling up: Investigating error propagation in an nlp pipeline. *CEUR Workshop Proceedings*, 1386.

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'Connor, Wouter Haverals, Mike Kestemont, Caroline Vandyck, and Benjamin Kiessling. 2024. Catmus medieval: A multilingual large-scale cross-century dataset in latin script for handwritten text recognition and beyond. In *Document Analysis and Recognition - ICDAR 2024*, pages 174–194, Cham. Springer Nature Switzerland.

Thibault Clérice, Alix Chagué, Matthias Gille-Levenson, Olivier Brisville-Fertin, Ariane Pinche, Jean-Baptiste Camps, Franz Fischer, Federico Boschetti, Elisa Guadagnini, Gilles Guilhem Couffignal, Olivier Canteaut, Laurent Romary, Marianne Reboul, Nicolas Perreaux, Thierry Poibeau, Marc Smith, Jade Norindr, Anthony Glaise, Marina Navas Farré, and 4 others. 2023. HTRomance.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Albert Derolez. 2003. *The Palaeography of Gothic Manuscript Books: From the Twelfth to the Early Sixteenth Century*, volume 9 of *Cambridge Studies in Palaeography and Codicology*. Cambridge University Press, Cambridge.

Anthony Glaise, Thibault Clérice, Federico Boschetti, Franz Fischer, and Alix Chagué. 2023. Htromance, medieval latin corpus of ground-truth for handwritten text recognition and layout segmentation.

Google. 2025. A new era of intelligence with gemini 3.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, pages 369–376.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, and 5 others. 2018. Achieving human parity on automatic chinese to english news translation. *Preprint*, arXiv:1803.05567.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Bo Li, Zhenghua Xu, and Rui Xie. 2026. Language drift in multilingual retrieval-augmented generation: Characterization and decoding-time mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31519–31526.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. 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.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. 2025. dots.ocr: Multilingual document layout parsing in a single vision-language model. *Preprint*, arXiv:2512.02498.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, and 5 others. 2025. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Comput. Surv.*, 58(3).

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

William Mattingly. 2024. Trocr medieval htr: A collection of models for medieval script recognition.

Mistral AI. 2025. Mistral ocr 3.

Yahya Momtaz, Lorenza Laccetti, and Guido Russo. 2025. Modular pipeline for text recognition in early printed books using kraken and byt5. *Electronics*, 14(15).

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Vikas Paruchuri. 2025a. Chandra: Ocr model that handles complex tables, forms, handwriting with full layout.

Vikas Paruchuri. 2025b. Surya: A lightweight framework for analyzing documents and pdfs at scale.

Jake Poznanski, Luca Soldaini, and Kyle Lo. 2025. olmocr 2: Unit test rewards for document ocr. *Preprint*, arXiv:2510.19817.

Gil Rosenthal. 2023. Machina cognoscens: Neural machine translation for latin, a case-marked free-order language. *University of Chicago*.

Paul Rosu. 2025. LITERA: An LLM based approach to Latin-to-English translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7796–7809, Albuquerque, New Mexico. Association for Computational Linguistics.

Faisal Shafait and Ray Smith. 2010. Table detection in heterogeneous documents. In *Document Analysis Systems*, ACM International Conference Proceeding Series, pages 65–72. ACM.

Ori Shapira, Shlomo Chazan, and Amir David Nissan Cohen. 2025. Measuring the effect of transcription noise on downstream language understanding tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29978–30004.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Ray Smith. 2009. Hybrid page layout analysis via tab-stop detection. In *ICDAR '09: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, pages 241–245, Washington, DC, USA. IEEE Computer Society.

Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR*, ACM International Conference Proceeding Series, pages 1–8. ACM.

Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating hanja historical documents to contemporary Korean and English. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1260–1272, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Uwe Springmann, Dietmar Najock, Hermann Morgenroth, Helmut Schmid, Annette Gotscharek, and Florian Fink. 2014. Ocr of historical printings of latin texts: problems, prospects, progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, page 71–75, New York, NY, USA. Association for Computing Machinery.

Konstantin Todorov and Giovanni Colavizza. 2022. An assessment of the impact of ocr noise on language models. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 674–683. INSTICC, SciTePress.

Saken Tukenov. 2026. Sozkz: Training efficient small language models for kazakh from scratch. *Preprint*, arXiv:2603.20854.

