NassEra: Temporal Analysis of Arabic Corpora with Transformer and Generative Models

Anonymous ACL submission

Abstract

NassEra presents a unified framework for chronological modeling and linguistic analysis of historical Arabic texts using modern neural architectures. The study formulates year prediction as a continuous regression task and compares two paradigms: Transformer-based encoders (DistilBERT, CAMeLBERT, AraBERT, mBERT) and generative models fine-tuned through Low-Rank Adaptation (LoRA). Using a balanced corpus of over 39,000 Arabic text segments spanning twelve centuries (203–1423 AH), Arabic-specific models outperform multilingual baselines. Among encoders, **AraBERT v2** achieved the lowest error (RMSE = 66.06, MAE = 36.93), confirming the benefit of Arabic-focused pretraining. Among generative models, **ALLaM-7B-Instruct** fine-tuned with LoRA achieved the highest temporal precision (MAE = 63.6, RMSE = 128.9, Pearson $r = 0.94$) while updating only 0.12% of parameters. The updated **Qwen2.5-7B-Instruct** model showed competitive results (MAE = 79.4, RMSE = 148.2, $r = 0.88$), indicating strong chronological sensitivity with reduced computational cost. Encoders achieved lower numeric error, whereas generative models attained higher correlation-based precision. NassEra demonstrates that both architectures effectively capture diachronic linguistic patterns in Arabic, offering scalable tools for manuscript dating and digital humanities research.

1 Introduction

The growing availability of large digitized Arabic corpora has created new opportunities to study linguistic change across Islamic history. Classical and post-classical Arabic, spanning fourteen centuries of written tradition, display significant variation in morphology, orthography, and style. Modeling how these linguistic forms evolve over time is valuable not only for natural language processing (NLP) but also for historical linguistics and digital philology. However, temporal modeling of Arabic remains limited due to uneven corpus coverage and linguistic complexity.

This paper introduces NassEra, a unified framework for estimating the date of composition of Arabic texts using neural architectures. The task is formulated as a regression problem predicting continuous year values instead of broad temporal categories. The framework combines Transformer-based encoders (such as AraBERT and CAMeLBERT) with generative models fine-tuned through Low-Rank Adaptation (LoRA). This dual design enables a balanced assessment of how contextualized encoders and generative reasoning capture historical linguistic signals.

Beyond performance evaluation, the study explores whether language models can learn and represent temporal patterns in Arabic writing, and how language-specific pretraining affects this ability. Through this investigation, NassEra connects computational modeling with philological inquiry, showing that large language models function as effective tools for historical text dating and the broader digital humanities.

2 Related Work

Research on temporal text modeling has evolved through several stages, moving from lexical heuristics to contextual neural representations. This section reviews prior approaches and situates our work within diachronic NLP, with emphasis on Arabic.

2.1 Rule-Based and Statistical Approaches

Early studies in temporal document dating relied on explicit lexical or stylistic markers. Dalli (2006) pioneered unsupervised dating using time-sensitive word frequencies, while Kanhabua and Núñez (2008) incorporated part-of-speech information and collocational entropy to detect period-specific expressions.

The emergence of supervised learning offered a more scalable alternative. García-Fernández et al. (2011) employed support vector machines and cosine similarity between temporal term vectors, achieving finer granularity in year estimation. Han et al. (2017) applied sliding interval classifiers to improve resolution over continuous time. Although these models captured broad diachronic tendencies, they depended heavily on engineered lexical features and struggled to generalize across languages or genres. Surveys such as Boldsen and Nørvåg (2021) confirmed that statistical systems plateau at coarse temporal resolution, motivating the shift to representation learning.

2.2 Neural and Embedding-Based Temporal Modeling

Neural architectures reduced reliance on manually designed features by learning distributed representations directly from text. Kulkarni et al. (2018) used neural models to infer composition periods through neologism frequency, and Liebeskind and Ordan (2020) showed that convolutional and recurrent networks could classify Hebrew texts by historical era.

