SeshatLM: Leveraging Synthetic Data for Hieroglyphic Reconstruction

Anonymous ACL submission

Abstract

Reconstructing damaged Egyptian manuscripts remains a critical challenge in archaeology and cultural heritage preservation. We present *SeshatLM*, a hieroglyphic reconstruction framework based on an mmBERT model fine-tuned with a custom masked language modeling objective that simulates realistic damage in hieroglyphic text. To improve restoration performance, we leverage synthetic training data generated through machine translation, achieving 58.22% accuracy in missing glyph prediction. We additionally introduce a large-scale synthetic hieroglyphic corpus generated by translating German Bible verses and general-domain German text into hieroglyphics using a fine-tuned TranslateGemma model. The dataset contains 150,000 synthetic samples and will be released to support future research in hieroglyphic language modeling and restoration. Our results highlight the potential of language-model-based systems for reconstructing fragmentary Egyptian inscriptions.

1 Introduction

Egyptology offers significant opportunities for computational analysis, particularly in the study of hieroglyphics, one of the earliest writing systems and a primary source of insight into ancient Egyptian language, culture, and history. However, many Egyptian artifacts are fragmentary or heavily damaged due to erosion, physical deterioration, or deliberate destruction, resulting in missing glyphs and incomplete inscriptions that hinder historical interpretation and translation.

To address this challenge, we introduce *SeshatLM*, a transformer-based framework for reconstructing missing hieroglyphic text. Our approach fine-tunes an mmBERT model using a custom masked language modeling objective that simulates realistic patterns of contiguous damage commonly observed in Egyptian artifacts, as described in Section 6.2. SeshatLM achieves a reconstruction accuracy of 58.22% in restoring 15% masked hieroglyphic sequences, demonstrating the effectiveness of language-model-based restoration for fragmentary inscriptions.

In addition, we introduce a novel hieroglyphic dataset extending the data used in HieroTransformer (De Cao et al., 2024), consisting of 178,164 hieroglyphic samples. The dataset combines both human-curated and synthetically generated data. To generate synthetic hieroglyphic text, we leverage the Thesaurus Linguae Aegyptiae (TLA) dataset (Thes. Ling. Aegyptiae, 2024), which provides aligned hieroglyphic sentences and German translations. Using this alignment, we train a translation model that maps German text into hieroglyphics. Our synthetic corpus is generated from both in-domain German Bible verses, selected for their thematic similarity to the religious content commonly found in hieroglyphic inscriptions (BibleNLP, 2023), and out-of-domain general German sentences (jonas-is coding). The synthesis process is described in Section 5. The resulting dataset contains 150,000 synthetic hieroglyphic samples, which we release to support future research in computational Egyptology and hieroglyphic language modeling.

Our contributions are as follows: *(a)* We introduce *SeshatLM*, a transformer-based system for hieroglyphic text reconstruction using a custom masking strategy tailored to realistic artifact damage; *(b)* We develop a model for German into Ancient Egyptian hieroglyphic translation to generate synthetic training data; *(c)* We demonstrate that synthetic data augmentation substantially improves hieroglyphic restoration performance under limited-resource conditions; *(d)* We release a large-scale hieroglyphic dataset containing both human-curated and synthetic samples generated from in-domain and out-of-domain German corpora.