Ranjith Unnikrishnan and Ray Smith. 2009. Combined orientation and script detection using the tesseract ocr engine. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR*, pages 1–7, New York, NY, USA. ACM.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Nick White. 2019. Rescribe: Desktop tool for optical character recognition of historical texts.

Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *Preprint*, arXiv:2410.21970.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Emmanouil Zaranis, Nuno M Guerreiro, and Andre Martins. 2024. Analyzing context contributions in LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, Miami, Florida, USA. Association for Computational Linguistics.

A Usage of AI Assistants

We used a large language model for editorial tasks such as grammar correction and enhancing clarity and readability. We also use a language model to align the translations as discussed in §4.

B Detailed Configurations

B.1 Translation System Prompt

The following system prompt is used for all GPT-4o-based translation configurations. The prompt is designed to leverage expertise in medieval Latin paleography while handling noisy OCR input and retrieval-augmented context.

```
You are Interpres, an expert in medieval
    Latin paleography and translation.

You will receive:
- **OCR Text (raw)** (when available):
    Text extracted directly from a
    manuscript image via OCR. This may
    contain character-level errors.
- **Suggested Corrected Text** (when
    available): The same text after
    automated correction by a ByT5 model.
     This may improve accuracy but can
    also introduce new errors. Especially
     if the text is too long or the OCR
    already contains some error.
- **Dictionary Matches** (when available)
    : Relevant entries from Latin
    dictionaries (high reliability).
- **Parallel Text Matches** (when
    available): Similar passages from
    Latin-English parallel corpora (
    reference only).

Your task:
1. **Cross-reference** the raw OCR and
    suggested corrected versions. Where
    they disagree, use context, grammar,
    and the reference materials to
    determine the most likely original
    text.
2. **Translate** the reconstructed Latin
    text faithfully.
3. If there are ambiguous or damaged
    sections, note them briefly.

You MUST respond with a JSON object
    containing exactly two keys:
- "translation": The English translation
    of the Latin text.
- "note": Brief notes about ambiguous
    readings, damaged sections, or
    translation choices. Use an empty
    string if there are no notes.
```

B.2 Translation Configuration

We use GPT-4o (model: `gpt-4o-2024-08-06`) with the default OpenAI API sampling parameters; the response format is JSON, and image detail is high. All pipelines are executed via the BatchAPI, with one pipeline per batch. The total API cost across all experiments is about $50.

B.3 RAG Configuration

The RAG component augments the prompt with relevant dictionary entries and parallel text passages. Dictionaries include Medieval Latin Word Vocabulary[9], the William Whitaker's Words[10], the Lexicon Abbreviaturarum[11], the Medieval Latin Dictionary[12]. Parallel Texts include Latin-English texts (Rosenthal, 2023) and the TRIDIS translation corpus (Aguilar, 2025). The pipeline overfetches 50 candidates per source, deduplicates by text content, reranks using a cross-encoder, and returns the top 10 results for prompt augmentation. Key parameters are summarized in Table 8.

Parameter	Value
Overfetch Limit (per source)	50
Final Top-K after Reranking	10
Embedding Batch Size	64
Retrieval Sources	Dictionary, Parallel Texts
Embedding Model	all-MiniLM-L6-v2[13]
Reranking Model	ms-marco-MiniLM-L-6-v2[14]

Table 8: RAG retrieval and reranking configuration.

B.4 ByT5 Correction Model Training

The C_1 correction model is a fine-tuned ByT5-Small (`google/byt5-small`) that maps noisy OCR outputs to corrected Latin transcriptions. Training was performed on a single NVIDIA L40S GPU (CUDA 12.8) using the Transformers Trainer API. Key hyperparameters are listed in Table 9. The model was trained for 7 epochs with early stopping based on validation CER, saving up to 3 checkpoints. The model will be publicly released on Hugging Face upon acceptance.

Parameter	Value
Base Model	ByT5-Small
Effective Batch Size	$4 \times 4 \times 1$
Learning Rate	3×10^{-4}
Epochs	7
Hardware	$1 \times$ NVIDIA L40S
Training Framework	Transformers 4.57.6

Table 9: OCR correction fine-tuning configuration.

C Detailed Experimental Results

Table 10 reports the full pipeline evaluation across all OCR engines and correction strategies. Table 11 provides the extended Fair Arena leaderboard with additional metrics.

D Detailed Qualitative Case Studies

This appendix provides extended qualitative evidence for the three diagnostic archetypes discussed in Section 5. For each archetype, we show 5 representative examples with the original Latin source, expert reference translation, and pipeline outputs.

