Computational Humanities Research

www.cambridge.org/chr

Fine-tuning large-language models for early modern Dutch translation

Gavin Lip[1], Victor de Boer[2], Arno Bosse[3] and David Grantsaan[4]

[1]Faculty of Science, Vrije Universiteit Amsterdam, Netherlands; [2]Vrije Universiteit Amsterdam, Netherlands; [3]HUC – Digital Infrastructure, KNAW, Netherlands and [4]Vrije Universiteit Amsterdam, Netherlands

Research Article

Cite this article: Lip Gavin, Victor de Boer, Arno Bosse and David Grantsaan. 2026. "Fine-tuning large-language models for early modern Dutch translation" *Computational Humanities Research*, **2**:e3, https://doi.org/10.1017/chr.2026.10022

Received: 29 January 2025
Revised: 30 October 2025
Accepted: 13 January 2026

Keywords:
digital humanities; early modern Dutch; machine translation; parameter-efficient fine-tuning

Corresponding author:
Gavin Lip;
Email: lip.gavin@gmail.com

Abstract

Large-language models (LLMs) have transformed natural language processing and opened new possibilities for the computational social sciences and digital humanities. Yet translating historical sources remains difficult because early modern varieties are scarcely represented in contemporary training corpora and because standard tokenizers fragment their non-standard orthography. This article tackles these gaps by adapting open LLMs to early modern Dutch-to-English translation and advances two concrete contributions: (i) a memory-efficient fine-tuning workflow that runs on a single consumer GPU, comparing order-reward policy optimization with the Unsloth supervised fine-tuning approach and (ii) a verifiable evaluation protocol that combines embedding-based metrics with systematic expert review. Experiments on testimonial texts (1680–1792) show that fine-tuning choice decisively shapes quality: the Unsloth-tuned Mistral model attains the highest BERTScore and METEOR values and most faithfully preserves historical nuance. The framework supports a collaborative workflow where machine-generated drafts accelerate expert translation, making archival texts more accessible while maintaining scholarly oversight through domain-expert validation.

Plain-language summary

Historical documents, such as the archives of the Dutch East India Company, contain invaluable information about the global past. However, these records are often written in older forms of languages, like Early Modern Dutch, which are difficult for most contemporary readers to understand. This research project aimed to solve this problem by training modern artificial intelligence (AI) systems, known as large-language models (LLMs), to accurately translate these historical texts into contemporary English. The goal was to create a reliable and accessible tool that could open these important archives to a broader audience of researchers, historians and the public, without requiring expertise in 17th-century Dutch.

To achieve this, the study used a technique called "fine-tuning." Rather than building an AI translator from scratch, which is incredibly resource-intensive, existing open-source AI models (like Mistral and Llama 3) were given a specialized training dataset created from translated court testimonies from the Dutch East India Company's archives in Cochin (Kerala) during 1680–1682 and 1707–1792 (van Rossum et al. 2020). Using two memory-efficient fine-tuning methods called "Unsloth" and "order-reward policy optimization (ORPO)" fine-tuning, this training was conducted on a regular consumer-grade computer, making the process more accessible for other researchers. The results of the fine-tuned models were then compared with other open-source models without specialized training, in addition to GPT-4o, a general-purpose AI that is often used as a standard for LLM performance comparisons.

Success was measured through a two-part evaluation process. First, automated scoring systems (BERTScore and METEOR) were used to compare AI-generated translations with professional human-created translation, measuring accuracy and fluency. Second, a historian specializing in Early Modern Dutch reviewed the translations to assess whether they captured the subtle nuances, legal terminology and historical tone of the original documents.

The findings clearly showed that specialized fine-tuning made a significant difference. The Mistral model fine-tuned with the Unsloth method produced the best results, consistently generating high-quality translations that were not only accurate but also preserved the historical integrity of the source texts. This approach was much more effective than the ORPO method and significantly outperformed the general-purpose models. This research demonstrates a successful and reproducible framework for adapting AI to translate low-resource historical languages, paving the way for making vast archives of historical knowledge more accessible to everyone.

Introduction

Archival records of the Dutch East India Company (Vereenigde Oostindische Compagnie [VOC]) comprise more than five million digitized pages and are recognized by UNESCO

This article was awarded Open Materials badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

as Memory of the World heritage (UNESCO 2003). These documents illuminate colonial trade, slavery, migration and environmental history across Asia, Africa and Australia, yet they are written in early modern Dutch, a historical form that only a relatively small number of specialists can now read with confidence. High-quality, verifiable English translations are therefore conducive for opening this corpus to a global community of scholars and informing public debate.

The GLOBALISE project is constructing an infrastructure that applies handwriting recognition, entity extraction and linked-data enrichment to render the VOC files searchable and machine-readable (GLOBALISE Project 2025). However, without reliable English translations, downstream analyses, such as topic modeling, spatial–temporal mapping or citizen-science annotation, remain inaccessible. Similar challenges arise in other historical translation initiatives, where translation functions as the critical bridge between disciplines, cultures and periods (Deng, Guo, and He 2023).

From a natural language processing (NLP) perspective, early modern Dutch diverges from its contemporary counterpart in orthography, syntax and lexicon, and is virtually absent from web-scale pre-training corpora. State-of-the-art large-language models (LLMs) therefore show reduced accuracy in low-resource settings because English-centric pre-training introduces lexical interference and bias. Task-specific adaptation through supervised fine-tuning (SFT) or reward-based alignment can mitigate but not fully resolve these limitations (Terryn and de Lhoneux 2024; Zhang et al. 2023). Crucially, recent *parameter-efficient fine-tuning* (PEFT) methods, such as low-rank adaptation (LoRA), update less than 1 percent of model weights, making domain adaptation feasible on everyday hardware.

This article outlines a framework; a PEFT workflow combining targeted data augmentation, lightweight fine-tuning and blended automatic–expert evaluation. Such a system can close much of the gap between raw machine output and expert translation for early modern Dutch, thereby rendering the VOC corpus computationally accessible and historically verifiable. To develop such a system, a modular pipeline (Figure 1) is designed that

1. augments early modern Dutch coverage via testimonial text data (1680–1792) and custom tokenizer training,
2. fine-tunes open LLMs with specialist parallel data using SFT and order-reward policy optimization (ORPO), and
3. evaluates the resulting translations with two complementary metrics: *BERTScore* and *METEOR* supplemented by expert judgment.

The outcome is a repeatable workflow for historical multilingual NLP: a model that renders 17th- and 18th-century Dutch courtroom testimony into modern English with fidelity comparable to human translators while retaining stylistic hallmarks, such as sentence length, register and punctuation.

Background

Current work on early modern Dutch translation intersects three domains: the historical breadth of the Dutch East India Company (VOC) archives, recent progress in domain-specific fine-tuning of LLMs and improvements in subword tokenization that enable LLMs to handle rare orthography and morphology (Sennrich, Haddow, and Birch 2016). The following sections situate the present study within this scholarship.

The Dutch East India Company and the VOC archives

Established in 1602, the VOC dominated early modern maritime commerce and colonial administration, particularly in Asia (Stoler 2002). Its surviving records, millions of pages of correspondence, ledgers and court proceedings, constitute a uniquely detailed source of economic, social and political history (Petram and Van Rossum 2022). Although large-scale digitization has facilitated access (Manjavacas and Fonteyn 2022) to these records, computational use of the corpus remains limited because the material is written in early modern Dutch. The variety diverges markedly from the present Dutch in vocabulary, syntax and idiom (Jiang and Zhang 2024), and there is no large bilingual corpus to train conventional neural-machine translation (NMT) systems (Huang et al. 2024). Specialized fine-tuning of multilingual LLMs therefore offers a promising, but still underexplored, pathway to research accessibility.

Testimonies of Enslavement

"*Testimonies of Enslavement*" comprises English translations of court cases heard by the VOC's Court of Justice in Cochin (Kerala) during 1680–1682 and 1707–1792 (van Rossum et al. 2020). This unique collection presents parallel texts in Dutch, Portuguese, Latin and local languages, systematically expanding historical abbreviations while preserving culturally specific legal and administrative terms that resist literal translation, precisely the linguistic challenges that complicate automated translation of Early Modern Dutch. The corpus's digitization includes detailed metadata covering case type, date, thematic tags and personal attributes, such as gender or legal status, making it a benchmark for studying slavery, migration and cross-cultural jurisprudence. Most critically for machine translation (MT) research, the collection's sentence-aligned parallel structure provides the expert-curated bilingual data essential for fine-tuning and evaluating historical translation models, bridging the gap between raw archival material and the structured datasets required for NMT.

Fine-tuning LLMs for specialized translation

MT remains a flagship application of transformer-based LLMs (Vaswani et al. 2017). NMT systems outperform earlier rule-based and statistical approaches because they learn cross-lingual patterns from large parallel corpora (Huang et al. 2024). A full fine-tune updates every weight in a multi-billion-parameter network, which demands expensive hardware. PEFT instead adjusts only a tiny slice of the model (often <1%), so researchers can adapt an LLM on a laptop-grade GPU while leaving the core weights untouched. Fine-tuning adapts these general models to low-resource historical domains through two successive steps:

1. *Instruction tuning* – also called SFT, exposes the model to curated sentence pairs so that it acquires the orthography, vocabulary and syntax of early modern Dutch.
2. *Preference alignment* reshapes the model's output style using human feedback. Popular methods include reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022), direct preference optimization (DPO) (Rafailov et al. 2023), lightweight LoRA adapters (LoRA adds a small trainable matrix to each layer, enabling parameter-efficient updates.) (Hu et al. 2022) and ORPO (Hong, Lee, and Thorne 2024).

Combined, these stages yield a translator that preserves historical nuance while meeting present-day quality expectations.