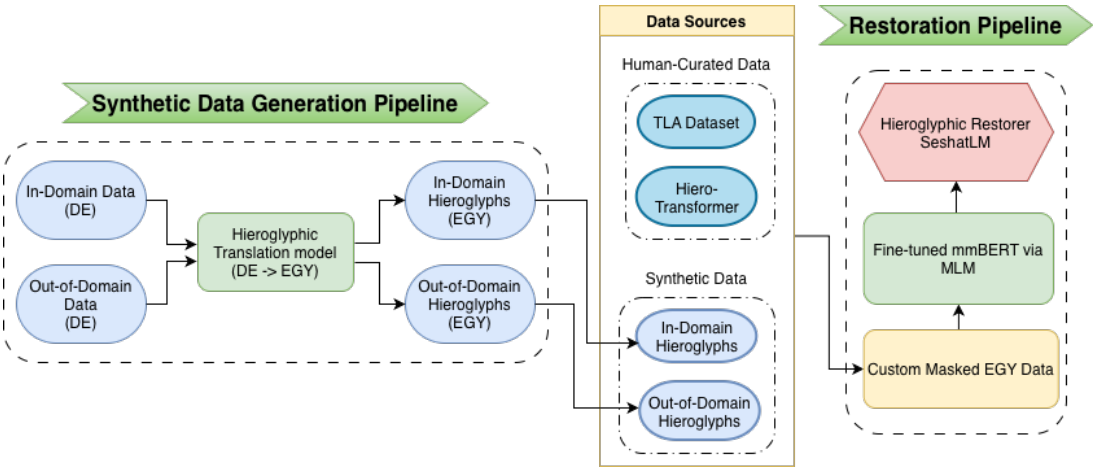

Figure 1: Overview of our reconstruction framework. SeshatLM is a language model trained with a custom Masked Language Modeling objective to reconstruct corrupted or missing token subsequences from the surrounding textual context. EGY denotes hieroglyphic text, while DE denotes standard German. The in-domain data consists of German Bible verses selected for their thematic similarity (religious texts) to the hieroglyphic texts in the TLA dataset, whereas the out-of-domain data consists of general-domain German sentences.

Setup/Data	5% masked		15% masked		30% masked	
	Top 1 Acc.	Top 5 Acc	Top 1 Acc.	Top 5 Acc	Top 1 Acc.	Top 5 Acc
TLA + HT + SYNT	66.66%	81.03%	58.22%	74.77%	47.47%	66.98%

Table 1: Language modeling accuracies on the TLA test subset using our custom masking scheme. Top-1 Acc. measures accuracy based on the model's highest-confidence prediction, while Top-5 Acc. considers whether the correct token appears within the model's top five predictions. The strong Top-5 Acc. scores support the use of SeshatLM and similar neural text restoration models as tools to assist experts in the restoration process, rather than as isolated black-box systems. TLA refers to the TLA dataset, HT refers to the HieroTransformer dataset described in Section 3.2, and SYNT refers to the overall synthetic data described in Section 3.3.

2 Findings

Table 1 presents performance on the test subset sampled from the TLA dataset. With 15% of contiguous text spans masked or "corrupted", the fine-tuned model trained only on the human-curated TLA and HieroTransformer datasets achieves an accuracy of 52.16%. Incorporating synthetic in-domain religious text translated from Bible verses improves performance to 54.69%. The best-performing model is trained on the most comprehensive dataset, consisting of TLA, HieroTransformer, synthetic translated Bible text, and synthetic general-domain German sentences, achieving an accuracy of 58.22%.

Details on the datasets used in training and testing are in Sections 3, 4, and 5. The modeling architectures are detailed in Section 6. Experimentation details are in Section 7.

3 Datasets

3.1 Thesaurus Linguae Aegyptiae

The dataset used in this study originates from the Thesaurus Linguae Aegyptiae (TLA) project (Thes. Ling. Aegyptiae, 2024) and was made accessible through the Hugging Face platform. This dataset contains 12,773 hieroglyphic text entries, providing a robust foundation for our research. Each entry includes metadata that enhances the understanding and analysis of the hieroglyphic content. The accompanying metadata includes Transliteration into Latin characters, Lemmatization, Universal Part-of-Speech (UPOS) tags, Glossing, and a German translation. A date estimate is provided via the "DateNotBefore" and "DateNotAfter" fields, which specify chronological boundaries for the text, aiding in historical contextualization.

3.2 HieroTransformer dataset

The HieroTransformer dataset builds upon the foundational work of the TLA project and introduces

a significant augmentation of new hieroglyphic text and its transliterations, creating a more extensive corpus (De Cao et al., 2024). We took 14,561 unique data points from the HieroTransformer dataset, ensuring no overlap with those in the TLA project. In addition, the dataset has been processed to include symbols that indicate destruction such as '*' and '?' for missing words or glyphs.

3.3 Synthetic Hieroglyphic Data

In addition to combining existing human-curated datasets from TLA and HieroTransformer, we generated synthetic hieroglyphic data to further augment the dataset.