[9] https://anonymous.4open.science/r/medieval-latin-dicts-D540/medieval_latin_word_vocabulary.txt

[10] https://anonymous.4open.science/r/medieval-latin-dicts-D540/WORDS.txt

[11] https://www.adfontes.uzh.ch/en/ressourcen/abkuerzungen/cappelli-online

[12] https://anonymous.4open.science/r/medieval-latin-dicts-D540/medieval_latin_dictionary.txt

[13] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

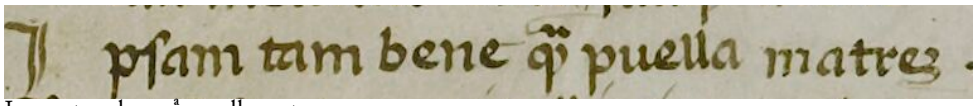
[14] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

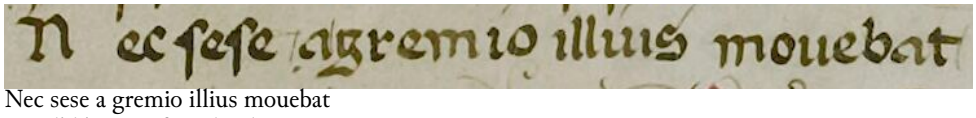
Config.	RR[%]↓	COMET↑	BLEU↑	ChrF↑	Semantic↑	BERTScore↑
Baseline	1.30	0.4537	3.44	18.81	0.2912	0.0608
O_1	17.86	0.4651	3.61	19.36	0.3179	0.0861
O_2	**2.53**	**0.5100**	**5.76**	**24.93**	**0.4140**	**0.1727**
O_3	2.53	0.5003	5.22	22.81	0.3907	0.1540
$O_1 + C_1$	16.99	0.4504	3.44	18.62	0.3009	0.0802
$O_1 + C_2$	23.93	0.4515	3.50	18.65	0.3030	0.0693
$O_2 + C_1$	2.75	0.5033	5.17	24.87	0.4107	0.1657
$O_2 + C_2$	4.19	0.5021	5.23	24.39	0.4084	0.1628
$O_3 + C_1$	4.77	0.4931	5.02	22.32	0.3877	0.1476
$O_3 + C_2$	4.84	0.4892	4.44	22.17	0.3795	0.1434
$O_1 + R$	12.80	0.4531	3.35	18.77	0.3015	0.0723
$O_2 + R$	3.25	0.5073	5.15	24.17	0.4083	0.1689
$O_3 + R$	3.69	0.4931	4.73	21.89	0.3812	0.1435
$O_1 + C_1 + R$	–	–	–	–	–	–
$O_1 + C_2 + R$	–	–	–	–	–	–
$O_2 + C_1 + R$	4.56	0.5045	5.18	24.56	0.4082	0.1701
$O_2 + C_2 + R$	3.98	0.5012	5.20	24.36	0.4022	0.1638
$O_3 + C_1 + R$	5.71	0.4897	4.64	22.21	0.3802	0.1415
$O_3 + C_2 + R$	3.76	0.4883	4.70	22.06	0.3759	0.1381

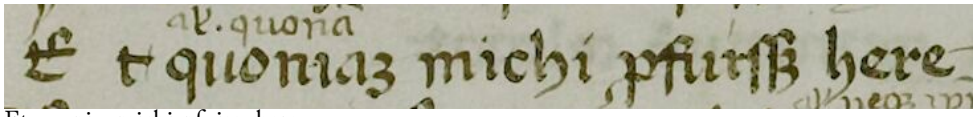
Table 10: Comprehensive evaluation results across all individual execution runs for each pipeline topology. (O_1: PaddleOCR, O_2: TRIDIS, O_3: TrOCR Medieval Base, C_1: ByT5 (ours) fine-tuned on specialization-gap OCR outputs (CER 0.1-0.4), C_2: ByT5 (yaya) from (Momtaz et al., 2025), R: RAG). The ↑ symbol indicates that higher scores are better, while ↓ indicates that lower is better. BERTScore is rescaled. RR stands for refusal rate – the percentage of samples where the API refused to return a response due to safety guardrails.