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

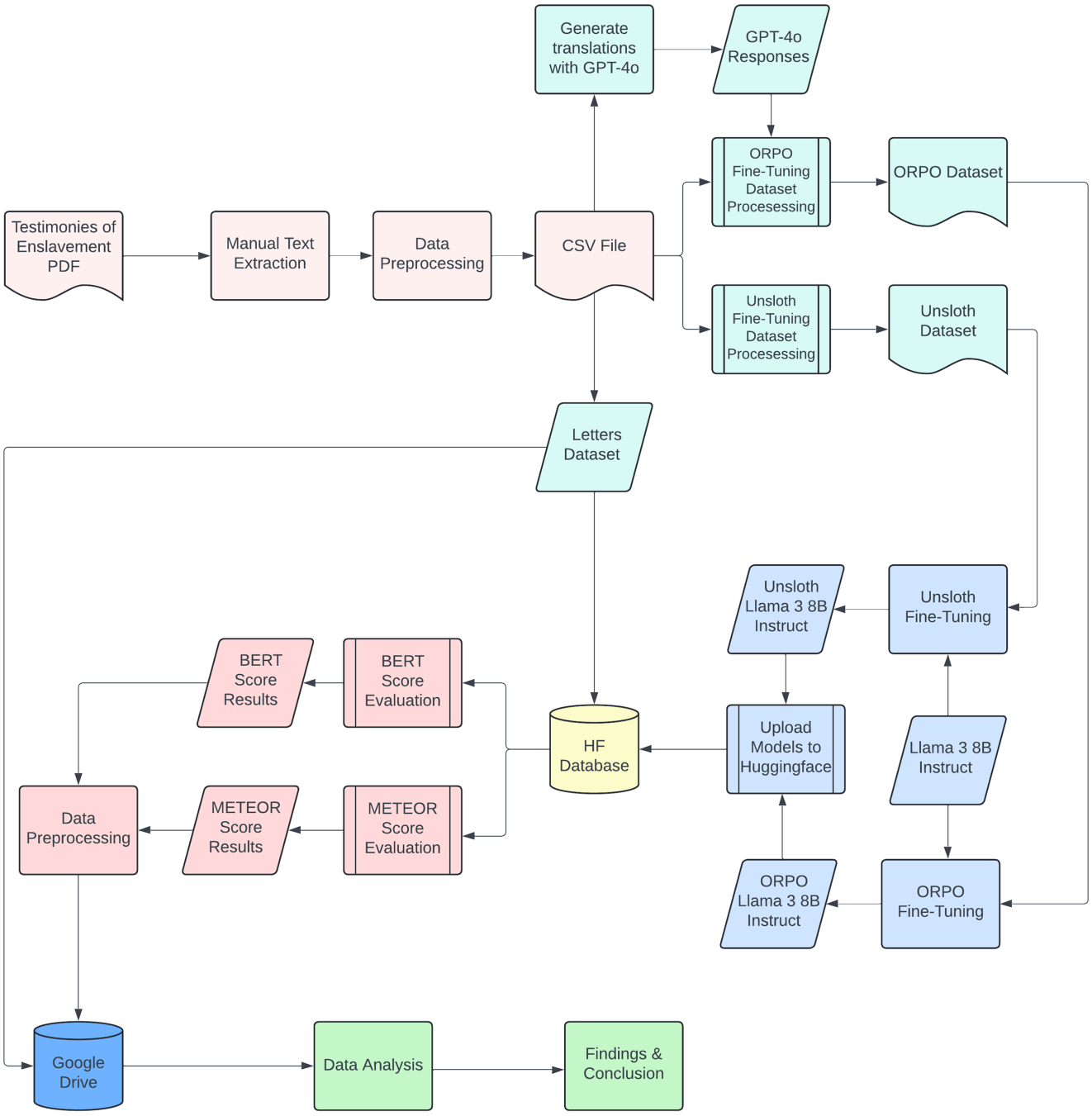


Figure 1. End-to-end training and evaluation pipeline.

Unsloth supervised fine-tuning

The first pipeline employs the Unsloth library, a wrapper around Hugging Face's `transformers` library that integrates PEFT modules (Unsloth AI 2024; Wolf et al. 2020). Unsloth accelerates training by fusing key matrix operations, re-ordering attention kernels and streaming gradients directly between CPU and GPU. On a single consumer GPU, it cuts training time by up to two times and memory usage by roughly 70 percent with no measurable loss in accuracy. It supports recent open models lines, such as Llama, Mistral, Phi and Gemma.

ORPO preference alignment

ORPO augments the language-model loss with an odds-ratio term that mildly penalizes rejected responses and strongly rewards preferred ones (Hong, Lee, and Thorne 2024). This regularization balances task learning and human alignment, improving generalization to previously unseen early modern Dutch inputs.

Machine translation

MT enables automatic transfer of text across languages and is indispensable for unlocking historical sources (Sennrich, Haddow, and

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

Birch 2016). Recent advances in NMT have markedly improved adequacy and fluency, even for linguistically complex and culturally embedded material such as early modern Dutch (Gao et al. 2022).

Three methodological stages frame the field:

1. *Rule-based MT*: Hand-crafted transfer rules and bilingual lexicons performed reliably in restricted domains, but proved brittle with orthographic variation, archaic morphology and long-distance dependencies (Wang et al. 2022).
2. *Statistical MT*: Word and phrase-based probabilistic models (Brown, Della, and Pietra 1993; Lopez 2008) replaced rules with corpus statistics, improving robustness but still struggling with reordering and rich morphology: complex inflectional patterns where words change form for grammatical functions, creating numerous variants that fragment translation probabilities.
3. *NMT*: Sequence-to-sequence architectures with attention (Bahdanau, Cho, and Bengio 2015), culminating in the Transformer (Vaswani et al. 2017), delivered large leaps in quality. Sub-word segmentation (Sennrich, Haddow, and Birch 2016) and massively multilingual pre-training (Aharoni, Johnson, and Firat 2019; Zhang and Sennrich 2020) further widened language coverage.

Unlike traditional MT systems that are explicitly trained on parallel corpora for translation tasks, contemporary LLMs acquire translation capabilities emergently through their general language modeling objectives. While specialized NMT systems remain optimized for translation quality, LLMs offer a more flexible approach that can handle multiple tasks, including translation, within a single model architecture. This study bridges these paradigms by adapting general-purpose LLMs to specialized historical translation through PEFT, combining the flexibility of LLMs with the domain specificity of traditional MT approaches.

- *Full fine-tuning* updates every parameter, achieving state-of-the-art accuracy but incurring high memory cost and risking catastrophic forgetting.
- *PEFT* freezes the backbone and trains lightweight modules – adapters (Houlsby et al. 2019), prefix-tuning (Li and Liang 2021) or LoRA matrices (Hu et al. 2022) that reduce compute without sacrificing quality.

LLMs offer two further advantages for historical MT: (i) zero- or few-shot performance on extremely low-resource pairs and (ii) unified multitask capability translation, summarization and extraction in one model. This study therefore tests two PEFT strategies, Unsloth SFT and ORPO, to inject early modern Dutch expertise into open-source LLMs while keeping training and inference budgets modest. The resulting models are benchmarked with embedding-based metrics that reward semantic fidelity and stylistic precision.

Translation evaluation

Given the complexity and historical significance of translating early modern Dutch texts, accurate evaluation is crucial. This study scores system outputs with two automatic metrics chosen for their ability to capture *semantic adequacy* and *fluency*:

(i) BERTScore (Zhang et al. 2020), a semantic-similarity metric that leverages contextual embeddings to gauge how closely a hypothesis preserves the reference meaning, an essential property for historically nuanced material.

(ii) METEOR (Banerjee and Lavie 2005; Denkowski and Lavie 2014), which aligns tokens at the exact, stem, synonym and paraphrase levels to balance lexical richness, fluency and stylistic authenticity.

Together these metrics overcome tokenization mismatches and length bias that affect traditional *n*-gram measures such as BLEU (Papineni et al. 2002; Post 2018), aligning with the goal of robust, historically-aware MT evaluation.

BERTScore

BERTScore compares a candidate sentence $C = \{c_1, \ldots, c_{|C|}\}$ with a reference sentence $R = \{r_1, \ldots, r_{|R|}\}$ in a contextual-embedding space. Each token is encoded by the multilingual `bert-base-multilingual-cased` model (Devlin et al. 2019) (layer 8 by default). Cosine similarities form the matrix $S_{ij} = \cos\big(\mathbf{e}(c_i), \mathbf{e}(r_j)\big)$. Keeping the maximum similarity for each token yields

$$\text{Precision} = \frac{1}{|C|}\sum_i \max_j S_{ij}, \qquad \text{Recall} = \frac{1}{|R|}\sum_j \max_i S_{ij}.$$

The harmonic mean gives the final score,

$$\text{BERTScore} = 2\,\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

after inverse-document-frequency weighting that down-weights very common words. Sentence-level scores are averaged to produce the system-level values reported in the "Results" section (range 0–1, where 1 denotes identical meaning).

METEOR

METEOR aligns hypothesis and reference tokens via exact, stem, synonym and paraphrase matches (Banerjee and Lavie 2005). Let m be the number of matched tokens; precision (P) and recall (R) are

$$P = \frac{m}{|C|}, \qquad R = \frac{m}{|R|}, \quad F_{mean} = \frac{10PR}{R + 9P}.$$

Fragmentation incurs a penalty,

$$\text{Penalty} = 0.5\Big(\tfrac{\text{chunks}}{m}\Big)^3, \qquad \text{METEOR} = F_{mean}\,(1 - \text{Penalty}),$$

yielding a score between 0 and 1. We use METEOR v1.5 (SacreBLEU implementation) with the English WordNet synonym list and the paraphrase table of Denkowski and Lavie (2014).

Tokenization

Tokenization is the initial parsing step in most NLP pipelines: It segments raw text into words, subwords or bytes that LLMs can process (Mielke et al. 2021). For multilingual or historical LLMs, the chosen scheme directly influences how well the network captures vocabulary, morphology and syntax (Kudo and Richardson 2018; Manjavacas and Fonteyn 2022; Sennrich, Haddow, and Birch 2016); translation quality therefore hinges on tokenization fidelity (Hu et al. 2024).

Efficient tokenizers compress more information into each context window, reducing training time and inference cost (Jiang and Zhang 2024). Early modern Dutch, however, exhibits variable spelling, archaic graphemes and mixed scripts that tokenizers optimized for present-day corpora often fragment. The resulting high out-of-vocabulary (OOV) rates lengthen sequences, split morpho-syntactic units and depress model performance (Manjavacas and

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

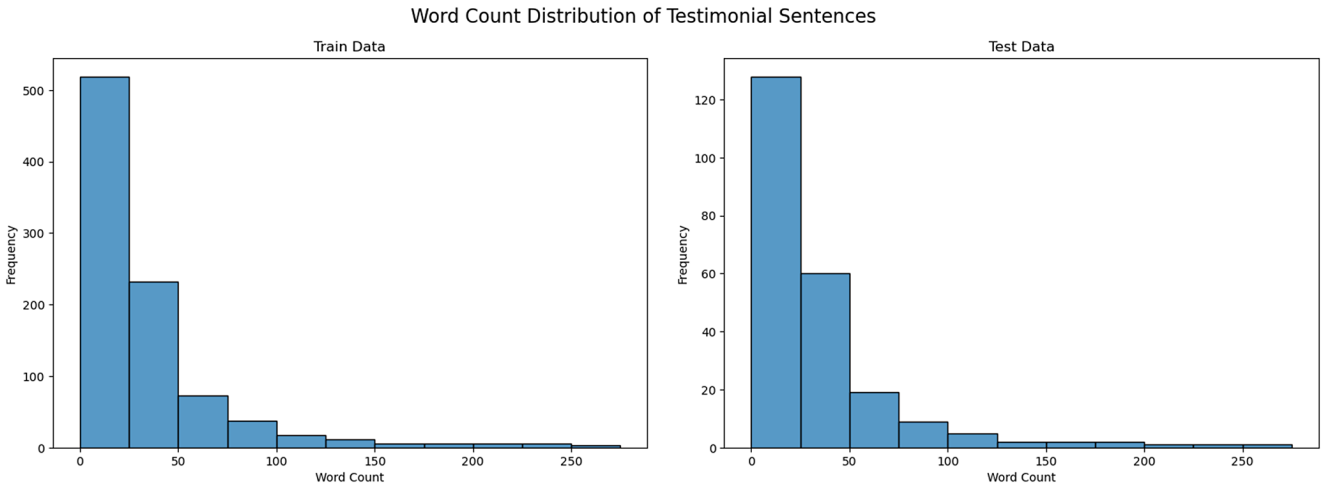


Figure 2. Word-count distribution for early modern Dutch sentences in the training and test splits.