3.3.1 In-Domain Synthetic Data

To generate in-domain synthetic data, we use the `biblenlp-corpus`[1] dataset from Hugging Face, selecting the German subset due to its thematic similarity to the religious and ceremonial content commonly found in ancient Egyptian inscriptions. Given that the hieroglyphic text in the TLA dataset primarily revolves around themes like life, the divine, and religion, we sourced additional text with similar themes. Thus, we use Bible verses as a semantically related source domain to preserve domain-specific linguistic structure during synthetic data generation.

We translate the German text into hieroglyphics using our translation system described in Section 5. The resulting corpus contains 31,096 synthetic hieroglyphic samples. These samples are used to augment the restoration training data and improve the model's ability to reconstruct hieroglyphic sequences under limited-resource conditions.

[1] huggingface.co/datasets/bible-nlp/biblenlp-corpus

Dataset	Total	Cleaned	Vocab.
TLA	-	12,773	1320
HieroTransformer	18,844	14,561	1458
Synthetic In-Domain	-	31,096	911
Synthetic Out-of-Domain	-	119,734	1071
Combined	-	178,164	-

Table 2: The datasets used in training SeshatLM. TLA is the data from Thes. Ling. Aegyptiae, 2024. HieroTransformer is a subset of the data used in De Cao et al. (2024). Vocab. refers to the count of distinct vocabulary units seen on the hieroglyphic side.

3.3.2 Out-of-Domain Synthetic Data

We additionally generate out-of-domain synthetic data using the `german-sentences`[2] dataset from Hugging Face, which further improves missing hieroglyph prediction accuracy. Unlike the Bible corpus, this dataset consists of general-domain German sentences spanning a broader range of vocabulary, sentence structures, and topics.

Using the same translation framework, we translate the German sentences into hieroglyphics, producing 119,734 synthetic samples. Incorporating both in-domain and out-of-domain synthetic corpora enables the restoration model to leverage domain-specific semantic patterns while improving generalization across broader linguistic contexts.

4 Data Pre-processing

We performed comprehensive data preprocessing steps to standardize and integrate data from multiple sources. These preprocessing tasks were critical for ensuring consistency and compatibility between datasets.

(a) Standardizing Glyph Representations Any glyph in the TLA dataset bounded by <g> and </g> tags is mapped into a single token corresponding to its Gardiner code, a standardized notation for Egyptian hieroglyphs.

To merge the Thesaurus Linguae Aegyptiae dataset with the HieroTransformer dataset, we converted all Gardiner code representations in the HieroTransformer dataset into their respective Unicode codepoints. In all datasets, including HieroTransformer's, each Gardiner code is set up to be treated as one token by the tokenizer. This transformation allowed for uniform representation across datasets. To achieve this, for the subset of Gardiner codes with no assigned Unicode codepoints in the latest standards, we applied a 1-to-1 mapping with a surrogate codepoint that did not appear elsewhere in the dataset. This conversion was facilitated using the Gardiner code table available on Wikipedia (Wikipedia contributors, 2024), based on multiple works published by Alan H. Gardiner such as Gardiner (1957).

The conversion process ensures that the datasets are aligned, as the HieroTransformer dataset exclusively used Gardiner codes instead of hieroglyphic Unicode codepoints.

[2] huggingface.co/datasets/jonas-is-coding/german-sentences

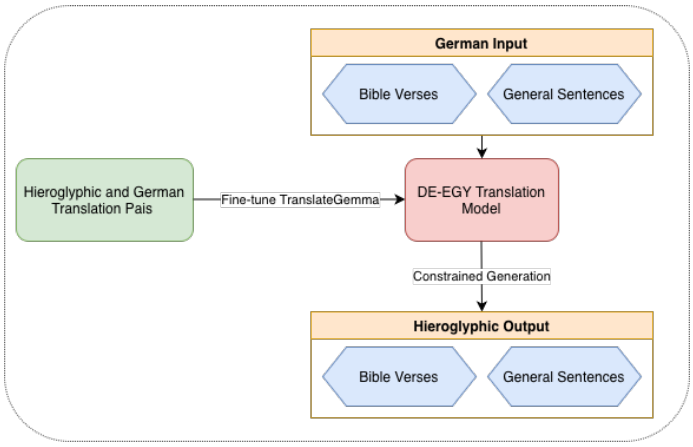

Figure 2: Overview of the synthetic data generation pipeline. DE denotes German text and EGY denotes hieroglyphic text. The DE-EGY translation model is first trained on aligned German–hieroglyphic pairs, after which existing German datasets are translated into hieroglyphics using constrained decoding.

