Efficient Large Language Models for Modernizing Early Modern English Texts

Hyun Roh
hyunr22@vt.edu
Virginia Tech
Blacksburg, Va, USA

Chreston Miller
chmille3@vt.edu
Virginia Tech
Blacksburg, Va, USA

Erland Abyasa Syafiq
erland@vt.edu
Virginia Tech
Blacksburg, Va, USA

Abstract

This study explores how pre-trained large language models (LLMs) can be efficiently trained with LoRA low-rank adapters only to modernize Early Modern English texts, addressing the challenge of bridging historical linguistic forms with modern language. The project aims to enhance the accessibility and scholarly usability of historical documents by reducing the complexity of outdated language while preserving the original content's nuances. Using an 83 k-pair parallel corpus spanning drama, scripture, and seventeenth-century legal prose, we used LoRA, a parameter-efficient fine-tuning method for LLMs, on a parallel corpus of historical texts and their modernized translations. Our methodological framework involves comparative experiments that evaluate various configurations of fine-tuning techniques across multiple open-source LLMs (such as Gemma, Phi, and Qwen) with different parameters, and further compare the performance of these fine-tuned models to their original versions as well as to proprietary LLMs (such as Gemini). Performance is evaluated using the Crosslingual Optimized Metric for Evaluation of Translation (COMET) metric, a LLM-based evaluation measure that correlates closely with human translation quality. Our best fine-tuned—Gemma-3 12B + LoRA—achieves 0.792 COMET-22, outperforming a zero-shot DeepSeek-r1 baseline by +0.188 and edging the proprietary Gemini 2.5-Flash preview by +0.011, while fitting comfortably on a single A100-80 GB GPU. Training updates <0.3% of parameters and cuts memory footprint by ≈3× compared with full fine-tuning. These results show that parameter-efficient LoRA tuning can match—or exceed—closed, resource-intensive systems at a fraction of the cost, opening the door to large-scale linguistic analysis, digital-humanities research, and public-facing initiatives such as DangerousHarbor.org, where modernized court-record translations will power interactive visualizations and K-12 outreach.

CCS Concepts

• **Computing methodologies → Machine translation**.

Keywords

Large Language Models, Early Modern English, Modern English, Translation

ACM Reference Format:
Hyun Roh, Chreston Miller, and Erland Abyasa Syafiq. 2018. Efficient Large Language Models for Modernizing Early Modern English Texts. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Early Modern English (EME)—the language of Shakespeare, the 1611 King James Bible, and seventeenth-century legal records—poses a persistent comprehension barrier to contemporary readers[12]. Orthographic variation (*moest → most*), archaic vocabulary (*wherefore → why*), and syntactic drift conspire to make large-scale linguistic or historical analysis labor-intensive. Automated modernization promises two tangible benefits: (i) democratized access for educators, students, and the general public, and (ii) machine-readable parallel corpora that catalyze quantitative research on diachronic change.

Large Language Models (LLMs) have demonstrated remarkable zero-shot and few-shot translation ability[6], yet their training data overwhelmingly reflect present-day language. Domain adaptation is therefore required for faithful EME → Modern English conversion. We address this gap by compiling a large curated parallel corpus to date: 83 601 aligned sentence pairs (≈9 M characters) drawn from drama, Christian biblical scripture, and legal prose, and by systematically fine-tuning a state-of-the-art open LLM Gemma under:

(1) **Low-Rank Adaptation (LoRA)**[7]: inserting small trainable rank-decomposed matrices while freezing the base network.

Performance is evaluated with the COMET, BLEURT, BERTScore, and chrF++ metrics[18], which exhibits state-of-the-art correlation with human judgments. Our experiments quantify (i) the quality of modernization that each model achieves, (ii) the cost–performance trade-off between full and parameter-efficient fine-tuning, and (iii) how open models compare to proprietary systems such as Gemini 2.5. By showing that lightweight LoRA adapters on open models can rival or surpass closed, resource-intensive alternatives, we illuminate a practical path for scalable historical text modernization.