Fonteyn 2022; Mielke et al. 2021). Targeted vocabulary augmentation and bespoke subword training are, therefore, prerequisites for accurate translation and analysis of historically significant documents. In this study, a custom tokenizer was developed to show what optimal processing of our source corpus would look like, the performance of which is reported as part of the methodological setup.

Data methods

Specialized data procedures were required at all stages of the pipeline (Figure 1), including extraction, cleaning, corpus construction and evaluation.

Early modern Dutch source sentences and their English renderings were transcribed from VOC Court-of-Justice testimonies recorded in Cochin (Kerala) for 1680–1682 and 1707–1792 (van Rossum et al. 2020). Because published PDFs contain complex layouts, the sentences were manually copied into a two-column comma-separated file comprising 1,281 aligned pairs.

The exploratory graphs showed that both Dutch and English sentences were typically shorter than 300 words (Figure 2); the frequency of sentences greater than 300 words declined significantly beyond that limit. After stripping extra whitespace, line breaks and other artifacts with Python regular expression scripts, we discarded pairs containing fewer than five or more than 300 tokens and split the remainder into training and test sets (80% and 20%, respectively).

Alignment corpora: The training split was duplicated to create two datasets: one for Unsloth SFT and one for ORPO fine-tuning (Hong, Lee, and Thorne 2024). ORPO requires a `chosen/rejected` format, so expert translations filled the `chosen` slots while GPT-4o outputs with identical prompts populated the `rejected` column (OpenAI 2024).

Consolidation and cleaning: Model outputs were stored separately and then merged so that each row contained (i) the Dutch source with word count and case ID, (ii) the expert reference, and (iii) each system's hypothesis. Residual anomalies (e.g., stray brackets) were removed before saving the consolidated file.

BERTScore and METEOR ("Translation evaluation" section) were computed for every hypothesis–reference pair. System-level statistics were exported as `.csv` files for subsequent analysis.

Tokenizer implementation

To mitigate the issues of orthographic fragmentation outlined in the "Tokenization" section, a custom byte-pair encoding (BPE) tokenizer was trained on the Early Modern Dutch corpus. This step served as a foundational preprocessing measure to ensure the subsequent fine-tuning models could efficiently process the historical texts. Its performance was compared against several standard tokenizers (GPT-2, mBERT, and XLM-R) on the test set. As shown in Table 1, the custom tokenizer achieved an impressive OOV coverage in contrast to other tokenizers and the highest sequence efficiency, representing the source texts in the fewest number of tokens. This optimization reduced computational overhead and potential context fragmentation during the fine-tuning and inference stages.

Model evaluation methods

This study combines two automatic metrics, *BERTScore* and *METEOR*, with a qualitative review by a specialist in early modern Dutch. (For metric definitions, see the "Translation evaluation" section.) To validate the effectiveness of our fine-tuning approach, we employed a comparative framework that contrasted adapted models against their base (unfine-tuned) versions, supplemented by domain expert evaluation. This dual validation strategy ensured that performance gains resulted from targeted adaptation rather than general model capabilities.

Automatic scores quantified semantic adequacy and fluency across all systems, while expert judgment specifically assessed whether fine-tuned models better preserved historical nuance compared to their base counterparts (Freitag et al. 2022; Zhang et al. 2020). BERTScore and METEOR were computed against reference translations prepared by scholars at Arsip Nasional Republik Indonesia and the Corts Foundation (van Rossum et al. 2020). The domain specialist reviewed parallel outputs from base and fine-tuned models, focusing on preservation of legal terminology, historical context and documentary authenticity.

This comprehensive evaluation framework, combining quantitative comparisons between base and adapted models with qualitative domain expertise, ensured that performance improvements reflected genuine enhancements in historical translation quality rather than metric optimization alone.

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

Table 1. Tokenizer performance comparison on early modern Dutch

Tokenizer	Avg. tokens/sentence	OOV rate	Compression ratio	Sequence efficiency
Custom_BPE	**64.04**	**0.00%**	**0.665**	**0.763**
GPT-2	109.25	0.00%	0.401	0.619
mBERT	91.80	0.45%	0.445	0.640
XLM-R	88.44	0.00%	0.467	0.660

Testing environment

All inference ran in Google Colab on NVIDIA L4 and A100 GPUs via a custom notebook. (Notebook archive: `https://anonymous.4open.science/r/EMDutchLLMFinetuneA67B`.) Each model received a one-shot prompt consisting of (i) an instruction to render an early modern Dutch sentence in contemporary English under specified guidelines and (ii) the sentence itself (Lip 2025). Generation parameters were fixed across systems: `do_sample=True`, `num_beams=5`, `penalty_alpha=0.6`. The models translated 230 sentences from 18th-century court cases; outputs were written to CSV for subsequent scoring.

Quantitative benchmark

Python implementations of BERTScore and METEOR (Lip 2025) produced scores for every hypothesis–reference pair. BERTScore captures sentence-level semantic similarity, while METEOR assesses lexical choice, phrase alignment and fluency. Their conjunction provides a more complete picture of translation quality than either metric alone.

Qualitative benchmark

For the qualitative benchmark, a domain expert performed a systematic review of all 230 test sentences across all model outputs. Rather than aiming for "perfect" automated translations – an unrealistic goal given the complexity of historical texts – our framework positions model outputs as *preliminary drafts* for expert refinement. In practice, historians would work with these imperfect translations, using them as starting points that significantly accelerate the traditional translation process while maintaining scholarly oversight.

Training

Two fine-tuning strategies were applied to the testimonial corpus (see the "Data methods" section): *SFT* implemented with the *Unsloth* parameter-efficient library (Unsloth AI 2024); and *ORPO* (Hong, Lee, and Thorne 2024).

Base models

Experiments start from publicly available, instruction-tuned checkpoints:

- `Llama-3-8B` (Meta) – SFT and ORPO;
- `Llama-2-7B` (Meta) – SFT;
- `Mistral-7B-Instruct-v0.3` (MistralAI) – SFT;
- `TinyLlama-1.1B` – SFT.

Compute environment

Fine-tuning was performed in Google Colab Pro on NVIDIA L4 (24 GB) and A100 (40 GB) GPUs. Notebooks and configuration files are archived in the project repository (Lip 2025). Training curves were logged to Weights and Biases for real-time monitoring (Biewald 2020).

Hyper-parameters

All runs use an accumulative batch size of 64 and a maximum sequence length of 512 tokens. For the full parameter list for each model, see Table A6.

- *SFT (Unsloth):* Nine epochs, AdamW optimizer, learning rate 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$.
- *ORPO:* Eight epochs, AdamW optimizer, learning rate 1×10^{-5}, same β values, KL-divergence weight 0.02.

Results

High-performing models

The three best systems were (i) Mistral-7B-Instruct fine-tuned with Unsloth, (ii) Llama-3-8B-Instruct with Unsloth, and (iii) GPT-4o. Mistral-Unsloth obtained the highest mean and median on both BERTScore and METEOR, with the narrowest inter-quartile range (Tables A10 and A11, marking it as the most reliable translator for early modern Dutch. Llama-3-Unsloth produced comparably high scores but showed slightly greater spread, while GPT-4o delivered a strong, untuned baseline.

GPT-4o required no additional training yet consistently approached expert quality: its low metric variance indicates robustness across sentence length and syntactic complexity (Table A10). Although it seldom tops individual metrics, its "plug-and-play" reliability makes it a practical benchmark when fine-tuning resources are scarce.

Mistral-Unsloth excelled not only in aggregate metrics but in domain expert assessment, demonstrating a strong capacity for handling the specialized lexicon of historical legal documents. The model consistently preserved critical semantic categories, including specific occupation titles (e.g., wasvrouw [laundress]), formal role titles (e.g., eijsscher [plaintiff/prosecutor]) and period-specific legal terms (e.g., schagerije [a type of dairy farm or settlement]). This proficiency is evidently illustrated in Table A2, where the translation accurately renders complex procedural language – such as "Aldus gedaan en gerelateert" to "Thus done and testified" and "gesw: clerq" to "sworn clerk" – while correctly preserving names, dates and the formal structure of the declaration. However, this high standard was not maintained across all outputs. Occasionally, the model exhibited inconsistencies with certain filler words and abbreviations, leading to a dip in translation quality. These specific shortcomings are documented in Tables A1 and 3, which showcase instances of untranslated words and verbatim copying of the source text, highlighting the areas where the model's performance remains variable.

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

Table 2. Example of a high-quality translation by Llama-3-Unsloth showing preservation of legal syntax and formal register

Early modern Dutch	English translation (Llama-3-Unsloth)
So is 't dat wij na rijpe deliberatie, en na overwogen te hebben al 't gene ter sake is dienende, ons met voorsz: sententie en gewijsde hebben geconformeert, gelijk wij deselve confirmeeren, en approberen bij desen, met last en ordre aan voorm: Raad van Justitie om voorsz: largatie op heden te doen effect sorteeren.	Thus it is that we after mature deliberation, and after having considered all that concerns this matter, conform ourselves with the aforementioned sentence and verdict, as we confirm and approve herewith, with instruction and order to the aforementioned Council of Justice to effectuate the aforementioned largatie immediately.

Table 3. Example of an unsatisfactory translation: the Dutch text is copied verbatim rather than translated

Early modern Dutch	English translation (Mistral-Unsloth)
Tevergeefs tragt men somtijds zig hunne natuurlijke genegentheit tegen te zetten.	Tevergeefs tragt men somtijds zig hunne natuurlijke genegentheit.

Llama-3-Unsloth ranked a close second. Metric gains over the unfine-tuned Llama-3-8B-Instruct confirm the benefit of domain adaptation. Expert review noted strong preservation of complex legal syntax and formal register, though the model maintained the same tendency to leave Dutch technical vocabulary untranslated (Table 2). The model demonstrated sophisticated handling of archaic abbreviations and intricate sentence structures, expanding abbreviated forms like "voorsz:" and "voorm:" to "aforementioned" while preserving the document's formal legal tone. However, specialized legal terms, such as "largatie," remained untranslated, reflecting the model's capability to preserve domain-specific terminology.

Both Unsloth-tuned models achieved high automatic scores yet share two limitations: (a) incomplete rendering of specialized terminology and (b) occasional omission of abbreviations or honorifics embedded in the source. Addressing these gaps, through the expansion of the targeted glossary or additional preference alignment, will be essential before deploying the systems for large-scale historical research.

In sum, automatic metrics and expert feedback converge: Unsloth fine-tuning markedly improves translation quality, with Mistral-7B-Unsloth the current best option, with domain-specific terminology remaining intact, preserving historical authenticity.