(b) Filtering Duplicates After obtaining a standardized representation of hieroglyphic data points in the form of Gardiner codes, we applied a duplicate filtering process to ensure an unbiased dataset. By removing duplicate entries, we eliminated redundant data points that could skew the learning process or introduce unintended biases when training the model. This step was essential for maintaining the integrity of the dataset and ensuring that the model was exposed to a diverse and representative set of hieroglyphic data points. Among the TLA and HieroTransformer datasets, we removed 4283 duplicate examples.

(c) Merging the Datasets By transforming all hieroglyphic text into a consistent format, we successfully merge the TLA and HieroTransformer datasets, resulting in a unified corpus of hieroglyphic data consisting of 27,334 data points.

(d) Word Separators To further accommodate the character-based tokenization scheme for the Language Modeling system, all word separators are removed. This aligns with the expectation that text corruption would not respect word boundaries.

5 Synthetic Data Generation Pipeline

To augment the training data and improve the prediction accuracy and robustness of the hieroglyphic restoration model, we develop a translation system that takes German text as input and outputs the corresponding Ancient Egyptian in hieroglyphs. We then use this system to generate synthetic hieroglyphic data for restoration training. Figure 2 summarizes the overall translation pipeline.

5.1 Hieroglyphic Translation Model

TranslateGemma Fine-Tuning We fine-tuned `TranslateGemma-4B` (Google Translate Research Team et al., 2026), a 4-billion-parameter decoder-only vision-language model built on Gemma 3 and pre-trained by Google for multilingual translation across 55 languages, to translate German text into hieroglyphics. Fine-tuning was performed on the TLA corpus, partitioned into 80% training, 10% validation, and 10% test sets.

Hieroglyph Tokenization Some hieroglyphs in the TLA corpus are encoded as JSesh Egyptological codes delimited by XML-style tags (e.g., `<g>M12B</g>`). To address tokenization issues arising from these alphanumeric codes, we constructed a deterministic bijective mapping from each unique JSesh token (e.g., `A1`, `D21`, `Aa1`) to a distinct CJK Unicode character. This ensures that each hieroglyphic glyph is represented as a single token rather than being fragmented into multiple subword pieces. Hieroglyph tokens were defined at the word level, treating each whitespace-delimited unit in the TLA source as an atomic token. The resulting vocabulary was then added to the model tokenizer, and the embedding matrix was resized accordingly.

Prompt Construction Input formatting required a non-trivial workaround: `TranslateGemma`'s built-in `apply_chat_template` interface enforces ISO 639-1 language codes for both source and target languages and raises an error when an unsupported language code is provided. Since Egyptian hieroglyphs do not have a registered ISO 639-1

code, the official template directly was not used. Instead, we manually constructed the Gemma 3 instruction-tuning prompt format using the model's raw special tokens (<bos>, <start_of_turn>, and <end_of_turn>), prepending a fixed system instruction directing the model to output only the hieroglyphic sequence given a German source sentence. During training, cross-entropy loss was computed exclusively over target hieroglyphic tokens, with all prompt positions masked to -100.

Training Configuration We trained the model for up to 50 epochs with early stopping (patience $= 5$) monitored on validation loss. Training used a per-device batch size of 8 with gradient accumulation over 2 steps (effective batch size of 16), a peak learning rate of 2×10^{-5} under a cosine decay schedule with 5% linear warmup, weight decay of 0.01, `bf16` mixed precision, and gradient checkpointing to accommodate the 4B-parameter model on a single NVIDIA A6000 GPU.