Advances in diachronic word embeddings enabled studying semantic drift across time. Hamilton et al. (2016) proposed aligned vector spaces to model lexical and cultural change, while Rosenfeld and Erk (2018) extended this framework to unsupervised temporal embeddings. These contributions deepened understanding of semantic evolution, though few addressed document-level dating or continuous regression.

Temporally contextualized models such as TALM (Ren et al., 2023) introduced explicit time alignment across embedding layers, improving historical inference for English and Chinese corpora. However, such methods remain underexplored for morphologically rich or low-resource languages, including Arabic.

2.3 Historical Arabic Corpora and Temporal Resources

Compared to English and European languages, diachronic modeling of Arabic is still in its early stages. Foundational corpora such as *al-Shamela* (Zeroual et al., 2016) and *OpenITI* (Miller et al., 2020) have made thousands of classical texts digitally accessible, forming the basis for large-scale linguistic analysis. Studies such as Belinkov and Glass (2018) examined lexical and syntactic evolution, while Abujabal et al. (2018) introduced the *KERTAS* dataset for manuscript dating using visual and textual cues.

Most Arabic-focused work has prioritized corpus building or coarse chronological classification, typically distinguishing broad eras (e.g., early vs. late classical). Fine-grained, year-level regression-especially using neural architectures-has received little attention. Moreover, the morpho-orthographic variability of Arabic complicates temporal modeling, requiring specialized preprocessing and language-aware representations.

2.4 Large Language Models and Temporal Inference

Transformer architectures revolutionized diachronic NLP through contextualized sequence modeling. Encoder models such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and DeBERTa (He et al., 2020) have been applied to historical classification and semantic change detection, though most treat time as discrete rather than continuous.

Generative large language models (LLMs) offer a complementary avenue for temporal inference, capturing stylistic and syntactic variation across long spans of text. Yet their use for chronological prediction remains rare, particularly for Arabic. Parameter-efficient tuning strategies such as LoRA (Hu et al., 2021)-which greatly reduce training cost-have not been systematically evaluated in diachronic contexts.

2.5 Gaps Addressed by This Study

Despite progress in temporal NLP, several challenges remain:

- Most prior work focuses on English and a few high-resource languages, leaving Arabic largely unexplored.
- Existing studies favor discrete classification (by century or decade) over continuous year-level prediction.
- The potential of generative and parameter-efficient models for diachronic inference remains underexplored.

NassEra addresses these gaps by combining regression-based Transformer encoders with LoRA-tuned generative models on a chronologically balanced Arabic corpus. This dual approach enables efficient fine-tuning while capturing the linguistic and stylistic evolution of Arabic across more than a millennium.

3 Dataset Curation

The dataset used in this study builds upon the *OpenITI* corpus and incorporates additional digitized sources to extend chronological, thematic, and sectarian coverage. The objective was to construct a representative collection of historical Arabic texts spanning the full Islamicate intellectual tradition and suitable for diachronic linguistic modeling.

3.1 Sources

We integrated several major digital repositories, each contributing distinct textual and sectarian perspectives: Al-Shamila data set is under Creative Commons Attribution 4.0 International (CC BY 4.0) and The datasets used under the OpenITI dataset are under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

- **al-Maktaba al-Shāmila (Shamela):** the core open-access digital library of Arabic texts, widely used in computational research.
- **Shamela 2019 Additions (Sham19Y):** updated texts collected from new releases and digital archives.
- **al-Jāmi^{c} al-Kabīr (JK):** curated editions published by Markaz al-Turāth in Jordan.
- **al-Maktaba al-Shī^{c}iyya (Shia Online Library):** a Shī^{c}ī collection maintained by the Ayatollah Milānī Foundation.
- **Shamela Gift Collection (ShamAY):** the official Shamela release enriched with privately digitized manuscripts.
- **Zaydī and Ibāḍī Collections:** specialized corpora representing alternative theological traditions.
- **Supplementary Sources:** including the *Graeco–Arabic Studies Corpus* (GRAR), *Tafsīr* commentaries, the *Filāḥa Project* (agriculture texts), the *Hindāwī* library, and a multilingual *Bible Corpus* for comparison.