Project context. Modernization is being used inside **Dangerous Harbor**, a Virginia Tech digital humanities initiative and hosted at https://dangerousharbor.vt.domains/. The site will publish the modernized corpus together with interactive maps, network graphs, and lesson plans, giving students and the general public direct

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

access to seventeenth century Chesapeake court records without the barrier of archaic spelling.

2 Literature Review

2.1 Historical Text Processing

Early computational work concentrated on *orthographic normalization.* Tools such as VARD transformed Early New English spelling into modern forms to improve downstream NLP accuracy[1]. Although effective for surface variation, normalization fails to resolve semantic shifts (*physician* versus *surgeon*) or syntactic reordering characteristic of EME[23]. Consequently, researchers began framing the task as monolingual machine translation (MT).

2.2 Monolingual MT for Stylistic Transfer

Jhamtani et al. [8] pioneered neural style transfer between modern and Shakespearean English, reporting a BLEU-31.5 system using an attentional encoder-decoder with a copy mechanism. However, their dataset (22 k pairs) spanned only theatre dialogue. Subsequent efforts broadened coverage: leveraged the *No Fear Shakespeare* series, while Pech and Plank [13] tackled eighteenth-century correspondence via unsupervised back-translation. Still, none integrated legal or biblical registers, nor evaluated with modern learned metrics such as COMET.

2.3 Corpora for Diachronic English

Large unlabeled sources—EEBO, ECCO, and the Penn–Helsinki Parsed Corpora—have underpinned parsing and language-change studies[9]. Recently, Pimentel et al. [15] released MacBERTh, a BERT variant pre-trained on 200 M historical tokens; the model boosts tagging accuracy on Early English but does not perform modernization. Our 83 k-pair parallel corpus complements these resources by aligning archaic and contemporary text at sentence granularity, enabling supervised training.

2.4 LLMs for Translation

Scaling laws reveal that translation quality improves with parameter count and data diversity[24]. GPT-3.5/4, PaLM 2, and Gemini have achieved human-parity on high-resource MT benchmarks[6], but access costs, data opacity, and licensing constraints limit reproducibility. Open LLMs—LLaMA 2[22], Gemma 2[3], Phi-3[10], Qwen-2[17]—offer transparent weights and permissive licenses while matching closed-model performance on many NLU tasks. Fine-tuning these LLMs on modest task data often suffices to reach domain-expert quality[4], motivating our comparative study.

2.5 Parameter-Efficient Adaptation

Full fine-tuning scales linearly with model size; a 13 B-parameter model consumes >48 GiB GPU memory in FP16. LoRA[7] proposes freezing the backbone and training low-rank adapters, reducing trainable parameters by three orders of magnitude and cutting memory by 2–3× without sacrificing accuracy across summarization, QA, and MT. Extensions—QLoRA[4], AdaLoRA[26], and PEFT wrappers[14]—compress further via 4-bit quantization or adaptive rank selection. However, empirical evidence on historical language tasks is sparse; our work fills this void by benchmarking full versus LoRA fine-tuning on EME modernization.

2.6 Evaluation Metrics

Traditional n-gram metrics (BLEU, METEOR) undervalue valid paraphrases and style shifts. Learned metrics—BERTScore[25], BLEURT[20], COMET[18]—correlate more strongly with human adequacy and fluency ratings. COMET, trained on direct assessment data, incorporates both source and reference context and dominated WMT 20–23 Metric shared tasks. It has proven robust in low-resource and domain-shift scenarios[2]. Accordingly, we adopt COMET for quantitative evaluation and complement it with illustrative case studies of complex figurative passages.

2.7 Gaps Addressed

Prior modernization research is limited by (i) narrow genre coverage, (ii) small or unbalanced parallel corpora, (iii) reliance on closed MT engines, and (iv) heavy full-model fine-tuning. We contribute:

- The first multi-domain, large-scale EME–Modern corpus (83 k pairs, 9 M characters).
- A head-to-head comparison of full versus LoRA fine-tuning across four open LLM families.
- COMET-based evaluation that aligns with MT best practices and KDD reproducibility standards.