Mid-range performance models

Aya-23-8B produces readable output, but quality fluctuates more than for the top systems. Roughly one sentence in 10 drifts far enough from the reference to require manual correction, as shown by the wider spread in Table A7. Error clusters in Table A3 reveal systematic problems with archaic verb endings and idiomatic noun compounds, suggesting that a modest domain-specific fine-tune could stabilize performance. Teams employing Aya-23-8B should budget a short post-editing pass and, where resources permit, curate several thousand additional sentence pairs before large-scale deployment.

The *Llama 2-Unsloth* and *Assistant* models deliver steadier output, but still suffer from untranslated sentences. Their higher mean BERTScore and METEOR values (Figures 3 and 4) allow editors to focus on stylistic polish rather than basic correctness. Nevertheless, the long tail of low-confidence sentences, again visible in Table A10, shows that perfect automation is still out of reach. This can be observed in Table A3 where the phrase "myself, the sworn clerk" has been left out of the translation. Furthermore, on longer sentences, the model will still leave the sentence fragmented and untranslated (Table A4). These results reflect the model's inconsistent performance, which requires further improvements for usability.

Qualitative review confirms these findings. Both models usually preserve grammatical structure and semantic content, yet they recurrently leave Dutch terms, such as *meijd*, *jongen* and *eijsser* untranslated, and they handle proper names or titles inconsistently. Another quirk is the insertion of the bracketed word "signed" in official documents. Table A3 illustrates a strong rendering, whereas Table A4 shows a complete failure in which the Dutch source is reproduced verbatim. The comparison reveals a catastrophic LLM failure where the model failed to translate, instead producing a "modernized" Dutch version that destroys historical authenticity through inappropriate spelling standardization and content truncation. This example illustrated a failure where an untrained LLM produced modernized Dutch text rather than English translation, destroying historical authenticity through inappropriate orthographic standardization and content truncation. The failure validates our core finding that PEFT is essential for preserving documentary integrity in early modern Dutch translation tasks.

Llama 2-Unsloth's metric gains over the base Llama-2-7B model indicate successful adaptation, yet the broader score distribution in Figures 3 and 4 implies vulnerability to syntactically dense passages.

Mistral-7B-Instruct-v0.3, despite the absence of domain fine-tuning, delivered impressively stable scores with minimal fluctuation (Figure 3), making it an attractive foundation for future adaptation. Expert review (Tables A8 and A9) shows that it preserves structure in simpler sentences but mistranslates abbreviations (*E:*, *achtb.*) and substitutes modern equivalents for historical measurements or currencies.

In sum, the mid-range systems achieve moderate automatic scores but share three limitations: (i) incomplete handling of specialized terminology, (ii) inconsistent treatment of proper nouns and titles, and (iii) occasional sentence-level failures on complex inputs. Addressing these gaps through targeted glossary expansion, additional preference alignment or confidence-based filtering will be essential before large-scale historical translation can proceed without human oversight.

Under-performing models

TinyLlama-1.1B-Unsloth scored well below all other fine-tuned systems (Table A7). Its low mean BERTScore and METEOR values, and the wide interquartile range, show that a one billion-parameter model lacks the capacity to capture early modern Dutch morphology and legal terminology. Occasional fluent sentences do appear, but the high variance suggests that extensive

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

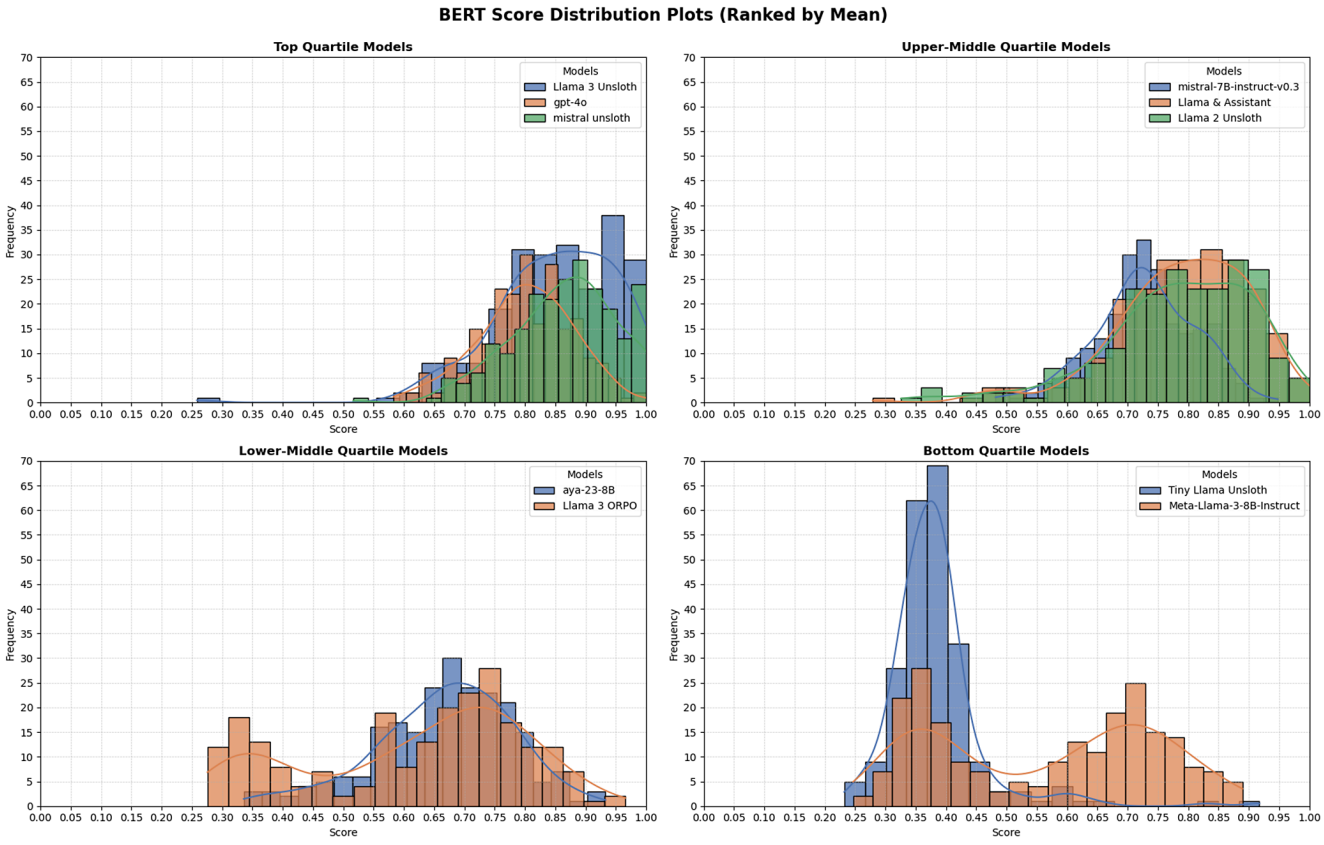


Figure 3. Distribution of BERTScore values for all systems, ranked by mean.

domain-specific augmentation or a larger backbone would be required before production use.

Llama-3-8B-ORPO exhibited similar volatility. Although individual outputs sometimes matched expert quality, many contained prompt remnants, repetitions or unfinished clauses, depressing overall metrics (Table A7). These errors point to miscalibrated reward weights: refining the preference data set or annealing the KL term could stabilize training.

The *base Meta-Llama-3-8B-Instruct* checkpoint delivered the weakest results among the untuned systems (Table A10). High score dispersion confirms that instruction tuning alone is insufficient for 17th-century Dutch and corroborates the large gains achieved through Unsloth SFT.

Evaluation

Fine-tuned model capabilities

The fine-tuned systems adapted with the Unsloth procedure markedly outperformed their baseline counterparts. Mistral-Unsloth achieved the closest alignment with expert reference translations, confirming that Unsloth can capture the lexical nuance and syntactic irregularity characteristic of early modern Dutch. Its low score variance further recommends it for applications in which scholars require dependable, minimally post-edited output.

Llama-3-Unsloth also profited from the same adaptation, although its wider dispersion of scores indicates residual sensitivity to sentence complexity and specialized terminology. Even so, the model's mean performance approaches that of Mistral-Unsloth, shown in Figure 5, underscoring the practical value of targeted fine-tuning for historical translation. Together, the two models demonstrate that parameter-efficient strategies can yield both accuracy and stability, provided the backbone has sufficient capacity and the domain data are carefully curated.

ORPO fine-tuning challenges

The Llama-3-8B model fine-tuned with ORPO delivered markedly lower and less reliable scores than the Unsloth SFT baselines. Its mean BERTScore was 0.617 (median = 0.665, SD = 0.180) and its mean METEOR 0.405 (median = 0.461, SD = 0.246; see Tables A10 and A11). Although the model occasionally produced near-expert translations, the wide dispersion of scores indicates that such high-quality outputs were unpredictable and infrequent. A review of the training notebook and data flow points to three primary implementation issues that likely caused this underperformance, rather than an intrinsic flaw in the ORPO algorithm for this task.

a) *Sparse, imbalanced preference pairs:* ORPO's efficacy depends on a sufficient volume of high-quality chosen-rejected example pairs. With only 1,281 sentence-translation tuples available, creating the preference dataset effectively halved the number of examples seen per epoch compared to SFT. This data scarcity forces the model to learn from a very narrow distribution, increasing the risk of overfitting to frequent lexical patterns in the training data rather than learning generalizable syntactic and semantic rules for translation. Furthermore, using a powerful model like GPT-4o for the "rejected" samples, while a common practice, may have introduced high-quality but stylistically