Additional works from smaller archives and private collections were incorporated to address temporal and sectarian imbalance. The final corpus covers texts from the early Abbasid period (~203 AH; 819 CE) to modern works beyond 1423 AH (2002 CE), representing nearly twelve centuries of Arabic written production.

3.2 Metadata and Structure

Each entry follows a standardized metadata schema adapted from OpenITI for consistent representation across sources. Core fields include: `date` (Hijri year or earliest attested reference), `author_lat` (normalized Latin transliteration), `book` (unique author–title code), `id` (source identifier), `tok_length` (tokenized word count), and `tags` describing subject area, period, and collection origin. Each record also stores the cleaned Arabic text (`clean_text`) and a short excerpt (`text_preview`) for inspection. All metadata fields were automatically validated and cross-referenced to prevent duplication and inconsistencies.

3.3 Text Cleaning and OCR Integration

To ensure uniformity across heterogeneous sources, all texts underwent a common preprocessing pipeline. Metadata markers (e.g., #META# sections), non-Arabic characters, and editorial headers were

removed; punctuation, diacritics, and spacing normalized; and redundant whitespace collapsed. Each document was stored as a UTF-8 string for efficient tokenization and modeling.

In addition to digital sources, one hundred printed works of historical and regional significance were scanned and digitized using optical character recognition (OCR). Scanning was performed at 300 dpi and processed with the Tesseract 5.4 engine fine-tuned for Arabic script. Roughly twelve percent of pages were manually corrected to reduce character-level error. These OCR additions expand the dataset's temporal and geographical coverage, especially for early and underrepresented genres.

3.4 Segmentation and Tokenization

After normalization, all texts were segmented into contiguous chunks of roughly 350 tokens to maintain semantic coherence and uniform model input lengths. Each segment inherited the metadata of its source document, including year and author identifiers. Tokenization and morphological segmentation were performed using `CamelTools`, yielding linguistically standardized input for both regression-based and generative modeling.

3.5 Corpus Overview

Table 1 summarizes the main characteristics of the curated corpus. The balanced temporal distribution enables consistent evaluation across historical periods.

Table 1: Overview of the curated Arabic corpus used for temporal modeling.

Category	Count	Coverage (Hijri / Gregorian)
Texts (unique works)	8,204	203–1423 AH / 819–2002 CE
Segments (chunks)	39,000	Balanced by half-century
Average segment length	347 tokens	–
Represented traditions	4 (Sunni, Shīᶜī, Ibāḍī, Zaydī)	–
Genres	Literature, theology, history, law, poetry	–

4 Methodology

This study examines the feasibility of predicting the year of composition of historical Arabic texts through neural language modeling. The framework combines two complementary paradigms: (1) regression-based fine-tuning of Arabic and multilingual Transformer encoders, and (2) parameter-efficient adaptation of large generative models using Low-Rank Adaptation (LoRA). The following subsections outline the modeling formulation, data processing, model architectures, and evaluation procedures.

4.1 Task Definition

The task is formulated as a supervised regression problem in which a model learns to predict the year of composition y_i for an Arabic text x_i. Formally, we seek a mapping $f_\theta(x_i) \rightarrow \hat{y}_i$, parameterized by θ, that minimizes the root mean squared error (RMSE) between predicted and true years:

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$$

Unlike categorical temporal classification (e.g., century- or decade-level), this continuous formulation preserves fine-grained chronological variation and captures gradual linguistic change.