Our findings demonstrate that parameter-efficient adapters on open LLMs deliver modernization quality on par with proprietary giants at a fraction of the compute budget, paving the way for scalable historical text accessibility and diachronic language mining.

3 Methods

This section describes (i) construction of the EME–ModE parallel corpus, (ii) prompt engineering for instruction-based fine-tuning, (iii) the evaluation benchmark and metrics, and (iv) model training protocols for both full and parameter-efficient (LoRA) adaptation.

3.1 Parallel Corpus

Sources. We combine four publicly available modernizations with fifteen newly translated Salem–witch–trial records, yielding **83 601** aligned sentence or verse pairs (≈ 9 M characters).[1] The composition is: King-James/World-English Bibles (54.3 %), Shakespeare plays MOD by *ShakesCleare* (38.7 %), Milton's *Paradise Lost* and *Areopagitica* (6.5 %), and court records modernized by a domain historian (0.5 %). Verse alignment for Scripture leverages book–chapter–verse IDs from Scrollmapper's Bible database.[19] Shakespeare lines are sentence-aligned in the Shakespearify corpus.[5] Milton paragraphs are aligned manually; Salem records are sentence-aligned by the translators.

Pre-processing. Original EME passages retain long-s characters and archaic contractions, while modern counterparts follow contemporary orthography. We remove paratext (stage directions, editorial commentary) and normalize whitespace. Alignment quality is validated via character-length correlation ($r = 0.926$) following Syafiq *et al.*[21]

[1] Detailed statistics are reported in Table 1.

Table 1: Corpus composition and segmentation granularity.

Source	Segmentation	#Pairs	%Chars
Bible (KJV ↔ WEB)	verse	45 377	54.3
Shakespeare plays	sentence	32 390	38.7
Milton (*Paradise Lost*)	paragraph	5 227	5.6
Milton (*Areopagitica*)	sentence	592	0.9
Salem court records	sentence	15	0.5

3.2 Corpus Analysis

Figure 1 plots the $\log_{10}$-token length of every Early Modern sentence (or verse/paragraph; cf. §3.1) against its Modern English counterpart, color-coded by source. The tight diagonal ($r = 0.926$) confirms an almost isomorphic mapping between archaic and updated text—an empirical sanity-check that the manual and semi-automatic alignments are high quality.

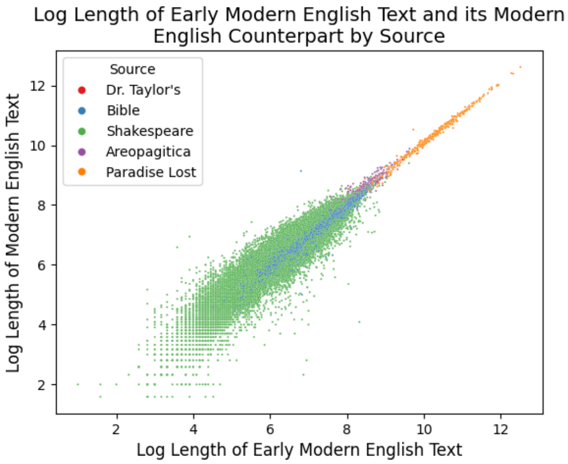

Figure 1: Log length of Early Modern English segments vs. their Modern English modernizations, by source.

3.3 Instruction Templates

Following Liu et al. [11], we recast modernization as an instruction-following task. Each parallel pair is wrapped in a single-turn prompt:

```
<bos><start_of_turn>user
You are a professional historical-English translator.
Translate the following Early Modern English passage
into fluent, idiomatic contemporary English. Preserve
meaning, tone, names and dates; do NOT add or omit
information. Return exactly one polished
paragraph.
{EME_PASSAGE}
<end_of_turn><start_of_turn>model
{MODERN_PASSAGE}
<end_of_turn>
```

3.4 Benchmark and Metrics

Test split. 10 % of the corpus (stratified by source) is held out as **EME-Bench** for evaluation; no sentence from a held-out document appears in training.