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

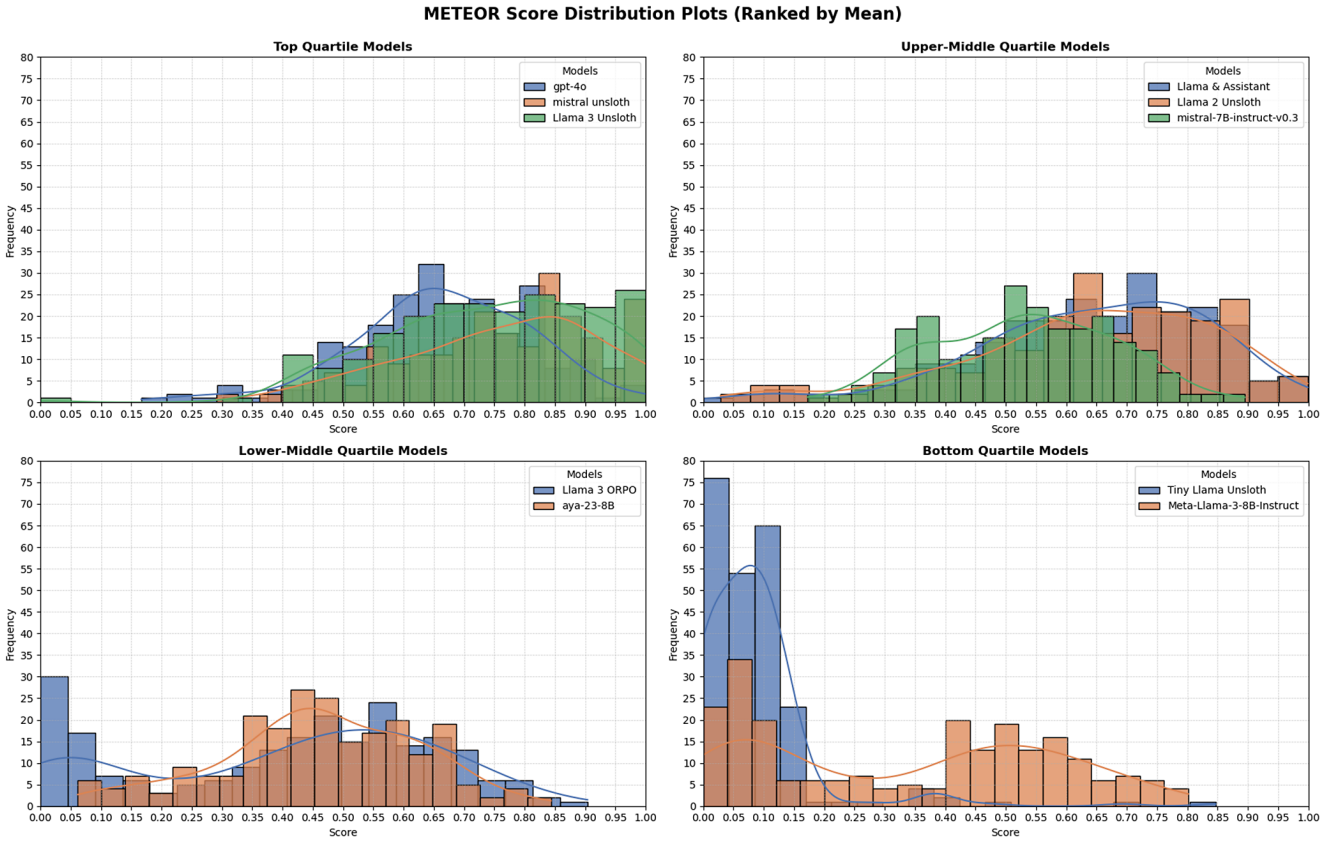


Figure 4. Distribution of METEOR scores for all models, ranked by mean.

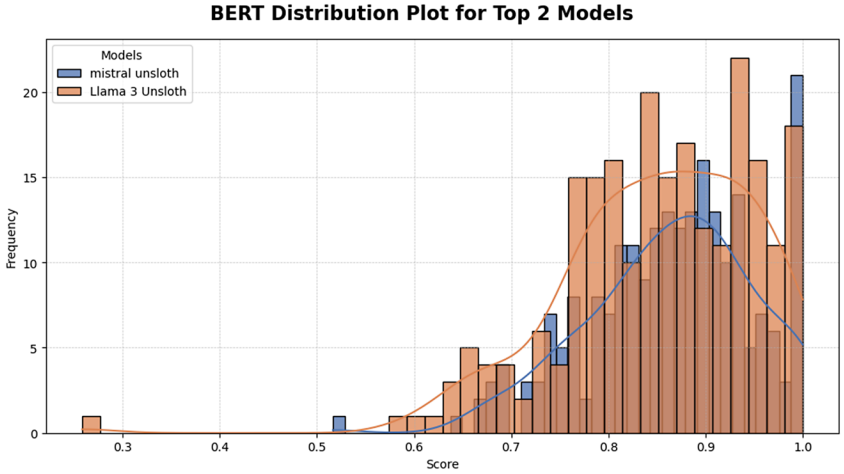


Figure 5. Sentence-level METEOR distributions for the two best systems.

different translations, potentially sending a noisy or conflicting signal to the model during alignment.

b) *Prompt-template drift:* A critical mismatch existed between the structured, multi-line instruction block used during training and the simple, single-sentence prompt used at inference. Instruction-tuned models are highly sensitive to the format of the prompt. This discrepancy meant that the model was not optimized for the task it was being asked to perform during evaluation. Consequently, the model often exhibited prompt leakage, reproducing parts of the training header or failing to complete the translation (shown in Table A5), which severely depressed the adequacy scores and contributed to the high variance in performance.

c) *Conservative hyper-parameters:* The training configuration, which ran for eight epochs with a low learning rate of 1×10^{-5} and a micro-batch size of two, was likely too conservative for this context. In ORPO, the language modeling loss is balanced by a KL-divergence regularizer that prevents the model from deviating too far from its original capabilities. When combined with a low learning rate, this regularizer can dominate the training process, effectively stifling the model's ability to adapt to the new domain. The model is thus prevented from fully

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

learning the nuances of Early Modern Dutch, leading to translations that may remain closer to the base model's generic output rather than acquiring specialized historical knowledge.

These three factors – data sparsity, prompt inconsistency and a restrictive training regime – collectively explain the 0.20–0.35 drop in mean scores and the nearly three-fold increase in variance relative to the Unsloth SFT models, without necessarily indicting ORPO's underlying algorithm for historical translation tasks.

Base model capabilities

The untuned *Mistral-7B-Instruct-v0.3* delivered stable BERTScore and METEOR values across the test set, confirming that it possesses a strong multilingual prior even without domain adaptation. This consistency makes it an attractive foundation for historical translation workflows. Once fine-tuned with Unsloth SFT the model's average scores rose by 8–10 points, and variance fell by roughly one third, indicating both higher accuracy and greater reliability (Table A10). The gain demonstrates the practical value of lightweight parameter-efficient tuning when adapting a general LLM to the lexical and syntactic idiosyncrasies of 17th- and 18th-century Dutch.

Among the other baselines, GPT-4o produced the most dependable high-quality output without any additional training. Although its mean scores lagged slightly behind the Unsloth-enhanced Mistral, the narrow dispersion of GPT-4o shows that it can serve as a robust standard alternative when compute or data for fine-tuning are unavailable (Figures 3 and 4). Taken together, these results suggest a two-tier deployment strategy: use GPT-4o as an immediate, low-overhead baseline and adopt a fine-tuned Mistral backend when project resources permit domain adaptation and post-editing.

Discussion

The comparative results underline how strongly translation quality varied between fine-tuning strategies. This is shown starkly in Figure A1, which highlights the jump in performance from base model to fine-tuned model through a heatmap visualization. Unsloth adaptation, exemplified by the Mistral-Unsloth and Llama-3-Unsloth models (Figures 3 and 4), consistently surpassed all alternatives. The ORPO variant remains less successful, but its shortcomings appear remediable ("ORPO fine-tuning challenges" section). Adjusting the preference corpus and harmonizing training and inference prompts may narrow the current performance gap. More broadly, combining automatic metrics with an authenticity checklist covering proper-noun preservation, date fidelity and punctuation style would yield a balanced scorecard for future experiments. A recurring error pattern across systems is the handling of very short lexical items and abbreviations in Dutch: when a sentence is both brief and semantically opaque, the model often leaves such tokens untranslated because their functional role is unclear. Targeted glossaries or context-window augmentation could mitigate this issue without sacrificing fluency.

Historical translators must decide which elements should remain unchanged to preserve documentary authenticity. Proper names, official titles (*schout*, *fiscaal*) and obsolete currency units (e.g., *stuiver*) are prime examples. Our expert evaluation shows that Mistral-Unsloth retains almost all personal and place names verbatim, whereas Llama-3-Unsloth occasionally anglicizes Dutch surnames or expands abbreviated titles. In contrast, GPT-4o sometimes over-translates institutional names, rendering *Hoogheemraadschap* as "Water Board," which may help modern readers but risks erasing historical nuance. A pragmatic compromise is to keep these critical tokens in their original form and provide a parenthetical gloss on the first occurrence. Implementing this policy would require a lightweight named-entity filter upstream of the decoder or a constrained decoding pass that protects whitelisted terms.

Findings

This study compared two parameter-efficient adaptation strategies, ORPO and Unsloth SFT for translating early modern Dutch into present-day English assessed through both qualitative and quantize frameworks. Both methods improved substantially on untuned foundation models, yet their performance profiles diverge in ways that matter for historical scholarship. Unsloth fine-tuned Mistral and Llama 3 systems delivered the highest BERTScore and METEOR values while preserving most syntactic and stylistic features observed in expert references. ORPO, by contrast, produced more variable output: when reward signals aligned with domain needs, it matched Unsloth quality, but miscalibrated preferences yielded prompt leakage or unfinished clauses. The TinyLlama experiment shows that extreme parameter reduction comes at a steep cost to adequacy and fluency, confirming that complex, low-resource language varieties still demand mid-size or larger backbones. Taken together, the results advocate for a two-step workflow, tokenizer customization, followed by Unsloth SFT for projects that require dependable, scalable translation of historical texts.

Future work

Scalable infrastructure

The present experiments were limited by GPU memory, restricting model depth, batch size and context window. Even a modest increase in compute would enable larger backbones and longer sequences; scaling-law studies indicate near-linear gains in accuracy under such conditions (Hoffmann et al. 2022; Kaplan et al. 2020). Beyond simply improving accuracy, scaling up the infrastructure would allow for fundamentally new research avenues. For instance, larger context windows (e.g., 32k or 128k tokens) would enable the translation of entire documents or chapters at once, rather than sentence by sentence. This would significantly improve coherence and consistency, allowing the model to track entities, terminology and narrative threads across long passages. Furthermore, employing larger backbone models (e.g., 70B or mixture-of-experts models) could enhance the system's ability to grasp complex legal reasoning, subtle irony or intricate rhetorical structures present in the source texts, which smaller models might oversimplify.

Richer, well-curated data

Translation quality is ultimately capped by the scarcity of parallel material. Expanding the corpus across genres and annotating each record with place, year and document type remain a priority. Automatic workflows that recognize 17th- and 18th-century typographic conventions can accelerate transcription, while a tokenizer retrained on the enlarged dataset should capture historical spelling variation without inflating sequence length (Kudo and Richardson 2018). While the current corpus of court testimonies is valuable, its legalistic and formal nature may bias the model. Future work

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

should prioritize expanding the training data to include a wider array of genres, such as personal letters, ship logs, administrative ledgers and pamphlets. This would create a more versatile translator capable of handling different registers and domains. Additionally, a significant challenge is the lack of a standardized transcription protocol for Early Modern Dutch. A key future task will be to develop and apply a semi-automated transcription workflow that can handle abbreviations, non-standard spelling and paleographic variations, perhaps using a combination of OCR/HTR models fine-tuned on specific scribal hands. This would not only accelerate corpus creation but also enrich the dataset with valuable metadata about scribal practices. This expanded and diversified corpus would, in turn, serve as the foundation for retraining the domain-specific tokenizer, further enhancing its ability to model historical spelling variation across genres and time periods.

Enhanced alignment strategies

ORPO lagged behind SFT chiefly because of sparse preference pairs and divergent prompt templates. Forthcoming work will (i) enlarge the preference pool with varied negative examples, (ii) harmonize training and inference prompts, and (iii) run ORPO as a post-SFT refinement step (Hong, Lee, and Thorne 2024; Ouyang et al. 2022). The suboptimal performance of ORPO highlights the sensitivity of preference-based alignment methods to data quality and prompt design. Future iterations will focus not just on enlarging the preference pool, but also on diversifying it. This includes generating "better" negative examples that are subtly incorrect (e.g., grammatically fluent but semantically inaccurate) rather than obviously flawed. This would force the model to learn finer-grained distinctions. Beyond ORPO, exploring other alignment techniques like Kahneman–Tversky optimization (KTO), which does not require paired preference data, could offer a more robust solution for low-resource scenarios. A comparative study of these different alignment methods could yield a clearer understanding of which strategies are best suited for the unique challenges of historical texts, where human preferences might involve balancing literal accuracy with historical authenticity.

