Semantic Alignment across Ancient Egyptian Language Stages via Normalization-Aware Multitask Learning

He Huang
Institut für Ägyptologie und Koptologie, Ludwig-Maximilians-Universität München
Katharina-von-Bora-Straße 10, 80333 München, Germany
h.huang@campus.lmu.de

Abstract

We study word-level semantic alignment across four historical stages of Ancient Egyptian. These stages differ in script and orthography, and parallel data are scarce. We jointly train a compact encoder-decoder model with a shared byte-level tokenizer on all four stages, combining masked language modeling (MLM), translation language modeling (TLM), sequence-to-sequence translation, and part-of-speech tagging under a task-aware loss with fixed weights and uncertainty-based scaling. To reduce surface divergence we add Latin transliteration and IPA reconstruction as auxiliary views. We integrate these views through KL-based consistency and through embedding-level fusion. We evaluate alignment quality using pairwise metrics, specifically ROC-AUC and triplet accuracy, on curated Egyptian–English and intra-Egyptian cognate datasets. Translation yields the strongest gains. IPA with KL consistency improves cross-branch alignment, while early fusion demonstrates limited efficacy. Although the overall alignment remains limited, the findings provide a reproducible baseline and practical guidance for modeling historical languages under real constraints. They also show how normalization and task design shape what counts as alignment in typologically distant settings.

Keywords: Multitask Learning, Semantic Alignment, Low-Resource Languages

1. Introduction

Ancient Egyptian remains understudied in natural language processing (NLP), despite preserving a rich set of etymological cognates through several historical stages. This paper focuses on aligning four representative stages grouped into two major branches: pre-Coptic Egyptian (Hieroglyphic and Demotic) and Coptic (Sahidic and Bohairic), the final stage of Ancient Egyptian. Although genealogically related, these stages differ significantly in script, vocabulary, and grammar. For the purpose of our cross-lingual alignment framework, we treat them as distinct languages and aim to align them within a unified model.

This task presents several challenges: (i) There are no closely related high-resource languages to support effective transfer learning. (ii) The scripts vary widely. Hieroglyphic and Demotic use Egyptological transliteration schemes with some special signs like ꜣ and š, while Coptic combines signs borrowed from Greek (e.g., ⲁ) with signs derived from Pre-Coptic Egyptian (e.g., ϣ). Only Coptic preserves vowels, while pre-Coptic Egyptian contains many homographs due to unwritten vowels, leading to semantic ambiguity. (iii) Egyptian texts often come from different thematic domains (e.g., indigenous Pre-Coptic Egyptian polytheistic versus Coptic Christian texts), and cross-stage parallel texts are rare, making direct alignment difficult.

We investigate how multitask learning can support word-level alignment under these constraints. Our setup includes masked language modeling (MLM), translation language modeling (TLM), sequence-to-sequence translation, and universal part-of-speech (UPOS) tagging, combined via dynamic loss weighting. To mitigate script inconsistency, we introduce normalization strategies (Latin-based and IPA) integrated through Kullback–Leibler (KL) regularization or early fusion. These also serve as data augmentation techniques for other low-resource languages facing similar transliteration issues. We evaluate alignment quality using AUC and triplet accuracy, which better reflect semantic similarity than top-k recall in noisy, misaligned spaces. All code, data samples, and evaluation scripts are available in a GitHub repository.[1]

Rather than aiming for perfect alignment, our goal is to establish a realistic baseline and analyze the core difficulties. This work provides both methodological insights and empirical evidence for modeling historical language stages under minimal supervision.

2. Related Work

NLP for Egyptian and Coptic. Research on pre-Coptic Egyptian and Coptic has mainly focused on resource building and basic linguistic processing. For a survey, see (Muñoz Sánchez, 2024). Earlier works on Ancient Egyptian use resources from the project Thesaurus Linguae Aegyptiae (TLA) (AV

[1] Egyptian-alignment: https://github.com/Merythuthor/Egyptian-alignment.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8923–8936
11–16 May 2026.