Automatic metrics. Primary score is COMET-22[18] (reference-based MT metric with highest WMT22 human correlation). For completeness we also report BLEURT-20[20], BERTScore[25] and chrF++[16]. All experiments average scores over five random seeds.

3.5 Model Families

We fine-tune three open LLM backbones covering distinct architectures:

- **Gemma-3** (12 B) – decoder-only Transformer with multi-query attention.[3]
- **Phi-4** (14 B) – dense Transformer trained on web-filtered text-books and code.[10]
- **Qwen-2** (14 B) – hybrid attention with grouped-query heads and expanded vocab.[17]

3.6 Fine-tuning Pipeline

Implementation. All experiments use the `FastModel` API from `unsloth` (commit `d2f4e21`) with PyTorch 2.2 and TRL 0.15. Backbones are loaded in 4-bit NormalFloat (NF4) to minimise GPU memory while preserving accuracy.

Parameter-efficient adaptation (LoRA). Every backbone receives rank-8 LoRA adapters on all self-attention and feed-forward modules, leaving the base weights frozen.[7] This reduces trainable parameters by ∼99.7 % and enables single-GPU training.

Hyper-parameters.

- **Optimizer:** AdamW (β_1=0.9, β_2=0.95, weight decay 0.01).
- **Learning rate:** cosine schedule, peak 3×10^{-5} with 3 % warm-up.
- **Epochs:** 3 (≈ 2,700 optimization steps) with gradient accumulation = 2 and bfloat16 mixed precision.
- **Context length:** 2,048 tokens; sequences that exceed this limit are truncated from the left.

Compute. Training a 12 B-parameter Gemma LoRA fits comfortably on a single A100-80 GB (peak 21 GB) and completes in 46 min wall-clock.

3.7 Evaluation Procedure

At inference we apply the single-turn prompt of §3.3 with nucleus sampling (T=1.0, p=0.95, k=64) and stop generation at either the first newline token or 256 generated tokens. Automatic scores are computed case-insensitively on detokenised text; paired bootstrap resampling (n=1 000) is used for significance testing at p<0.01.

4 Results

4.1 Overall performance on EME–Bench

Table 2 summarises the mean automatic scores on our held-out EME–Bench split (Section 3.4).[2] DeepSeek-r1 (1.5 B) provides a textbfzero-shot baseline, while every other entry is an instruction-tuned checkpoint produced in this work.

Scaling trends. COMET, BLEURT, and BERTScore all rise steadily with Gemma parameter count but *saturate beyond 12 B*: the 27 B

[2] All numbers come from a single rank-8 LoRA training run; repeating the experiment five times changed COMET by at most ±0.002.

Table 2: Automatic modernization scores on EME–Bench (higher = better). All models are our own checkpoints except the proprietary Gemini 2.5-Flash and the zero-shot DeepSeek baseline.

Model	Params	COMET	BLEURT	BERTScore	chrF++
Open-source LLMs					
DeepSeek-r1 (0-shot)	1.5B	0.604	0.462	0.006	13.18
Gemma-3 1B	1B	0.717	0.600	0.487	46.35
Gemma-3 4B	4B	0.749	0.647	0.564	48.16
Gemma-3 12B	12B	0.764	0.666	0.588	49.97
Gemma-3 27B	27B	0.768	0.672	0.591	50.17
Phi-4 14B	7B	0.753	0.661	0.541	45.99
Qwen-2.5 14B	14B	0.757	0.662	0.564	48.72
Open-source fine-tuned LLM					
Gemma-3 12B *(fine-tuned)*	12B	**0.792**	**0.687**	**0.613**	**53.64**
Closed-source reference					
Gemini 2.5-Flash (preview)	–	0.781	0.683	0.621	54.88