Evaluation The checkpoint achieving the lowest validation loss was selected for final evaluation. Evaluation on the held-out 10% test split was performed using full autoregressive generation with greedy decoding, generating up to 256 new tokens per source sentence. Translation quality was measured using chrF++, achieving a score of 36.86 on the test set.

5.2 Data Creation

Having established a fine-tuned German-to-hieroglyphs translation model, we leverage it to construct a large-scale synthetic parallel corpus for training the downstream restoration model.

Vocabulary-Constrained Decoding To prevent hallucination and ensure that every generated token belongs to the valid hieroglyph vocabulary, we employ vocabulary-constrained decoding at inference time. We pre-compute a validity mask over the full model vocabulary by verifying that each token decodes exclusively to characters within either the Egyptian Hieroglyphs Unicode block or the CJK surrogate range used during training to represent JSesh codes. At each decoding step, any token falling outside this valid set is assigned a logit of $-\infty$, hard-constraining the search space to well-formed hieroglyphic output. Termination tokens are unconditionally preserved to ensure generation can conclude normally. As a further safeguard, any output sequence that is empty or does not contain at least one valid hieroglyph character is discarded in a post-generation filtering step.

Corpus Construction Using the constrained decoding pipeline, we translated German Bible verses and general-domain German sentences into hieroglyphics. Generation was performed using greedy decoding with a maximum output length of 256 tokens. The resulting corpus comprised 150,000 synthetic samples, which were used to augment restoration training, improving both robustness and reconstruction accuracy.

6 Restoration Pipeline

6.1 Tokenization Scheme

Our model uses character-level tokenization, where the input consists of a single continuous string of hieroglyphic text without spaces. The model's vocabulary is composed of all unique glyphs present in the training dataset. Figure 3 gives a detailed example that illustrates how an input data point is tokenized.

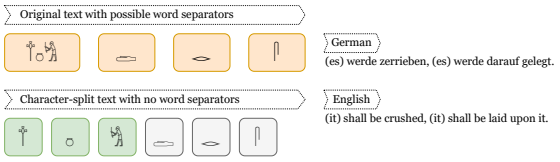

Figure 3: Example of a tokenized data point.

6.2 Custom Masking Scheme

Rather than adopting the standard masking strategy used in the original BERT pre-training objective, we introduce a span-based masked language modeling scheme that masks multiple contiguous token spans within each sequence. In standard BERT-style Masked Language Modeling (MLM), tokens are masked independently with a fixed probability (typically 15%), resulting in masked positions that are uniformly distributed throughout the sequence without preserving structural relationships between neighboring tokens.

Our approach instead masks contiguous spans of tokens, more closely reflecting the localized, spatial damage patterns commonly observed in fragmentary hieroglyphic inscriptions. Span lengths are sampled from a Poisson distribution with mean span length $\lambda = 2$ by default, inspired by the span-corruption objectives of SpanBERT (Joshi et al., 2020) and T5 (Raffel et al., 2023). For each sequence, eligible non-special token positions are

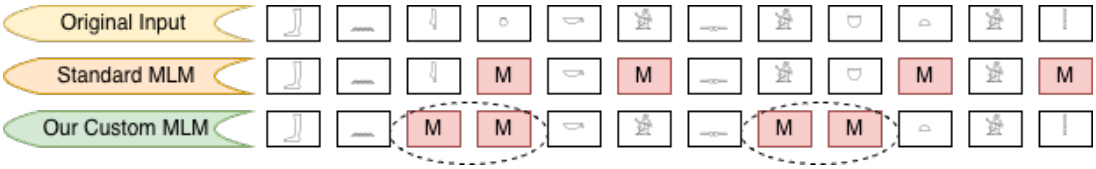

Figure 4: An example comparison between the standard MLM scheme and our proposed custom multi-span masking scheme used for hieroglyphic restoration training. M denotes a masked token. Multiple contiguous token spans are masked to simulate realistic damage patterns commonly observed in fragmentary artifacts.

first identified. Random span starting positions are then iteratively sampled, and contiguous spans are masked until 15% of the sequence tokens are corrupted. Once a span is selected, its token positions are removed from the candidate pool to prevent overlap and ensure unique masked tokens. Figure 4 shows a comparison between standard MLM and our proposed masking technique.