4.2 Data Representation and Preprocessing

The corpus described in Section 3 was used to build a balanced temporal dataset for regression modeling. Each document was divided into coherent segments of roughly 350 tokens, producing 111,672 labeled samples. To balance temporal representation, half-centuries with fewer than 1,000 samples were removed, and up to 1,500 retained per interval, yielding 39,000 balanced records. The dataset spans 203–1423 AH (819–2002 CE), covering nearly twelve centuries of Arabic textual history.

Auxiliary temporal variables—`decade` and `half_century`—were derived from year annotations for stratified sampling and chronological evaluation. All texts were preprocessed as detailed previously, ensuring

normalized orthography and consistent tokenization.

4.3 Input Encoding

Each text segment was tokenized using the model's native tokenizer. Encoder-based models (e.g., AraBERT, CAMeLBERT) used a maximum sequence length of 512 tokens, while generative architectures (e.g., LLaMA-2, Qwen2.5, ALLaM) accepted up to 2,048 tokens.

For regression models, the text served as input and the composition year as a scalar target. For generative models, year prediction was reformulated as a numeric generation task via Arabic instruction prompts such as:

أُعطيت نصًا عربيًا. تنبّأ بسنة التأليف (هجري).
اكتب الرقم فقط دون أي نص آخر:

This approach aligns with instruction-tuned paradigms, allowing generative LLMs to infer temporal information directly from linguistic cues.

4.4 Model Architectures

Two model families were evaluated.

4.4.1 Encoder-based Regression Models

Pretrained Transformer encoders were fine-tuned using a regression head to output continuous year predictions:

- `distilbert-base-multilingual-cased`
- `bert-base-multilingual-uncased`
- `CAMeL-Lab/bert-base-arabic-camelbert-da`
- `aubmindlab/bert-large-arabertv02`

Fine-tuning employed the Hugging Face `Transformers` framework with `problem_type='regression'` to yield scalar outputs.

Generative LoRA Models. Large Arabic-capable generative models were adapted using parameter-efficient fine-tuning (PEFT) with Low-Rank Adaptation (LoRA), updating only a small fraction of parameters:

- `meta-llama/LLaMA-2-7B-Chat-HF`
- `Qwen/Qwen2.5-7B-Instruct`
- `humain-ai/ALLaM-7B-Instruct`

LoRA applied a low-rank decomposition with $r \in \{4, 8\}$, scaling factor $\alpha \in \{8, 16\}$, and dropout $p = 0.05$, targeting the attention modules (`q_proj`, `k_proj`, `v_proj`, `o_proj`). This configuration reduced trainable parameters to approximately 0.12–0.15% of total model capacity, enabling efficient fine-tuning on standard GPUs while maintaining representational stability.

4.5 Training Configuration

Encoder models were trained with the AdamW optimizer (learning rate 1×10^{-5}, batch size 32) for up to 30 epochs with early stopping (typically converging around 15 epochs) based on validation loss. Generative LoRA models were fine-tuned for four epochs with an effective batch size of 8 (via gradient accumulation), learning rate 1×10^{-4}, and mixed-precision FP16 computation. Evaluation occurred every 20 steps, tracking both loss and numeric prediction accuracy. All experiments were executed on NVIDIA CUDA GPUs with automatic device mapping.

4.6 Evaluation Metrics

Model performance was assessed using both regression and generation-based metrics to reflect the two modeling paradigms.

For encoder-based regressors:

- **Root Mean Squared Error (RMSE)** - overall deviation from true years.
- **Mean Absolute Error (MAE)** - average prediction discrepancy.
- R^2 **Score** - proportion of variance explained.

For generative LoRA models:

- **MAE** and **RMSE** - computed from decoded outputs.
- **Median Absolute Error** - robust central tendency measure.
- **Exact / Off-by-One Accuracy** - decade-level tolerance.
- **Pearson Correlation** (r) - correlation with true years.

This dual evaluation scheme ensures fair comparison: regression metrics assess numeric precision, while correlation and accuracy capture qualitative generation behavior. All results are computed on a held-out validation split (10–20% of the corpus) with no text overlap between training and testing.