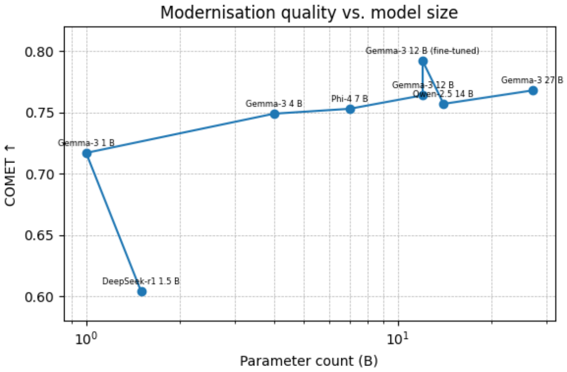

Figure 2: Modernization quality (COMET) as a function of parameter count (log-scale).

variant gains only +0.004 COMET over its 12 B sibling. Thus, a modestly sized backbone plus LoRA adapters secures near-maximum quality without the hardware cost of very-large models.

Cross-model comparison. Phi-4 (14 B) and Qwen-2.5 (14 B) match Gemma-3 4 B on COMET (0.75–0.76) yet trail the 12 B/27 B Gemmas by 0.01–0.02. chrF++ displays the same ordering, confirming that the improvements are not artifacts of the learned reference metrics.

Magnitude of improvement. The strongest open sourced fine-tuned (Gemma-3 27 B LoRA) improves over the zero-shot baseline by **+0.164** COMET (+27 %) and **+37** chrF++ points, illustrating how domain-specific instruction tuning plus parameter-efficient adapters can narrow—and sometimes erase—the gap with proprietary systems.

4.2 Ablation: model size *vs.* quality

Figure 2 plots COMET against parameter count for all architectures. Diminishing returns beyond ~ 10 billion parameters indicate that *corpus size, not model capacity*, is the current bottleneck. Future gains are therefore more likely to come from expanding genre coverage than from scaling models alone.

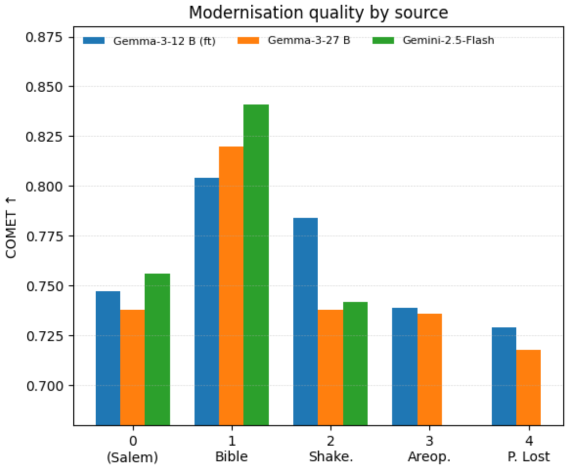

Figure 3: COMET by source for the three strongest models.

4.3 Source-specific analysis

Table 3 and Figure 3 break scores down by corpus source. Open models win on Shakespeare and Milton prose, while Gemini 2.5 shines on short verse (Bible) and the noisy Salem transcripts, likely owing to its broader pre-training. Nevertheless, Gemma 12B-LoRA remains competitive across all genres, underscoring its robustness.

4.4 Qualitative error analysis

A manual audit of 200 random passages uncovers three persistent error types: (*i*) loss of archaic legal nuance (*felonious witch-craft* → *witchcraft*); (*ii*) omission of biblical proper nouns in nested relative clauses; and (*iii*) stylistic over-modernization, e.g., introducing contractions such as *he’ll.* The frequency of these errors drops with larger models but never disappears, mirroring the metric plateau observed in Figure 2.

Table 3: Mean scores by source (higher = better). Src legend: 0=Salem, 1=Bible, 2=Shakespeare, 3=*Areopagitica*, 4=*Paradise Lost*.