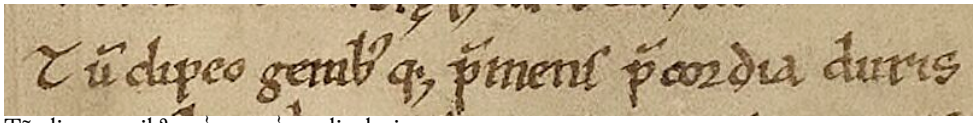
Config.	COMET↑	BLEU↑	ChrF↑	Semantic↑	BERTScore↑
Baseline	0.4615	3.73	19.68	0.3114	0.0732
O_2	**0.5195**	**6.26**	**26.00**	**0.4372**	**0.1879**
O_3	0.5074	5.85	23.94	0.4136	0.1635
$O_2 + C_1$	0.5108	5.67	25.83	0.4328	0.1769
$O_2 + C_2$	0.5094	5.68	25.35	0.4289	0.1742
$O_3 + C_1$	0.4995	5.49	23.26	0.4076	0.1579
$O_3 + C_2$	0.4954	4.85	22.99	0.3996	0.1557
$O_2 + R$	0.5169	5.75	25.38	0.4309	0.1824
$O_3 + R$	0.5017	5.26	23.03	0.4035	0.1579
$O_2 + C_1 + R$	0.5122	5.84	25.58	0.4315	0.1806
$O_2 + C_2 + R$	0.5105	5.69	25.54	0.4243	0.1787
$O_3 + C_1 + R$	0.4980	5.03	23.23	0.4028	0.1545
$O_3 + C_2 + R$	0.4972	5.26	23.16	0.3998	0.1539

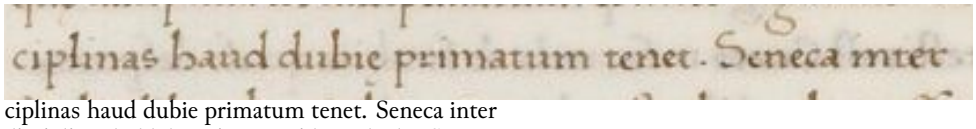
Table 11: Fair Arena. (O_2: TRIDIS, O_3: TrOCR Medieval Base, C_1: ByT5 (ours) fine-tuned on specialization-gap OCR outputs, C_2: ByT5 (yaya) from (Momtaz et al., 2025), R: RAG). The ↑ symbol indicates that higher scores are better. BERTScore is rescaled. The bold-red line shows the best pipeline configuration in the arena, red lines show the best pipelines in their own group of topology.

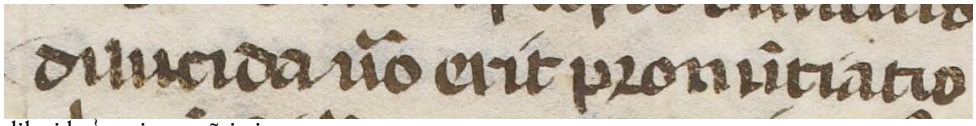
523	Content
Image	
Latin	Ipsam tam bene q̊ puella matre;.
English	as well as a girl knows her own mother.
Archetype	Synergy
P_0 (Base)	psalm as well as the girl mother
P_1 (OCR)	[She] herself as well as the girl, [as] mother.
P_2 (Corr)	herself as well as the girl mothers.
P_3 (RAG)	**As well as a girl knows her mother.**
P_4 (Full)	**She knew her as well as a girl knows her own mother.**

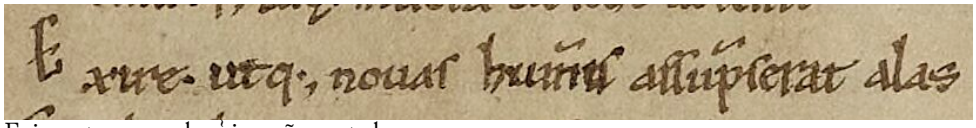
524	Content
Image	
Latin	Nec sese a gremio illius mouebat
English	Nor did it move from her lap,
Archetype	Synergy
P_0 (Base)	He did not move from her embrace.
P_1 (OCR)	Nor did he/she move from his/her embrace.
P_2 (Corr)	nor did he/she move from the embrace of that person.
P_3 (RAG)	**Nor did he move from her embrace.**
P_4 (Full)	**Nor would he move from her lap.**