4.7 Reproducibility

All experiments were executed with fixed random seeds and deterministic PyTorch behavior. Preprocessing scripts, configurations, and logs were preserved, and all experiments used open-source libraries (`transformers`, `peft`, and `datasets`). This design follows best practices for transparency and replicability in computational linguistics.

5 Results

This section summarizes the main experimental outcomes, including model performance and temporal error patterns. The findings are presented objectively, with interpretation reserved for Section 6.

5.1 Performance Overview

Table 2 reports the results for all encoder-based regression models at their final training epochs. Evaluation metrics include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Lower RMSE and MAE values, and higher R^2 scores, indicate better temporal prediction performance.

Table 2: Performance of encoder-based regression models on Hijri year prediction (203–1423 AH; 819–2002 CE). Lower RMSE/MAE and higher R^2 indicate better performance.

Model	RMSE	MAE	R^2
DistilBERT (15 epochs)	78.49	46.90	0.5285
CAMeL-BERT (15 epochs)	207.57	118.22	0.7609
AraBERT v2 (Large)	**66.06**	**36.93**	0.6561
Multilingual BERT (15 epochs)	78.66	46.01	0.5265

Among encoder models, **AraBERT v2 (Large)** achieved the best overall performance, showing that Arabic-specific pretraining improves the capture of historical and stylistic variation. CAMeL-BERT obtained the highest R^2, suggesting broader variance explanation, though at the cost of considerably higher residual error.

Table 3 presents the outcomes for LoRA-tuned generative models, which reformulate year prediction as a numeric generation task. Models were evaluated using regression metrics and generation accuracy.

The **ALLaM-7B-Instruct** model achieved the highest overall precision (MAE = 63.6, Pearson $r = 0.94$), demonstrating fine-grained chronological accuracy within several decades while updating only 0.12% of parameters. The updated **Qwen2.5-7B-Instruct** model showed strong intermediate performance (MAE = 79.4, RMSE = 148.2, $r = 0.88$), confirming its efficiency as a multilingual alternative with lower computational cost.

5.2 Temporal Error Patterns

Table 4 reports average prediction errors by century. Earlier centuries (3rd–7th AH; 9th–13th CE) show higher errors due to diverse orthography and morphology, while later centuries (8th–14th AH; 14th–20th CE) exhibit improved accuracy, reflecting greater linguistic standardization and richer corpus coverage.

Overall, results show that model accuracy improves in later centuries (post–8th AH), confirming that linguistic standardization and corpus density support better temporal prediction.

5.3 Summary of Findings

Key findings include:

- **AraBERT v2** delivered the best encoder-based performance, confirming the value of Arabic-specific pretraining.
- **ALLaM-7B-LoRA** achieved the overall highest precision ($r = 0.94$, RMSE = 128.9).
- **Qwen2.5-7B-LoRA** provided competitive mid-level results ($r = 0.88$, RMSE = 148.2) with lower training cost.
- All models showed stable convergence and consistent training behavior.

Table 3: Performance of LoRA-tuned generative models on Hijri year prediction (203–1423 AH). Best values per column are bolded.

Model	MAE	RMSE	Median AE	Exact (%)	r
LLaMA-2-7B-Chat + LoRA	322.7	443.8	218.0	4.7	0.466
Qwen2.5-7B-Instruct + LoRA	79.4	148.2	62.1	20.0	0.880
ALLaM-7B-Instruct + LoRA	**63.6**	**128.9**	**7.0**	**32.5**	**0.942**

Table 4: Average absolute error (MAE) by century for representative models. Lower is better.

Century (ah)	CAMeL-BERT	AraBERT	ALLaM-LoRA
200–700	412.5	398.2	512.0
800–1100	275.3	241.7	165.4
1200–1400	131.4	94.6	**63.6**

- Temporal accuracy increased toward the modern period (after 800 AH / 1400 CE), demonstrating that linguistic regularization enhances model reliability.