This masking formulation encourages the model to reconstruct coherent adjacent token segments rather than scattered tokens, producing a substantially more challenging but meaningful restoration objective. By modeling contiguous gaps instead of independently masked positions, the proposed scheme better captures real-world hieroglyphic damage patterns and improves the model's ability to recover missing text from surrounding context.

6.3 Model Architecture

Our restoration model is built upon the mmBERT-base architecture (Marone et al., 2025), a massively multilingual encoder trained on over 3 trillion tokens spanning 1,800+ languages. mmBERT-base is built on the ModernBERT architecture (Warner et al., 2024), incorporating several improvements over classical BERT-style encoders (Devlin et al., 2019) including alternating local and global attention, Flash Attention 2 (Dao, 2023), and support for sequences up to 8,192 tokens. The model contains 307M parameters with 22 transformer layers, a hidden size of 768, and a vocabulary of 256,000 tokens derived from the Gemma 2 tokenizer.

To adapt mmBERT for hieroglyphic restoration, we fine-tune the pre-trained model under our custom MLM objective and using character tokenization on our training corpus, treating restoration as a masked token prediction task. Missing or corrupted hieroglyphic positions are represented as `MASK` tokens, and the model predicts the original symbols from their bidirectional context. Since the original mmBERT vocabulary does not include ancient Egyptian hieroglyphic tokens, we construct a domain-specific vocabulary from the training corpus, augment the tokenizer with these additional tokens, and resize the model's embedding matrix accordingly.

7 Experimental Design

7.1 Training Parameters

Training was performed using the model described in Section 6 and the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.01. We employed a cosine learning rate scheduler with a linear warmup over 3% of the total training steps. The model was trained for up to 100 epochs using a batch size of 32 for both training and evaluation. Validation was performed at the end of each epoch, and the checkpoint with the lowest validation loss was automatically selected as the final model.

To prevent overfitting, we applied early stopping with a patience of 10 steps based on validation loss. All experiments were conducted using `bfloat16` mixed-precision training to improve memory efficiency and training speed. The restoration model is evaluated on a held-out 10% test split consisting of 1,277 human-curated samples from the TLA dataset.

7.2 Training from scratch

Instead of initializing the model with pretrained mmBERT weights, we additionally train mmBERT from scratch on the same human-curated datasets consisting of the TLA and HieroTransformer corpora. The results show that the pretrained model outperforms the model trained from scratch, achieving an accuracy of 52.16%, while the model trained from scratch achieves 51.19%. This demonstrates the benefit of transfer learning and suggests that pretrained multilingual representations may already capture useful cross-lingual patterns that transfer to hieroglyphic text, improving the model's ability to learn semantic and contextual structure in low-resource settings.

Training Set	Top-1 (%)	Top-5 (%)
C: 30k bible only	29.73	51.16
D: 30k general only	25.62	44.33
Δ (C − D)	+4.11	+8.83

Table 3: Comparison of in-domain (biblical) vs. out-of-domain (general) training data on hieroglyphic restoration accuracy. Both conditions use 31,096 training sentences and are evaluated on a 10% human-annotated held-out split with 15% masking.

7.3 In-Domain vs Out-of-Domain Data

To isolate the effect of training data domain on restoration performance, we compare two conditions using identical data quantities: model C, trained exclusively on biblical text, and model D, trained exclusively on general-domain sentences. Both are evaluated on the human-annotated 10% held-out split of the HF dataset.

The results, shown in Table 3, demonstrate a clear advantage for in-domain training data. A Welch's t-test confirms that this difference is statistically significant ($t = 2.19$, $p = 0.029$ for Top-1; $t = 3.59$, $p = 0.0003$ for Top-5), indicating the performance gap is unlikely to arise by chance.

This finding is consistent with the thematic proximity between biblical corpora and ancient Egyptian hieroglyphic texts. Both register types are characterized by religious, funerary, and ceremonial discourse—a domain markedly underrepresented in general-purpose text collections. The higher lexical and topical overlap between biblical text and hieroglyphic inscriptions provides the model with more transferable co-occurrence statistics, resulting in more accurate predictions at masked positions.