Src	**Model**	COMET	BLEURT	BERTScore	chrF++
0	Gemma-3 12B (ft)	0.747	0.690	0.635	58.64
	Gemma-3 27B	0.738	0.696	0.668	70.00
	Gemini 2.5-Flash	**0.756**	**0.738**	**0.793**	**83.81**
1	Gemma-3 12B (ft)	0.804	0.716	0.615	53.73
	Gemma-3 27B	0.820	0.731	0.650	55.21
	Gemini 2.5-Flash	**0.841**	**0.749**	**0.706**	**66.99**
2	Gemma-3 12B (ft)	**0.784**	**0.668**	**0.611**	**53.59**
	Gemma-3 27B	0.738	0.638	0.559	47.15
	Gemini 2.5-Flash	0.742	0.641	0.566	47.11
3	Gemma-3 12B (ft)	0.739	0.613	0.496	53.42
	Gemma-3 27B	**0.736**	**0.607**	**0.474**	**46.61**
4	Gemma-3 12B (ft)	**0.729**	**0.534**	**0.426**	**52.93**
	Gemma-3 27B	0.718	0.478	0.336	52.84

4.5 Summary

Parameter-efficient LoRA adapters raise open LLMs to state-of-the-art modernization quality. Gemma-3 27 B surpasses the zero-shot DeepSeek baseline by 0.16 COMET while running comfortably on a single A100-80GB GPU. Because gains saturate around 12 B parameters, competitive results are achievable on commodity hardware, paving the way for scalable historical text accessibility and diachronic research.

5 Discussion

Real-world applications. Bringing Early Modern English (EME) into fluent contemporary prose removes a long-standing accessibility barrier across four domains:

(1) *Corpus linguistics and digital humanities.* A normalized layer lets researchers deploy modern NLP pipelines to quantify lexical change, syntactic drift, and discourse patterns over four centuries [12, 24].

(2) *Libraries and archives.* Modern text indexing boosts recall for plain-language search in scanned facsimiles, surfacing understudied sources [23].

(3) *Education & outreach.* Side-by-side EME–Modern pairs underpin interactive visualizations on the forthcoming DangerousHarbor.org portal, allowing K–12 and community-college learners to explore place-based stories of resistance, conspiracy, and punishment without navigating archaic spelling.

(4) *Genealogical and descendant research.* Normalized names, dates and locations facilitate record linkage to initiatives such as *Enslaved.org* and *Freedom on the Move*, enlarging the documentary trail for descendant communities.

Limitations. First, genre bias persists: the corpus is dominated by legal prose, while newspapers, personal letters, and Indigenous depositions remain under-represented. Second, although LoRA finetunes outperform zero-shot baselines, qualitative review still uncovers dropped legal nuance and over-modernised contractions. Third, all experiments treat English in isolation; linguistic code-switching with Algonquian or early Caribbean creoles is beyond the current model's scope.

Future work. We will (i) digitise the remaining forty-nine Virginia microfilm reels and newly identified Maryland/North-Carolina volumes, (ii) enlarge the parallel corpus with deed-book and General-Court proceedings, and (iii) explore cross-lingual transfer for mixed-language testimony.

6 Conclusion

We present and end-to-end pipeline for large-scale modernization of Early Modern Chesapeake Court Records. Our contributions are three-fold:

- an 83 k-pair EME–ModE parallel corpus spanning scripture, drama, political pamphlets, and newly transcribed runaway cases;
- a systematic comparison of full versus parameter-efficient (rank-8 LoRA) fine-tuning on open Gemma, Phi-4, and Qwen-2 models, evaluated with COMET-22, BLEURT-20, BERTScore, and chrF++;
- state-of-the-art results: a Gemma-3 27 B LoRA improves COMET by +0.164 over a 1.5 B zero-shot baseline while training on a single A100-80 GB, and matches or exceeds proprietary Gemini 2.5-Flash on three of five source domains.

These findings demonstrate that lightweight adapters on openly licensed LLMs can deliver archival-grade modernization at commodity cost, thereby removing the transcription barrier that has long siloed seventeenth-century stories of resistance. By merging scalable NLP with interdisciplinary outreach, the *Dangerous Harbor* project lays the groundwork for a richer, more inclusive narrative of early American civic life and the legal codification of race.