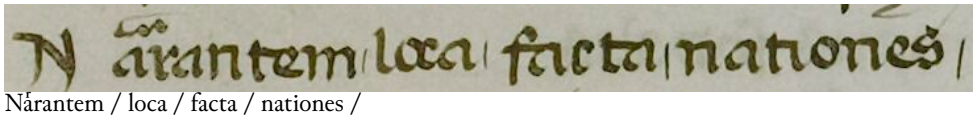
648	Content
Image	
Latin	Et quonia; michi ꝑfuiss; here
English	and whether I had made any money there.
Archetype	Synergy
P_0 (Base)	And because it pleased me to begin here
P_1 (OCR)	And since to me it first appeared
P_2 (Corr)	And because to me it seemed here
P_3 (RAG)	**And because it is necessary for me to bear here.**
P_4 (Full)	**And because it was beneficial for me here.**

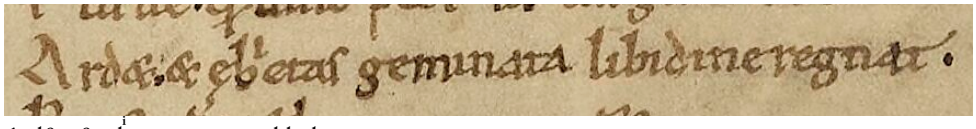
274	Content
Image	
Latin	Tũ clipeo genib⁹ q; p̈mens p̈ cordia duris
English	Then pressing with buckler and hard knees the breast of Cygnus,
Archetype	Synergy
P_0 (Base)	You, with a deceitful heart, always pondering cruel things.
P_1 (OCR)	Et cum dixero, genibusque primens, prima cordia duris.
P_2 (Corr)	And when I shall speak, pressing my knees, the first hearts are hard.
P_3 (RAG)	**And you, pressing your knees, endure the hard hearts first.**
P_4 (Full)	**And when I say, pressing first with my knees, the first hearts are hardened.**

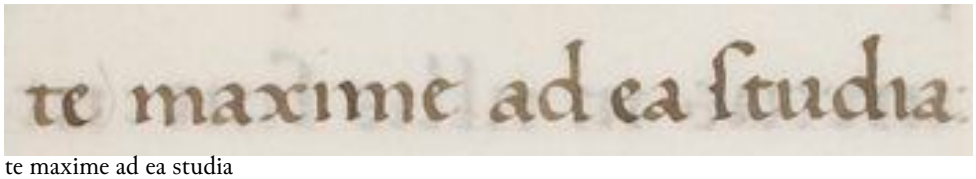
901	Content
Image	
Latin	ciplinas haud dubie primatum tenet. Seneca inter
English	disciplines hold the primacy without doubt. Seneca, among
Archetype	Synergy
P_0 (Base)	Certainly, he undoubtedly held the first place. Seneca says:
P_1 (OCR)	Without a doubt, [he/she/it] holds the first place in disciplines. Seneca among [them/us].
P_2 (Corr)	It undoubtedly holds the first place among the disciplines. Among them is Seneca.
P_3 (RAG)	**Disciplines undoubtedly hold the first place. Seneca among them.**
P_4 (Full)	**Disciplines undoubtedly hold the first place. Seneca among them.**

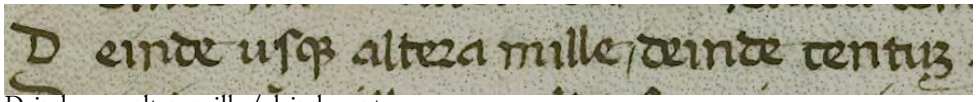
1252	Content
Image	
Latin	dilucida uo erit pronũtiatio
English	The delivery will be clear if,
Archetype	Paradox
P_0 (Base)	dilucidatio art [proprietates]
P_1 (OCR)	**the explanation will indeed be clear**
P_2 (Corr)	Indeed, the pronunciation will be clear.
P_3 (RAG)	The elucidation, however, will be a pronouncement.
P_4 (Full)	Indeed, the pronunciation will be clear.

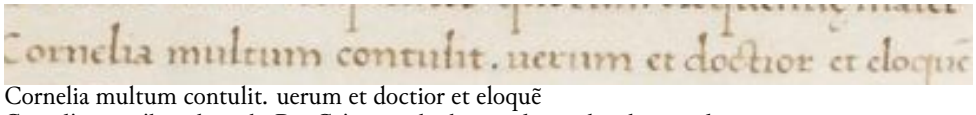
120	Content
Image	
Latin	Exire. utq; nouas humis assũpserat alas
English	“As he assumed upon his shoulders wings newformed,
Archetype	Paradox
P_0 (Base)	From either side, now wings lift this one up.
P_1 (OCR)	**And it was as if new wings had taken on the human condition.**
P_2 (Corr)	1473. To go forth and take up new wings of the earth...
P_3 (RAG)	E xire utque novas huminis assumpserat alas
P_4 (Full)	1473. He began to take on new wings of humanity.