7.4 Size of Synthetic Data

Augmenting the base corpus with synthetic data yields significant improvements across both evaluation metrics, as illustrated in Figure 5. The human-annotated baseline (HF + HieroTransformer) achieves 52.16% Top-1 and 71.48% Top-5 accuracy. Adding as few as 10k synthetic examples already raises performance to 56.47% Top-1 and 73.59% Top-5, gains of +4.31 pp and +2.11 pp respectively, demonstrating that even modest synthetic augmentation meaningfully enriches the training distribution.

Performance continues to improve as synthetic volume increases, with Top-1 accuracy reaching 58.22% (+6.06 pp over baseline) and Top-5 accuracy reaching 74.77% (+3.29 pp) at 150k synthetic examples, confirming that synthetic data provides a reliable and scalable source of supervision for hieroglyphic restoration.

The overall trajectory is consistently positive, with minor fluctuations across intermediate data volumes attributable to stochastic training dynamics rather than any systematic degradation. Beyond approximately 80k synthetic examples, marginal gains begin to stabilize: Top-1 accuracy across the 80k–150k range spans 57.11%–58.22%, with individual increments rarely exceeding 0.5 pp. This saturation pattern suggests that future improvements are more likely to come from increasing the quality and diversity of synthetic generation than from scaling data volume alone.

8 Related Work

There has been limited prior work addressing the reconstruction of missing or damaged hieroglyphs at the text level. HieroLM (Cai and Zhang, 2025) frames hieroglyph recovery as a next-word prediction task using an LSTM-based language model, arguing that purely vision-based approaches cannot exploit contextual and grammatical information. While HieroLM represents an important step, it relies on a unidirectional sequential model and does not exploit bidirectional context, a task our Transformer-based masked language model solves.

More broadly, lacuna reconstruction in ancient manuscripts has attracted growing interest across other writing systems. (Levine et al., 2024) apply a bidirectional RNN trained with masked language modeling to reconstruct missing characters in Coptic manuscripts, achieving 72% accuracy on single-character gaps. Similarly, the Ithaca project (Assael et al., 2022) applies deep learning to damaged Ancient Greek inscriptions, achieving 62% top-1 accuracy.

The broader computational Egyptology literature has largely focused on vision-based tasks. (Barucci et al., 2021) propose GlyphNet, a dedicated CNN architecture for hieroglyph classification, and (De Cao et al., 2024) frame hieroglyphic text as a source language for sequence-to-sequence translation into modern languages, both of which presuppose complete, legible input and do not address lacunae.