Adaptation to other historical languages

The presented workflow, comprising metadata-rich corpus curation, custom tokenizer training and two-stage alignment with continuous domain expert oversight, offers a transferable methodological blueprint for other low-resource historical languages (e.g., Middle High German, early modern Spanish and regional Malay manuscripts). Although the current investigation focuses specifically on Early Modern Dutch legal texts, the core components of this framework, particularly the integration of domain expertise throughout the pipeline, provide a reproducible approach adaptable to diverse historical contexts and text types.

It is acknowledged that the present implementation utilizes a specialized legal corpus; consequently, future applications would benefit from validating the approach across multiple genres and domains. The essential transferable elements include the PEFT strategies, evaluation protocols combining automatic metrics with expert review and tokenizer adaptation methodologies. However, successful application to new languages necessitates language-specific resources, particularly curated parallel corpora developed in collaboration with subject matter experts.

By emphasizing verifiable translation through sustained domain expert involvement, this pipeline model could help unlock extensive archives for linguists, historians and the broader public. The framework's ultimate adaptability will depend on addressing language-specific challenges while maintaining the core principle that historical translation requires continuous scholarly validation to ensure both accuracy and preservation of documentary authenticity.

Interactive and verifiable translation

Future systems should move beyond static translation and toward an interactive, human-in-the-loop framework. This could involve developing an interface where historians can not only correct translations but also see the model's confidence scores for specific words or phrases. The system could offer multiple translation options for ambiguous terms, annotated with explanations drawn from historical dictionaries or linguistic corpora. Such a tool would transform the model from a "black box" into a transparent research assistant. Furthermore, integrating a mechanism for the model to learn from these expert corrections in real-time (online learning) would create a continuously improving system that becomes more attuned to the specific needs and nuances of a particular research project or archive.

Interdisciplinary collaboration

This research highlights the critical importance of sustained interdisciplinary collaboration between computational methods and humanities scholarship. Future work should develop structured frameworks for domain experts and technical researchers to co-design translation systems that balance technical efficiency with historical accuracy. Such partnerships are essential for addressing epistemological questions about how MT shapes historical interpretation while ensuring tools enhance rather than replace traditional scholarly practices. Establishing sustainable collaboration models will be crucial for adapting this approach to diverse historical contexts and text types.

Conclusion

This study investigated PEFT strategies for translating early modern Dutch into present-day English.

Unsloth adaptation proved most effective: the Mistral-7B-Unsloth model achieved the highest mean BERTScore and METEOR values with the narrowest spread, while Llama-3-Unsloth followed closely but showed greater variance. By contrast, Llama-3-ORPO lagged behind, its lower and less stable scores traceable to sparse preference data and prompt drift rather than to ORPO's core algorithm.

Among the untuned baselines, GPT-4o offered the strongest off-the-shelf performance, furnishing a dependable benchmark when compute or parallel data are scarce. Mid-range systems, such as Aya-23-8B and Mistral-7B-Instruct, delivered moderate but fluctuating quality, and the TinyLlama-Unsloth experiment confirmed that extreme parameter reduction cannot yet capture the lexical and syntactic nuance of 17th- and 18th-century Dutch.

Three broad conclusions follow. First, careful, PEFT, backed by a domain-specific tokenizer and curated parallel corpus, can raise translation quality to near-expert levels while containing hardware costs. Second, robustness still hinges on sentence length and specialized terminology; further gains will require richer training data provided by domain-experts, context-aware prompts and glossaries for period-specific vocabulary. Third, hybrid evaluation that

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

combines embedding-based metrics with domain-expert review remains essential: automatic scores alone under-report errors in proper-name handling, abbreviations and legal terms.

Future work should (i) scale Unsloth tuning to larger backbones and longer contexts, (ii) expand the parallel corpus across genres and regions, and (iii) revisit ORPO as a post-SFT refinement once preference data and prompt alignment are improved. The framework outlined here, metadata-rich corpus curation, custom tokenizer training and domain-expert-aided two-stage alignment, offers a transferable blueprint for other low-resource historical languages, such as Middle High German, Early-Modern Spanish and regional Malay manuscripts. By systematically addressing the twin challenges of orthographic fragmentation and domain-specific terminology that characterize historical texts, such pipelines could unlock extensive archives for linguists, historians and the broader public. This research envisions a collaborative workflow where machine-generated translations serve as initial drafts for domain experts, reducing the time-intensive process of translating archival materials from scratch. Rather than replacing human expertise, our approach provides historians with rough translations that preserve historical terminology and context, which they can then refine and verify. This hybrid human–machine process makes large-scale archival translation feasible while ensuring scholarly accuracy through continuous expert validation.

This methodological framework promises to enable new research on migration, cultural exchange and social change that remains hidden behind linguistic barriers, while ensuring that computational tools preserve rather than compromise the documentary integrity essential to historical scholarship. The combination of PEFT with rigorous evaluation protocols thus provides a scalable pathway toward more accurate and transparent access to the world's multilingual archival heritage.

Acknowledgements. A special thanks to David Grantsaan for evaluating the translation quality of the LLMs, their expertise was crucial for judging the results.

Data availability statement. Computation notebook files, supplementary material, and replication of open-access data and code are available at this anonymous Github Repository: https://github.com/glp500/EMDutch-LLM-Finetune.

Disclosure of Use of AI Tools. No AI was used to generate content or conduct any part of the research.

Author contributions. Conceptualization: G.L., V.d.B. and A.B.; Data curation: G.L.; Formal analysis: G.L.; Investigation: G.L.; Methodology: G.L.; Resources: V.d.B. and A.B.; Software: G.L.; Supervision: V.d.B. and A.B.; Validation and formal analysis: D.G.; Visualization: G.L.; Writing – original draft: G.L.; Writing – review and editing: V.d.B. and A.B.

Funding statement. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The research was conducted as a partnership between Vrije Universiteit Amsterdam and the KNAW Humanities Cluster. Open access funding provided by Vrije Universiteit Amsterdam.

Competing interests. The authors declare none.

Ethical standards. The authors affirm that this research did not involve human participants.

References

Aharoni, Roee, Melvin Johnson, and Orhan Firat. 2019. "Massively Multilingual Neural Machine Translation." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), edited by **Jill Burstein, Christy Doran, and Thamar Solorio**, 3874–84. Minneapolis, MN: Association for Computational Linguistics.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate." In *3rd International Conference on Learning Representations (ICLR 2015)*, edited by **Yoshua Bengio and Yann LeCun**. San Diego, CA: ICLR.

Banerjee, Satanjeev, and Alon Lavie. 2005. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization*, edited by **Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss**, 65–72. Ann Arbor, MI: Association for Computational Linguistics.

Biewald, Lukas. 2020. "Weights and Biases." Software available from https://wandb.ai.

Brown, Peter F., Stephen A. Della, and Vincent J. Pietra. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics* 19: 263–311.

Deng, Ningyuan, Shuao Guo, and Yanqing He. 2023. "Istic Neural Machine Translation System for EvaHan 2023." In *Proceedings of ALT2023: Ancient Language Translation Workshop*, edited by **Chao-Lin Liu, Shervin Malmasi, and Chao-Jun Xuan**, 34–42. Macau SAR: Asia-Pacific Association for Machine Translation.

Denkowski, Michael, and Alon Lavie. 2014. "Meteor Universal: Language Specific Translation Evaluation for any Target Language." In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, edited by **Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna**, 376–80. Baltimore, MD: Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), edited by **Jill Burstein, Christy Doran, and Thamar Solorio**, 4171–86. Minneapolis, MN: Association for Computational Linguistics.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-Kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. "Results of WMT22 Metrics Shared Task: Stop Using BLEU—Neural Metrics Are Better and More Robust." In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, edited by **Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, and Roman Grundkiewicz**, 46–68. Abu Dhabi: Association for Computational Linguistics.

Gao, Luyu, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. "Precise Zero-Shot Dense Retrieval without Relevance Labels." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by **Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki**, 1762–77. Toronto: Association for Computational Linguistics.

GLOBALISE Project. 2025. "Globalise: Unlocking the History of Early Globalisation and Colonialism." Accessed July 17, 2025. https://globalise.huygens.knaw.nl/project_overview/.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, et al. 2022. "Training Compute-Optimal Large Language Models." In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, edited by **Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine**, 30016–30. New Orleans, LA: Curran Associates, Inc.

Hong, Jiwoo, Noah Lee, and James Thorne. 2024. "ORPO: Monolithic Preference Optimization without Reference Model." In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, edited by **Ruslan Salakhutdinov, Zico Kolter, and Kathellen Heller**, 18882–903. Vienna: PMLR.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. "Parameter-Efficient Transfer Learning for NLP." In *Proceedings of the 36th International Conference on Machine Learning (ICML*

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

2019), edited by **Kamalika Chaudhuri and Ruslan Salakhutdinov**, 2790–9. Long Beach, CA: PMLR.

Hu, Edward J., Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. 2024. "Amortizing Intractable Inference in Large Language Models." In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*. Vienna: ICLR.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*. Online: ICLR.

Huang, Xu, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. "Lost in the Source Language: How Large Language Models Evaluate the Quality of Machine Translation." In *Findings of the Association for Computational Linguistics: ACL 2024*, edited by **Lun-Wei Ku, Andre Martins, and Vivek Srikumar**, 3546–62. Bangkok: Association for Computational Linguistics.

Jiang, Zhaokun, and Ziyin Zhang. 2024. "Convergences and Divergences between Automatic Assessment and Human Evaluation." Preprint, arXiv:2401.05176 [cs].

Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. 2020. "Scaling Laws for Neural Language Models." Preprint, arXiv:2001.08361.

Kudo, Taku, and John Richardson. 2018. "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Generation." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by **Eduardo Blanco and Wei Lu**, 66–71. Brussels: Association for Computational Linguistics.

Li, Xiang Lisa, and Percy Liang. 2021. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by **Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli**, 4582–97. Online: Association for Computational Linguistics.

Lip, Gavin. 2025. "Emdutch-llm-finetune." https://github.com/glp500/EMDutch-LLM-Finetune. This repository is for the paper" Fine-Tuning Large-Language Models for Early Modern Dutch Translation" and contains all the notebooks and data used to run the experiments. By using this repository it is possible to recreate and rerun the trainings, tests, and evaluations.

Lopez, Adam. 2008. "Statistical Machine Translation." *ACM Computing Surveys (CSUR)* 40, no. 3: 1–49.

Manjavacas, Enrique, and Lauren Fonteyn. 2022. "Adapting Vs. Pre-Training Language Models for Historical Languages." *Journal of Data Mining & Digital Humanities*, *2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, 123–134.

Mielke, Sabrina J., Zaid Alyafeai, Elizabeth Salesky, et al. 2021. "Between Latin and Bytes: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP." Preprint, arXiv:2112.10508 [cs].