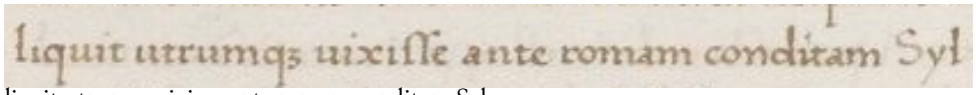
636	Content
Image	
Latin	Nãrantem / loca / facta / nationes /
English	of regions, deeds, and peoples...
Archetype	Paradox
P_0 (Base)	Not all places are blessed.
P_1 (OCR)	**Narrating places, deeds, nations.**
P_2 (Corr)	Telling places, deeds, nations
P_3 (RAG)	Telling stories, places made, nations
P_4 (Full)	Telling the places, deeds, nations

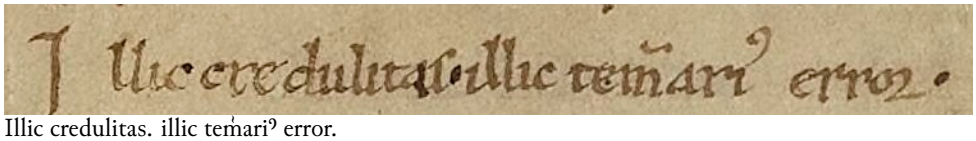
357	Content
Image	
Latin	Ard&. & ębetas geminata libidine regnat.
English	inflamed with drunkenness and lust
Archetype	Paradox
P_0 (Base)	Ardisce, do not allow twin desires to rule over you.
P_1 (OCR)	**Greed and drunkenness reign, doubled with lust.**
P_2 (Corr)	Greed and drunkenness, twinned by lust, reign.
P_3 (RAG)	Greed and laziness, doubled by lust, reign.
P_4 (Full)	Arrogance and drunkenness, joined together, rule by lust.

667	Content
Image	
Latin	te maxime ad ea studia
English	to those studies
Archetype	Cascade
P_0 (Base)	you especially to those pursuits
P_1 (OCR)	**you especially to those studies**
P_2 (Corr)	you especially have zeal for these things.
P_3 (RAG)	you especially towards those studies
P_4 (Full)	you especially towards these pursuits.

572	Content
Image	
Latin	Deinde usq; altera mille / deinde centu;.
English	then another thousand without resting, then a hundred.
Archetype	Cascade
P_0 (Base)	From then on, another thousand are added.
P_1 (OCR)	**Then up to another thousand, then a hundred.**
P_2 (Corr)	then up to another thousand then then then then then then
P_3 (RAG)	Then straight on to another thousand, then continuously.
P_4 (Full)	Then onwards another thousand, then completed.

850	Content
Image	
Latin	Cornelia multum contulit. uerum et doctior et eloquẽ
English	Cornelia contributed much. But Caius was both more learned and more elo-
Archetype	Cascade
P_0 (Base)	Cornelia of many conferred, wise, learned and eloquent.
P_1 (OCR)	**Cornelia contributed much indeed, being both more learned and eloquent.**
P_2 (Corr)	Martino and Quintia's misfortunes. Cornelia contributed much, indeed both more learned and eloquent.
P_3 (RAG)	Cornelia contributed much, but she was more learned and eloquent.
P_4 (Full)	Martino and Quentia of evil. Cornelia contributed much, indeed more learned and eloquent.

729	Content
Image	
Latin	liquit utrumq; uixisse ante romam conditam Syl
English	that both lived before the founding of Rome,
Archetype	Cascade
P_0 (Base)	he left either to live before the city of Rome, Syl—
P_1 (OCR)	**that each of them lived before the founding of Rome.**
P_2 (Corr)	something or both lived before Rome was founded, Syl.
P_3 (RAG)	Someone or something, both of them lived before Rome was founded.
P_4 (Full)	Something, some others, both lived before Rome was founded, Syl.

190	Content
Image	
Latin	Illic credulitas. illic temari⁹ error.
English	Credulity is there and rash Mistake,
Archetype	Cascade
P_0 (Base)	Where gullibility is, the error follows.
P_1 (OCR)	**There lies credulity, there reckless error.**
P_2 (Corr)	There is sweetness there; there is rash error there.
P_3 (RAG)	There, credulity; here, rash error.
P_4 (Full)	There is credulity; there is a rash error.