Altägyptisches Wörterbuch, AV Wortschatz der ägyptischen Sprache and others, 2025b,a) and focuses on automated transliteration and automated translation from hieroglyphic signs or transliteration to English (Rosmorduc, 2020; De Cao et al., 2024; Miyagawa, 2025). For Coptic, the project Coptic SCRIPTORIUM (Caroline T. Schroeder, Amir Zeldes, et al., 2013-2025) serves as a foundational ecosystem for digital studies (Schroeder and Zeldes, 2020), providing annotated corpora, POS tagging, (Zeldes and Schroeder, 2015) dependency treebanks (Zeldes and Abrams, 2018; Zeldes et al., 2025), and core NLP pipelines (Zeldes and Schroeder, 2016). These resources have subsequently facilitated the development of low-resource language models, such as the development of the monolingual MicroBERT (Gessler and Zeldes, 2022). While a few studies attempt to model multiple stages of Egyptian in a unified framework (Sahala and Lincke, 2024; Miyagawa, 2025), semantic modeling across stages remains rare. Word-level semantic representation for Ancient Egyptian is largely understudied(Georgakopoulos and Polis, 2021), which motivates our focus on word-level semantic alignment.

Alignment objectives and multitask learning. Multitask learning with objectives like MLM, TLM, and supervised translation promotes shared cross-lingual representations (Conneau and Lample, 2019; Ruder et al., 2019; Kendall et al., 2018). Dynamic task weighting using uncertainty scaling has proven effective in balancing heterogeneous loss signals (Kendall et al., 2018).

For ancient languages, research primarily targets Ancient Greek and Latin. Studies have compared post-hoc alignment of monolingual spaces versus joint multilingual training for Latin-Greek alignment (Wang et al., 2020). Large language models have been evaluated for cross-lingual generalization across classical languages, demonstrating zero-shot transfer on various tasks (Akavarapu et al., 2025). Regarding word-level alignment, multilingual models fine-tuned on Greek-Latin parallel data via successive training stages have achieved strong alignment accuracy (Yousef et al., 2022).

Normalization for historical texts. Orthographic normalization plays a key role in historical text processing. Prior work highlights that normalization can improve tagging and retrieval but may affect linguistic fidelity depending on the phonological depth and sparsity of the data (Ruder et al., 2019; Amrhein and Sennrich, 2020). More recent work has demonstrated the benefits of transliteration for addressing script barriers in multilingual models (Liu et al., 2024). Consequently, our work treats normalization as an auxiliary representation and evaluates its effect on semantic alignment.

3. Methods

3.1. Dataset

The datasets of this study come from two main sources. For the pre-Coptic Egyptian branch, we utilize Hieroglyphic and Demotic texts provided by the Thesaurus Linguae Aegyptiae (TLA) project (AV Altägyptisches Wörterbuch, AV Wortschatz der ägyptischen Sprache and others, 2025b,a). The Coptic SCRIPTORIUM offers Sahidic and Bohairic texts from the Coptic branch (Caroline T. Schroeder, Amir Zeldes, et al., 2013-2025). Details regarding specific licensing constraints and reproducibility are provided in Section 9.

These four languages represent distinct stages of Ancient Egyptian, characterized by differences in script, orthography, and grammar. For illustrative examples of these diverse writing systems and their transliterated forms, see Appendix A. Hieroglyphic encompasses multiple historical stages, but we merge them into one unit due to unclear internal boundaries and limited data. While Sahidic and Bohairic share many lexical items, they differ in spelling and morphosyntax. Therefore, we treat them as two languages in our experiments.

The language grouping into two branches informs both our model design and evaluation strategy. Alignment patterns are later analyzed at both intra- and cross-branch levels.

The original datasets include translations in English or German. Only entries with translations are selected for this study. German texts were translated into English using OPUS-MT (de→en) (Tiedemann and Thottingal, 2020). UPOS tags are retained where available. Lemmatization is not used, as such annotations are missing for the Coptic data.

3.2. Model

We adopt a lightweight encoder-decoder architecture adapted to the four stages of Ancient Egyptian. Instead of fine-tuning a large multilingual model such as XLM-R, which is over-parameterized for our data and possesses a vocabulary incompatible with Egyptian transliteration, we train a shared, compact BERT encoder and a byte-level BPE tokenizer from scratch jointly on all four Egyptian stages. This design prevents out-of-vocabulary (OOV) fragmentation for special Unicode characters (e.g., ꜣ and ꜥ). English is incorporated in the encoder input through translation-related tasks, serving as a semantic pivot among different Egyptian languages.

The decoder is a standard Transformer initialized randomly and used only for sequence-to-sequence translation. The encoder and decoder share both the tokenizer and embedding matrix to enhance cross-representation consistency.

Fig. 1 summarizes our experimental workflow, including the four Egyptian language stages (illustrated by cognates for "consecration" in their respective orthographies), multitask setup, and evaluation process.