OpenAI. 2024. "GPT–4o Technical Report." Accessed July 15, 2024. https://openai.com/research/gpt-4o

Ouyang, Long, Jeff Wu, Xu Jiang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, edited by **Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine**, 27730–44. New Orleans, LA: Curran Associates, Inc.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, edited by **Pierre Isabelle**, 311–18. Philadelphia, PA: Association for Computational Linguistics.

Petram, Lodewijk, and Matthias Van Rossum. 2022. "Transforming Historical Research Practices—A Digital Infrastructure for the VOC Archives (GLOBALISE)." *International Journal of Maritime History* 34, no. 3: 494–502.

Post, Matt. 2018. "A Call for Clarity in Reporting Bleu Scores." Preprint, arXiv:1804.08771.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, edited by **Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine**, 27336–54. New Orleans, LA: Curran Associates, Inc.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. "Neural Machine Translation of Rare Words with Subword Units." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by **Katrin Erk and Noah A. Smith**, 1715–25. Berlin: Association for Computational Linguistics.

Stoler, Ann Laura. 2002. "Colonial Archives and the Arts of Governance." *Archival Science* 2, nos. 1–2: 87–109.

Terryn, Ayla Rigouts, and Miryam de Lhoneux. 2024. "Impact of English Bias of Generative Large Language Models in Dutch and French." In *Proceedings of HumEval @ LREC-COLING 2024*, 12–27. Turin: ELRA and ICCL.

UNESCO. 2003. "Archives of the Dutch East India Company (VOC), 1602–1795: Memory of the World Register." https://www.unesco.org/en/memory-world/archives-dutch-east-india-company. Inscribed in 2003.

Unsloth AI. 2024. "Unsloth: Efficient Fine-Tuning & Reinforcement Learning for Large Language Models." https://unsloth.ai. GitHub repository: https://github.com/unslothai/unsloth.

van Rossum, Matthias, Alexander Geelen, Bram Van den Hout, and Merve Tosun. 2020. *Testimonies of Enslavement: Sources on Slavery from the Indian Ocean World*. London, UK: Bloomsbury Publishing.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, edited by **Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S.V.N. Vishwanathan, and Roman Garnett**, 5998–6008. Red Hook, NY: Curran Associates, Inc.

Wang, Haifeng, Wu Hua, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. "Progress in Machine Translation." *Engineering* 18: 143–53.

Wolf, Thomas, Lysandre Debut, Victor Sanh, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by **Qun Liu and David Schuurmans**, 38–45. Online: Association for Computational Linguistics.

Zhang, Biao, and Rico Sennrich. 2020. "Improving Massively Multilingual Translation with Soft Decoupled Encoding." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by **Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault**, 7994–8005. Online: Association for Computational Linguistics.

Zhang, Tianyi, Varsha Kishore, Felix Wu, et al. 2020. "BERTScore: Evaluating Text Generation with BERT." In *8th International Conference on Learning Representations (ICLR 2020)*. Addis Ababa: ICLR.

Zhang, Xuan, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. "Machine Translation with Large Language Models: Prompting, Few-Shot Learning, and Fine-Tuning with QLoRA." In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, edited by **Philipp Koehn, Loïc Barrault, Ondřej Bojar**, et al., 468–81. Singapore: Association for Computational Linguistics.

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

Appendix

Table A1. Bad translation from early modern Dutch to English by [*Unsloth Mistral*]

Early modern Dutch text	English output
[Den Raad met aendagt gelesen end[e] geresumeert hebbende den schriftelijken crimineelen eijsch door den koopman en fiscaal deses commandements Nicolaas Bowijn ex off: eijsser op ende jegens den gev: christen lascorijn Jano, over het wegvoeren en verkopen van een Poeliase jongen, in naeme Chacken, toebehoorende sijn hoogheis den coning van Cochim, g'exhibeert, benevens de stucken en bewijsen tot adstru[. . .] [. . .]ewonnen, mits gad[. . .] gelet op 't geene te [. . .] en in de [. . .] ende doende [. . .] ex off: ged[a]en[e]n eijsch, en verklaert over sulx den gev: te sijn een man des doods,hem ten dien eijnde condemneerend[e] om g[e]bragt sijnde ter ordinaire geregts plaetse, alwaer men hier gewoon is crimineel sententien te executeeren, aldaer aen den scherpregter overgelevert, en door denselven met de koorte aan den galge gestraft te werden, dat'er de dood nae volgt, en desselvs doode lichaem nae het buijten geregt gebragt, en weder aan de aldaer staende galge tot proije van de lugt en de vogelen des hemels opgehangen, cum expensis; advoueert wijders des eijssers sijn gedaen versoek om met het believen van de overigheid deser custe [. . .].]	[['The Council having with attention read and resumed the written criminal eijsch by the merchant and fiscal of this district, Nicolaas Bowijn, ex officio eijsscher, against the imprisoned Christian lascorin Jano, about the abduction and selling of a Poeliase jongen, named Chacken, belonging to his highness the king of Cochim, exhibited, together with the documents and proofs collected for the validation of, if possible, [. . .] [. . .]ewonnen, mits gad[. . .] gelet op 't geene te [. . .] en in de [. . .] ende doende [. . .] ex off: ged[a]en[e]n eijsch, and declares the executioner of this sentence, whether man or woman, to be brought to the ordinary place of execution, where criminal sentences are usually executed, there to be handed over to the executioner by the eijsscher, and to be punished by him with the rope on the gallows, until death follows, and his dead body to be brought to the outer place of execution, and there once again to be hung up on the there-standing gallows as prey for the wind and the birds of the sky, cum expensis; furthermore endorses the eijsscher's request with the favour of the government of this coast, [. . .].']]

Table A2. Good translation from early modern Dutch to English by [*Unsloth Mistral*]

Early modern Dutch text	English output
[Aldus gedaan en gerelateert binnen destad Cochim ter ordinaris Raad camer op vrijdag den 2:en september anno 1757: in presentie van d'e Louis Quintin Martinsart en Laurens Trogh beijde leeden uijt den Raad voorm: die de minute deses nevens den relatant en mij gesw: clerq hebben onderteekent.]	[Thus done and testified in the city of Cochim at the ordinary chamber of Council on Friday, 2 September 1757 in the presence of the honourable Louis Quintin Martinsart and Laurens Trogh, both members of the aforementioned Council, who signed the original of this besides the testifier and me the sworn clerk.]

Table A3. Good translation from early modern Dutch to English by [*Llama & Assistant*]

Early modern Dutch text	English output
[Aldus gedaan en gerelateert binnen destad Cochim ter ordinaris Raad camer op vrijdag den 2:en september anno 1757: in presentie van d'e Louis Quintin Martinsart en Laurens Trogh beijde leeden uijt den Raad voorm: die de minute deses nevens den relatant en mij gesw: clerq hebben onderteekent.]	[Thus done and testified within the city of Cochim at the ordinary Chamber of Council on Friday, 2 September 1757 in the presence of the honourable Louis Quintin Martinsart and Laurens Trogh, both sworn members of the aforementioned Council, who have signed the original of this, besides the]

Table A4. Poor translation from early modern Dutch to English by [*Llama & Assistant*]

Early modern Dutch text	English output
[Compareerde andermaal voor de na te noemene gecommiteerde E:s leden uit denachtb:Raad van Justitite deezer stede degevangen Barkie, indevoorenstaande confessie breeder vermeld, welke hem nu wederom ten overstaan van den koopman en fiskaal deezes goevernements de heer m:r Jan Willem Hendrik van Rossum door mij gezw: klerk van woorde te woorde distinc [. . .] voorgeleezen, en door vertaaling van den tolk Dionijs Alewijn duidlijk te verstaan gegeven zijnde, bleef daarbij ten vollen persisteeren, eerlijk met deeze verandering opde afvraage van d:e: fiskaal, hoe hij konde opgeeven dat het meisje 2: a 3: dagen bij Maleparte Marien is geweest daar zij verklaard heeft van smorgens te 5 tot savonds te 5 uur bij haar gehouden te hebben, zo verklaard hij dat het de waarheid was, Äòt geen hij opgegeeven heeft Maliaparte Marian binnen gescheld en afgevraagd zijnde hoe lange het kind bij haar gebleven is, zeide zij maar een dag en beriep zich op het kind, die toen in judicio geciteerd en haar facie staande hield, dat zij vijf dagen bij Maliaparte Marian geweest is.]	[Compareerde andermaal voor de na te noemene gecommiteerde E:s leden uit den achtb: Raad van Justitie desezer stededee gevangen Barkie, in de voorstaande confessie breeder vermeld, welke hem nu wederom ten overstaan van den koopman en fiscaal deseves goevernement de heer m:r Jan Willem Hendrik van Rossum door mij gezw: klerk van woorde te woorde distincteerende, en door vertaaling van den tolk Dionijs Alewijn duidlijk te verstaan gegeven zijnde, bleef daarbij ten vollen persisteeren, eerlijk met deeze verandering op de afvraage van den eersten gevraagden fiscaal, hoe hij konde opgeeven dat het meisje 2 a 3 dagen bij Maleparte Marien is geweest daar zij verklaard heeft van smorgens te 5 tot savonds te 5 uur bij haar gehouden te hebben, zo verklaard hij dat het de waarheid was 't geen hij opgegeeven heeft]

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

Table A5. Translations of early modern Dutch text by different models

Early modern Dutch text: Aldus gedaan en verklaard binnen de stad Koetsiem ter ordinaire Raadkamer der Justietsie op woensdag den 25e juli 1792. in presentsie van d'E:s Pieter Elstendorp en Johan Christiaan Frischbier, leden, die de minute deezes neevens de comparante, voorsz: translateur en mij gezwoorene klerk hebben ondertekend.	
Model	Translation output
Human Expert	Thus done and testified within the city of Koetsiem in the ordinary Chamber of Council of Justice on Wednesday, 25 July 1792 in [the] presence of the honourable Pieter Elstendorp and Johan Christiaan Frischbier, members, who signed the original of this, besides the deponent, above-mentioned interpreter and me the sworn clerk.
GPT-4o	Thus done and declared within the city of Koetsiem in the ordinary Council Chamber of Justice on Wednesday the 25th of July 1792, in the presence of the Honourables Pieter Elstendorp and Johan Christiaan Frischbier, members, who have signed the minute hereof alongside the appearing, aforementioned translator and myself, sworn clerk.
Tiny Llama Unsloth	(Truncated raw output, contained instructions rather than a direct translation.)
Aya-23-8B	Thus done and declared within the city of Koesem to the ordinary Court of Justice on Wednesday the 25th of July 1792. In the presence of the E:s Pieter Elstendorp and Johan Christiaan Frischbier, members, who have signed this minute along with the comparant, the translator, and . . .
Llama & Assistant	Thus done and stated within the city of Koetsiem in the ordinary Chamber of Council of Justice on Wednesday, 25th July 1792 in the presence of d'E:s Pieter Elstendorp and Johan Christiaan Frischbier, members, who signed the original of this, besides the confessant, . . .
Llama 3 Unsloth	Thus done and stated within the city of Koetsiem . . . on Wednesday, 25 July 1792 . . . who signed the original of this, besides the deponent, above-mentioned interpreter and me the sworn . . .
Meta-Llama-3-8B-Instruct	Thus done and declared within the city of Koetsiem in the ordinary Council Chamber of the Justices on Wednesday, July 25th, 1792. In the presence of Mr. Pieter Elstendorp and Johan Christiaan Frischbier, members who have signed this minute along with the translator and myself, sworn clerk.
Llama 2 Unsloth	Thus done and stated within the city of Koetsiem . . . in the presence of d'E:s Pieter Elstendorp and Johan Christiaan Frischbier . . . who have signed the original of this, besides the confessant, aforement
Mistral Unsloth	Thus done and stated within the city of Koetsiem at the ordinary Chamber of Council of Justice on Wednesday, 25 July 1792 . . . who signed the original of this, besides the deponent, above-mentioned interpreter and me the sworn . . .
Mistral-7B-Instruct-v0.3	Thus done and declared within the city of Coetsiem at the ordinary Council Chamber of Justice on Wednesday the 25th of July 1792, in the presence of the Es:s Pieter Elstendorp and Johan Christiaan Frischbier, members, who also signed this minute . . .
Llama 3 ORPO	(Truncated raw output, contained instructions rather than a direct translation.)