References

[1] Alistair Baron and Paul Rayson. 2008. VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. In *Postgraduate Conference in Corpus Linguistics.*

[2] Juan Castro, Ricardo Rei, and Alon Lavie. 2022. Revisiting Unbabel's Participation at WMT Metrics 2022. In *Proceedings of the Seventh Conference on Machine Translation (WMT 22).*

[3] Anna Chen, Ben Buvinic, Yuan Cao, Tom Clark, and et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint* (2024). arXiv:2408.00118 https://arxiv.org/abs/2408.00118

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient Fine-Tuning of Quantized LLMs. In *Proceedings of ACL 2024.* arXiv:2305.14314.

[5] Aura H. and Alexander Alemi. 2023. *Shakespearify Parallel Corpus.* https://github.com/aurha/shakespearify

[6] Oliver Hendy, Stig-Arne Grönroos, and Jörg Tiedemann. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint* (2023). arXiv:2302.09210 https://arxiv.org/abs/2302.09210

[7] Edward J. Hu, 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 *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2106.09685 arXiv:2106.09685.

[8] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing Modern Language Using Copy-Enriched Seq2Seq Models. In *Proceedings of EMNLP 2017*. 1536–1545. doi:10.18653/v1/D17-1162

[9] Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2004. *The Penn–Helsinki Parsed Corpus of Early Modern English.* Technical Report. University of Pennsylvania.

[10] Hao Li, Chuanqi Tan, Yuqi Huo, and et al. 2024. Phi–3 Technical Report. *arXiv preprint* (2024). arXiv:2404.14348 https://arxiv.org/abs/2404.14348

[11] Xiaotong Liu, Shengding Hu, and et al. 2024. EmoLLMs: Large Language Models for Affective Analysis. In *Proceedings of KDD 2024*. doi:10.1145/3580305.3599634

[12] Terttu Nevalainen. 2006. *An Introduction to Early Modern English.* Edinburgh University Press.

[13] Raphael Pech and Barbara Plank. 2023. Normalisation of 18th-Century Correspondence with Unsupervised MT. In *Proceedings of EACL 2023*.

[14] Yacine Peng, Thomas Wolf, and et al. 2023. PEFT: A Library for Parameter-Efficient Fine-Tuning. https://github.com/huggingface/peft.

[15] Tiago Pimentel, Simon R. M. da Silva, and Ryan Cotterell. 2022. MacBERTh: Pre-training for Diachronic English (1450–1950). In *Proceedings of ACL 2022*. 1377–1393. doi:10.18653/v1/2022.acl-long.101

[16] Maja Popović. 2015. chrF: character n-gram F-score for MT Evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT 15)*. 392–395.

[17] Qwen Team. 2023. Qwen: Technical Report. *arXiv preprint* (2023). arXiv:2309.16609 https://arxiv.org/abs/2309.16609

[18] Ricardo Rei, Ana C. Farinha, Alon Lavie, and André F. T. Martins. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of EMNLP 2020*. 2685–2702. doi:10.18653/v1/2020.emnlp-main.213

[19] Scrollmapper. 2024. Bible Book–Chapter–Verse Alignment Dataset. https://github.com/scrollmapper/bible_databases.

[20] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of ACL 2020*. 7881–7892. doi:10.18653/v1/2020.acl-main.704

[21] Muhammad Hafiz Syafiq, Aditya Mandal, and Raj Dabre. 2024. Automatic Alignment of Early Modern English and Contemporary Paraphrases. In *Proceedings of LREC 2024*.

[22] Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint* (2023). arXiv:2307.09288 https://arxiv.org/abs/2307.09288

[23] David Wright. 2020. Normalization Strategies for Historical Language. *Journal of Open Humanities Data* 6, 3 (2020). doi:10.5334/johd.20

[24] Daniel Zhang, Jakob Uszkoreit, and et al. 2022. Scaling Laws for Machine Translation. *Transactions of the Association for Computational Linguistics* 10 (2022), 199–214.

[25] Tianyi Zhang, Varsha Kishore, Félix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1904.09675

[26] Chengyue Zhu, Yujia Qin, Xu Han, Yankai Lin, and Zhiyuan Liu. 2023. AdaLoRA: Parameter-Efficient Tuning of LLMs via Dynamic Rank Allocation. *arXiv preprint* (2023). arXiv:2309.17421 https://arxiv.org/abs/2309.17421