These quantitative results form the basis for the comparative discussion in Section 6.

6 Comparative Analysis and Discussion

This section interprets the experimental findings, focusing on model behavior, linguistic implications, and broader insights.

6.1 Model Behavior and Language Specialization

Arabic-specific models, especially **AraBERT** and **CAMeL-BERT**, consistently outperformed multilingual counterparts due to pretraining on extensive Arabic corpora capturing morphological and orthographic variation across historical registers. This confirms that language-specific pretraining is essential for temporal modeling in Arabic, a language marked by diachronic and stylistic diversity. Among evaluated models, **ALLaM-7B** fine-tuned with LoRA achieved the highest precision, showing that instruction-tuned generative architectures effectively learn temporal cues from limited data while requiring minimal parameter updates. The **Qwen2.5-7B** model also showed strong intermediate performance, indicating that multilingual instruction-tuned LLMs generalize well to Arabic diachronic data even without exclusive Arabic pretraining. These findings suggest that generative LLMs internalize diachronic linguistic knowledge without explicit temporal supervision.

6.2 Effect of Model Scale and Adaptation

Model scale and adaptation strategy influenced performance differently. Encoder-based regressors required full fine-tuning of all parameters for stability, whereas LoRA-based generative models achieved comparable or better accuracy while updating under 0.2% of parameters. This highlights the efficiency of parameter-efficient adaptation for historical and low-resource tasks. Scaling model size alone proved insufficient; targeted fine-tuning methods like LoRA produced more reliable temporal representations even with moderate data.

6.3 Linguistic and Diachronic Insights

Prediction errors revealed interpretable linguistic trends. Texts from early centuries (3rd–7th AH) showed higher errors due to orthographic inconsistency, morphological variation, and regional diversity. In later centuries (8th–14th AH), the standardization of spelling, syntax, and lexicon-driven by education and print culture-correlated with higher accuracy. These results indicate that models capture linguistic stabilization over time, providing quantitative evidence for patterns noted in Arabic philology. Neural models thus complement traditional analysis by quantifying diachronic variation at scale.

6.4 Generative vs. Regression Paradigms

Encoder regressors produced stable and interpretable outputs suitable for controlled evaluation. Generative models, by contrast, captured broader stylistic and discourse-level cues that signaled temporal variation indirectly. While their numeric predictions sometimes deviated, they often preserved correct chronological order, implying an internalized sense of linguistic temporality.

Combining regression precision with generative contextual sensitivity may further enhance temporal modeling.

6.5 Error Patterns and Biases

A mild regression-to-the-mean bias appeared across models, with predictions clustering near median years. High-error samples often came from edited or composite works combining features from multiple eras. This highlights the need for detailed textual provenance and metadata to improve temporal inference. Future work may include stylistic segmentation and contextual embeddings to better capture intra-textual temporal variation.

6.6 Broader Implications

Transformer-based models infer chronological and linguistic patterns from Arabic texts without manual feature design. By integrating regression-based encoders and parameter-efficient generative models, NassEra establishes a scalable framework for diachronic modeling across twelve centuries (203–1423 AH). This approach enables applications in manuscript dating, linguistic periodization, and large-scale exploration of historical corpora. Future extensions may incorporate regional or genre metadata, multimodal sources, and interpretability tools to identify features most predictive of temporal change.

7 Conclusion

This study demonstrates that predicting the year of composition of historical Arabic texts is both feasible and linguistically meaningful using modern Transformer and generative models. Arabic-specific encoders such as **CAMeL-BERT** and **AraBERT** performed strongly, underscoring the importance of language-focused pretraining for diachronic modeling. Parameter-efficient fine-tuning of large generative models, particularly **ALLaM-7B** and **Qwen2.5-7B** with LoRA, improved temporal accuracy while updating under 0.2% of parameters.