Table A6. Training arguments for different models

Model	Training arguments
Meta-Llama-3-8B Instruct	per_device:train_batch_size = 2, gradient_accumulation_steps = 4, warmup_steps = 5, num_train_epochs = 9, learning_rate = 2e-4, fp16 = not torch.cuda.is_bf16_supported(), bf16 = torch.cuda.is_bf16_supported(), logging_steps = 1, optim = "adamw_8bit", weight_decay = 0.01, lr_scheduler_type = "linear", seed = 3407
Meta-Llama-2-7B Instruct	per_device:train_batch_size = 2, gradient_accumulation_steps = 4, warmup_steps = 5, num_train_epochs = 9, learning_rate = 2e-4, fp16 = not torch.cuda.is_bf16_supported(), bf16 = torch.cuda.is_bf16_supported(), logging_steps = 1, optim = "adamw_8bit", weight_decay = 0.01, lr_scheduler_type = "linear", seed = 3407
Mistral-7B-Instruct-v0.3	per_device:train_batch_size = 2, gradient_accumulation_steps = 4, warmup_steps = 5, num_train_epochs = 9, learning_rate = 2e-4, fp16 = not torch.cuda.is_bf16_supported(), bf16 = torch.cuda.is_bf16_supported(), logging_steps = 1, optim = "adamw_8bit", weight_decay = 0.01, lr_scheduler_type = "linear", seed = 3407
TinyLlama_-v1.1	per_device:train_batch_size = 2, gradient_accumulation_steps = 4, warmup_steps = 5, num_train_epochs = 9, learning_rate = 2e-4, fp16 = not torch.cuda.is_bf16_supported(), bf16 = torch.cuda.is_bf16_supported(), logging_steps = 1, optim = "adamw_8bit", weight_decay = 0.01, lr_scheduler_type = "linear", seed = 3407
Meta-Llama-3-8B Instruct	learning_rate = 8e-6, lr_scheduler_type = "linear", max_length = 1024, max_prompt_length = 512, beta = 0.1, per_device:train_batch_size = 2, per_device:eval_batch_size = 2, gradient_accumulation_steps = 4, optim = "paged_adamw_8bit", num_train_epochs = 8, evaluation_strategy = "steps", eval_steps = 0.2, logging_steps = 1, warmup_steps = 10

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

Table A7. Translation responses of various models

Early modern Dutch text: Weledele achtbaare heer en heeren!	
Model	Translation output
Human Expert	Right honourable lord and lords!
GPT-4o	Most honorable sir and sirs!
Tiny Llama Unsloth	(Truncated raw output: removed extra code)
Aya-23-8B	Dearest noble lord and lords!
Llama & Assistant	[signed] Honourable lord and lords
Llama 3 Unsloth	Honourable lord and lords!
Meta-Llama-3-8B-Instruct	(Truncated: it just prompted to translate)
Llama 2 Unsloth	[signed] Honourable lord and lords
Mistral Unsloth	Honourable lord and lords!
Mistral-7B-Instruct-v0.3	Worthy honorable sirs and lords
Llama 3 ORPO	(Truncated: it just prompted to translate)

Table A8. Good translation from early modern Dutch to English by [*Mistral-7B-Instruct-v0.3*]

Early modern Dutch text	English output
[Interrogatoria voor Maleparte Marian, C.]	[Interrogatories for Maleparte Marian]

Table A9. Good translation from early modern Dutch to English by [*Mistral-7B-Instruct-v0.3*]

Early modern Dutch text	English output
[Antw: Ik heb het bedrogh al lang van te vooren ontdekt.]	[Answer: I have long since discovered the deception.]

Table A10. Statistics of BERT scores for all models

Model	Mean	Median	Std	Min	Max	Range
gpt-4o	0.8011	0.8044	0.0784	0.5831	1.0000	0.4169
Tiny Llama Unsloth	0.3822	0.3749	0.0769	0.2320	0.9178	0.6857
aya-23-8B	0.6669	0.6761	0.1126	0.3357	0.9330	0.5973
Llama & Assistant	0.7908	0.8027	0.1111	0.2790	1.0000	0.7210
Llama 3 Unsloth	0.8522	0.8613	0.1026	0.2583	1.0000	0.7417
Meta-Llama-3-8B-Instruct	0.5635	0.6046	0.1812	0.2468	0.8898	0.6430
Llama 2 Unsloth	0.7858	0.7981	0.1240	0.3257	1.0000	0.6743
mistral unsloth	0.8640	0.8717	0.0863	0.5164	1.0000	0.4836
mistral-7B-instruct-v0.3	0.7281	0.7261	0.0835	0.4822	0.9480	0.4658
Llama 3 ORPO	0.6170	0.6649	0.1796	0.2757	0.9661	0.6904

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

Table A11. Statistics of METEOR scores for all models

Model	Mean	Median	Std	Min	Max	Range
gpt-4o	0.6648	0.6646	0.1467	0.1667	0.9993	0.8326
Tiny Llama Unsloth	0.0847	0.0721	0.1002	0.0000	0.8470	0.8470
aya-23-8B	0.4604	0.4605	0.1624	0.0617	0.8436	0.7819
Llama & Assistant	0.6454	0.6669	0.1904	0.0000	0.9995	0.9995
Llama 3 Unsloth	0.7428	0.7538	0.1705	0.0000	0.9999	0.9999
Meta-Llama-3-8B-Instruct	0.3361	0.4006	0.2414	0.0000	0.8018	0.8018
Llama 2 Unsloth	0.6412	0.6523	0.2054	0.0287	0.9995	0.9708
mistral unsloth	0.7585	0.7763	0.1612	0.2909	1.0000	0.7091
mistral-7B-instruct-v0.3	0.5317	0.5355	0.1478	0.1724	0.8951	0.7226
Llama 3 ORPO	0.4050	0.4611	0.2457	0.0000	0.9041	0.9041

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press

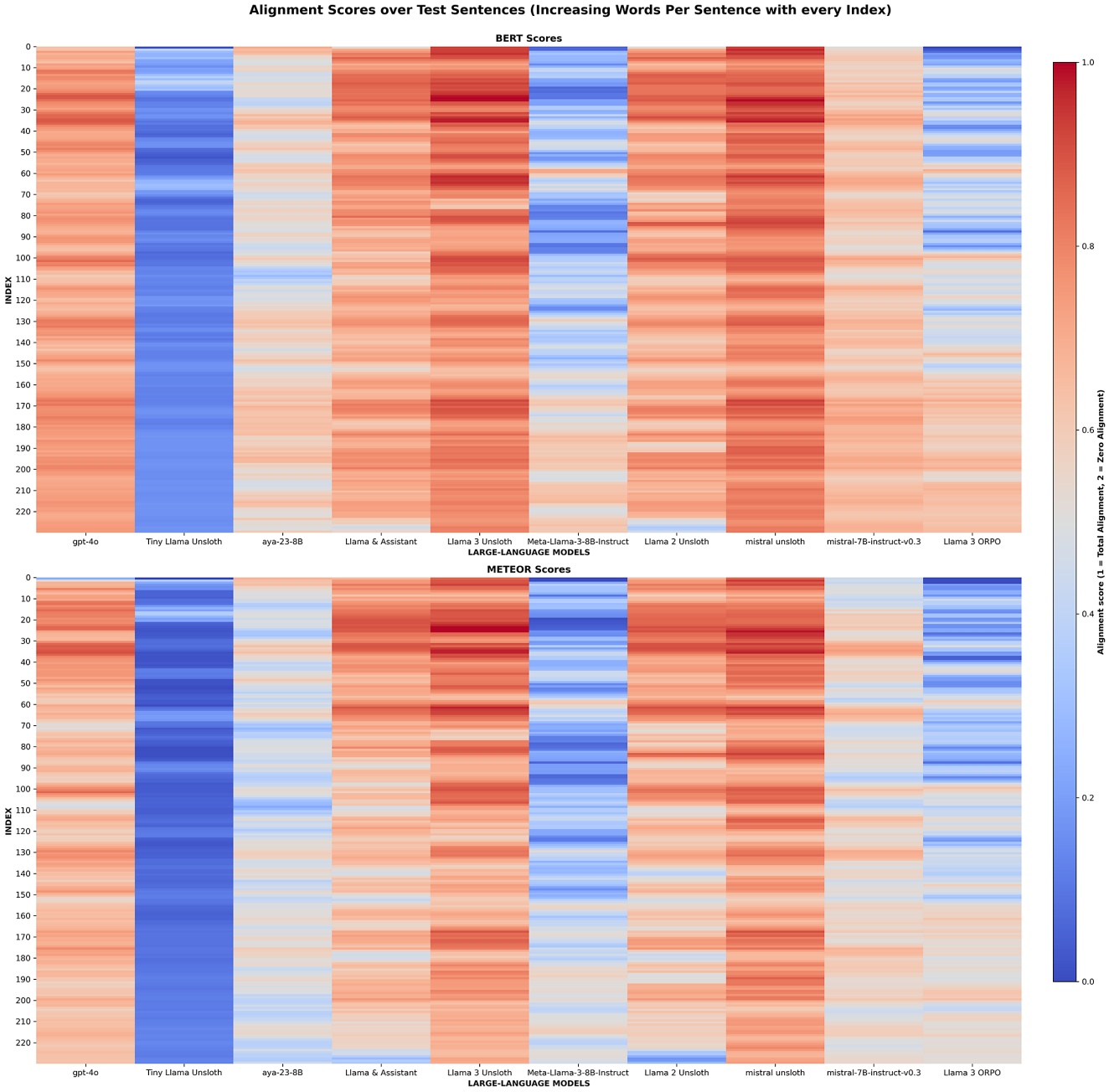


Figure A1. BERT and METEOR score heat map.

https://doi.org/10.1017/chr.2026.10022 Published online by Cambridge University Press