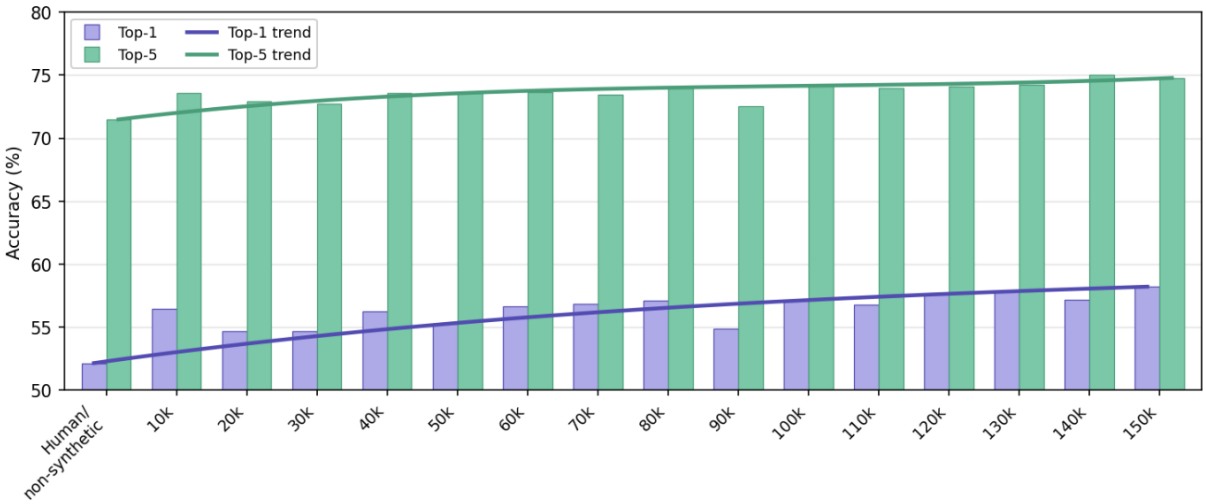

Figure 5: Effect of synthetic data scale on restoration performance. Aside from the human-annotated baseline, the x-axis represents the approximate amount of synthetic data added during training. Increasing the volume of synthetic training data improves both Top-1 and Top-5 accuracy, although the gains begin to plateau at larger data scales.

9 Conclusion

We presented SeshatLM, a transformer-based framework for the restoration of missing and damaged hieroglyphic text in ancient Egyptian manuscripts. By fine-tuning mmBERT under a custom span-based masked language modeling objective that closely mimics realistic patterns of artifact damage, SeshatLM achieves 58.22% Top-1 and 74.77% Top-5 accuracy in restoring 15% masked hieroglyphic sequences, a substantial improvement over the human-annotated baseline and over models trained from scratch, confirming the value of pretrained multilingual representations even in low-resource, non-modern-language settings.

A central contribution of this work is the demonstration that synthetic data augmentation is a useful and scalable strategy for improving hieroglyphic restoration under limited-resource conditions. We developed a German-to-hieroglyphics translation pipeline based on a fine-tuned TranslateGemma model with vocabulary-constrained decoding, and used it to generate a large-scale corpus of 150,000 synthetic hieroglyphic samples. Our ablation experiments confirm that in-domain synthetic data, drawn from thematically proximate sources such as biblical text, provides stronger supervision signal than general-domain alternatives, owing to the shared religious and ceremonial register with ancient Egyptian inscriptions.

We can additionally release our combined dataset of 178,164 hieroglyphic samples that were used for training SeshatLM, integrating human-curated TLA and HieroTransformer data with both in-domain and out-of-domain synthetic text, to support future research in computational Egyptology and hieroglyphic language modeling. We hope SeshatLM serves as a practical assistive tool for Egyptologists and historians engaged in the reconstruction of fragmentary inscriptions, and that the resources introduced here lower the barrier to entry for further work at the intersection of natural language processing and the digital preservation of ancient cultural heritage.

10 Limitations

SeshatLM is trained on a combination of natural and synthetic Hieroglyphic data. The synthetic data was generated by a model trained on a relatively small set of natural data, from the available public data sources. In general, the performance should be considered in the framing of an assistive tool rather than a standalone system.

References

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando De Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

Andrea Barucci, Costanza Cucci, Massimiliano Franci, Marco Loschiavo, and Fabrizio Argenti. 2021. A

deep learning approach to ancient egyptian hieroglyphs classification. *IEEE Access*, 9:123438–123447.

BibleNLP. 2023. eBible: Curated corpus of parallel data derived from versions of the bible provided by eBible.org. GitHub repository.

Xuheng Cai and Erica Zhang. 2025. Hierolm: Egyptian hieroglyph recovery with next word prediction language model. *Preprint*, arXiv:2503.04996.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Mattia De Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: a hieroglyphic transformer for translating Ancient Egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 71–86, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Alan H Gardiner. 1957. *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*. Griffith Institute, Ashmolean Museum, Oxford.

Google Translate Research Team, Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma Technical Report.

jonas-is coding. german-sentences.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Lauren Levine, Cindy Li, Lydia Bremer-McCollum, Nicholas Wagner, and Amir Zeldes. 2024. Lacuna language learning: Leveraging RNNs for ranked text completion in digitized Coptic manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 61–70, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Thes. Ling. Aegyptiae, 2024. 2024. Thesaurus linguae aegyptiae, original earlier egyptian sentences, corpus v18, premium.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Wikipedia contributors. 2024. List of egyptian hieroglyphs — Wikipedia, the free encyclopedia. [Online; accessed 5-February-2025].