Model performance reflected clear sensitivity to historical linguistic variation across the corpus (203–1423 AH; 819–2002 CE). Accuracy was highest in later centuries, where textual conventions and orthography became more standardized. These findings show that neural models internalize diachronic patterns in Arabic, enabling large-scale quantitative analysis of linguistic change across twelve centuries.

Beyond temporal prediction, NassEra provides a practical framework for manuscript dating and historical philology. Future work will extend this approach to multilingual corpora, integrate regional and genre metadata, and explore authorship attribution as a complementary task for chronological inference.

Overall, the results confirm that Arabic-oriented and multilingual instruction-tuned language models serve as effective tools for historical research, offering new insights into the evolution of Arabic language and intellectual tradition from the early Abbasid period to the modern era.

Limitations

Despite its broad scope, the study faces several limitations. The corpus, while extensive, remains unevenly distributed across centuries and genres, with early periods (3rd–7th AH; 9th–13th CE) underrepresented relative to later ones. All experiments were conducted on normalized digital texts, which enhance consistency but reduce orthographic, dialectal, and manuscript variation. Evaluation focused on numeric year prediction, leaving stylistic and regional dimensions unexplored. Large generative models also pose practical and ethical challenges, including computational cost, limited reproducibility, and data imbalance.

Future research should expand corpus coverage for earlier centuries, adopt richer evaluation metrics (e.g., century-level classification or stylistic clustering), and explore interpretable fine-tuning methods. Even with these constraints, the present work establishes a solid foundation for quantitative Arabic historical linguistics and demonstrates the promise of neural models—both Arabic-specific and multilingual—in capturing the temporal and stylistic evolution of Arabic across more than a millennium.

References

Abdalghani Abujabal, Hend Al-Khalifa, and Khaled Shaalan. 2018. Kertas: A dataset for arabic manuscript dating using textual and visual features. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*.

Yonatan Belinkov and James Glass. 2018. Arabic language modeling and analysis of historical variation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*.

Sigurd Boldsen and Kjetil Nørvåg. 2021. A survey of text dating methods: From rule-based heuristics to deep learning. *ACM Computing Surveys*, 54(3):1–36.

Angelo Dalli. 2006. Temporal text mining: the application of temporal text mining in the analysis of news. In *Proceedings of the 2nd International Conference on Web Information Systems and Technologies*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Alberto García-Fernández, Alberto García Núñez, Víctor Fresno, and Rafael Martínez. 2011. Publication date estimation for web pages using temporal language models. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1211–1212.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121.

Jialong Han, Yizhou Sun, and Xifeng Yan. 2017. Dating documents with sliding intervals. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1351–1360.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Nattiya Kanhabua and Alberto García Núñez. 2008. Improving temporal language models for determining the time of non-timestamped documents. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1053.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2018. Temporal text classification and analysis of neologisms in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 145–155.

Chaim Liebeskind and Noam Ordan. 2020. Diachronic hebrew text classification using neural models. *Digital Scholarship in the Humanities*, 35(4):911–927.

Sarah Bowen Miller, Maxim Romanov, and Yuri Takhteyev. 2020. Openiti: A machine-readable corpus of classical arabic texts. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Xiang Ren, Bowen Zhang, Xin Lu, Qi Li, and Fei Wu. 2023. Talm: Time-aware language models for diachronic nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Adam Rosenfeld and Katrin Erk. 2018. Deep diachronic embeddings for historical linguistics. In *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics, pages 146–157.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Imad Zeroual, Abdelhak Lakhouaja, and Laurent Romary. 2016. Building and analyzing a large-scale arabic digital library: The shamela corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

A Appendix

A.1 Dataset and Repository Access

To promote transparency and reproducibility, we release anonymized datasets and source code associated with this study:

- **Dataset and Code Repository (Anonymized)**: Available via `anonymous.4open.science`[1]. The repository includes:
  - The anonymized dataset comprising more than 39,000 Arabic text segments annotated with Hijri and Gregorian years (203–1423 AH; 819–2002 CE).
  - Structured metadata (source repository, token length, authorship century, and genre).
  - All preprocessing scripts, model fine-tuning notebooks, evaluation pipelines, and experiment logs.

All resources comply with double-blind review requirements and will be made publicly available upon paper acceptance.

A.2 Extended Experimental Results

Table 5 reports complete regression and generation outcomes for all models, including training epochs, performance metrics, and qualitative notes. Arabic-specific encoders (AraBERT, CAMeLBERT) consistently show lower error rates than multilingual baselines, while generative LoRA-tuned models outperform both in overall temporal precision.

A.3 Training and Evaluation Configuration

All experiments were conducted using PyTorch with the Hugging Face `Transformers` and `PEFT` libraries, executed on CUDA-enabled GPUs. Table 6 summarizes the primary hyperparameters and hardware environment.

A.4 Per-Century Error Analysis

To assess temporal variation, Table 7 presents mean absolute error (MAE) values by Hijri century. Early centuries (3rd–7th AH; 9th–13th CE) exhibit higher prediction errors due to greater morphological and orthographic variability, while later centuries (8th–14th AH; 14th–20th CE) show reduced error as written Arabic becomes more standardized.

A.5 Reproducibility Statement

All datasets, configurations, and model training pipelines are reproducible and documented in the anonymized repository. Each experiment includes version-controlled scripts for preprocessing, fine-tuning, evaluation, and analysis, alongside detailed logs and checkpoint management. Model weights and generated predictions will be publicly released upon acceptance to support continued research in diachronic Arabic NLP and historical language modeling.

[1] https://anonymous.4open.science/r/NassEra-FFFD/README.md

Table 5: Comprehensive regression and generation results for all evaluated models. Bold values denote best results per metric.

Model	Type	Epochs	RMSE	MAE	R^2 / r	Params Updated	Notes
DistilBERT	Encoder	15	78.5	46.9	0.5285	100%	Baseline multilingual model
CAMeL-BERT	Encoder	15	207.6	118.2	0.7609	100%	High variance explained (R^2)
AraBERT v2 (Large)	Encoder	15	**66.1**	**36.9**	0.6561	100%	Best encoder overall
mBERT (uncased)	Encoder	15	78.7	46.0	0.5265	100%	Multilingual reference baseline
ALLaM-7B-Instruct (LoRA)	Generative	4	**128.9**	**63.6**	**0.942**	**0.12%**	Strongest performance overall
Qwen2.5-7B-Instruct (LoRA)	Generative	4	148.2	79.4	0.880	0.15%	Stable multilingual performer
LLaMA-2-7B-Chat (LoRA)	Generative	4	443.8	322.7	0.466	0.12%	Learned valid numeric outputs

Table 6: Training configuration summary for encoder-based and LoRA-adapted generative models.

Parameter	Value or Description
Framework	PyTorch + Hugging Face Transformers / PEFT
Optimizer	AdamW
Learning Rate	1e-5 (Encoders), 1e-4 (LLMs)
Batch Size	32 (Encoders), 4 (LLMs)
Loss Function	Mean Squared Error (MSE) for regression objective
Max Sequence Length	512 (Encoders), 2048 (LLMs)
Precision	FP16 (mixed precision)
LoRA Configuration	Rank r: 8 (ALLaM), 4 (Qwen); Scaling α: 8–16; Dropout: 0.05
Hardware	NVIDIA RTX GPU (CUDA-enabled, 24–48 GB memory)
Average Training Time	2–3 hours per LoRA model

Table 7: Average absolute error (MAE) by Hijri century for representative models. Lower values indicate better temporal prediction accuracy.

Century (AH)	CAMeL-BERT	AraBERT v2	ALLaM-7B (LoRA)
200–700	412.5	**398.2**	512.0
800–1100	275.3	241.7	**165.4**
1200–1400	131.4	94.6	**63.6**