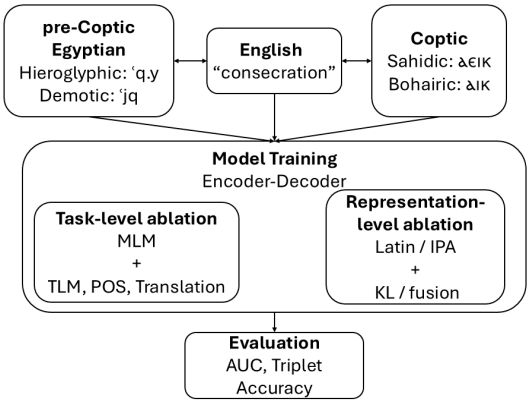

Figure 1: Overview of the experimental pipeline. The ablation study isolates two independent dimensions: (1) task-level supervision, without normalization; and (2) representation-level inputs under a standalone MLM task.

3.3. Baseline and Task-Level Comparisons

As a baseline, we trained the model using MLM on the original texts from all four Egyptian languages. To help the model distinguish language identity despite using a shared tokenizer, each input sequence is prepended with a language-specific token (e.g., `<hiero>`, `<dem>`, `<sah>`, `<boh>`).

We implement a multitask framework combining MLM, TLM, translation and UPOS tagging. While MLM learns token-level representations independently per language, the cross-lingual tasks (TLM, translation) use English (which is prepended with the `<eng>` token) to provide alignment signals. TLM jointly masks the Egyptian and English sequences, encouraging attention-based token alignment across the encoder (see Appendix B).

Example of TLM input: `[CLS] <hiero> nfr sw [SEP] <eng> he is good [SEP].`

Sequence-to-sequence translation leverages decoder cross-attention to directly guide alignment. UPOS tagging is used as an auxiliary task to provide syntactic supervision. We do not apply dictionary-based contrastive supervision due to noise and lack of lemma information in Coptic.

To balance the influence of each task, we apply a hybrid loss combining static weights and uncertainty-based adaptive scaling (Kendall et al., 2018). For task i, the total loss is:

$$\mathcal{L}_{\text{total}} = \sum_i W_i \left(\frac{1}{\sigma_i^2} \mathcal{L}_i + \log \sigma_i^2 \right), \qquad (1)$$

where W_i is a fixed prior weight and σ_i is a learnable parameter capturing task uncertainty. This design improves robustness in multitask training while allowing dynamic adjustment based on learning difficulty. Further implementation details and ablation settings are described in Section 4.

3.4. Representation-Level Normalization Strategies

To mitigate surface divergence caused by the heterogeneous character sets across Egyptological transliteration and Coptic script, we apply two normalization methods: Latin transliteration and IPA-based phonemic reconstruction. These are used as auxiliary views rather than replacements, with the original form preserved throughout training.

Unlike Egyptological transliteration with its special phonograms, the Latin normalization maps all graphemes to standard Latin characters (e.g., š, ꜩ → sh) (see sample text in Appendix B). Though not fully phonologically precise, it increases token overlap and highlights etymological similarities. The IPA scheme uses approximate phonemic representations to restore missing vowels in the pre-Coptic stages and unify consonant encoding. For instance, we render weak consonants like ꜣ, ꜥ, *i̯*, and w between two consonants as vowels *aa*, *a*, *i* and *u*, and fill remaining consonant gaps with *e*. This transforms *nfr* (Hieroglyphic, "good") into *nefer*, aligning it more closely with its Coptic cognate ⲛⲟⲩϥⲉ "good", converted into *noufe*.

Full normalization mappings are provided as `training/normalization.py` in the GitHub repository (see footnote 1). English text is not normalized into IPA to prevent misleading the model with coincidental sound overlap across unrelated language families.

We acknowledge that such normalization introduces simplifications and potential semantic drift, especially given the complexity of Egyptian phonology and its transliteration systems. Nonetheless, this rudimentary attempt serves as a first step toward understanding how orthographic normalization may aid cross-lingual semantic modeling, offering valuable guidance for future refinement.

3.5. View Integration Methods

To preserve the original orthography to retain semantic richness while providing phonological regularity, we compare two methods for integrating these auxiliary views into training:

(A) Consistency via KL Divergence. We adopt symmetric Kullback–Leibler (KL) divergence to enforce bidirectional alignment between the predictive distributions of original and normalized forms, effectively performing mutual distillation across orthographic views.

Given the token-level MLM predictions from both views, we minimize:

$$\mathcal{L}_{\text{consistency}} = \frac{1}{2}\Big[D_{\text{KL}}(p_{\text{orig}} \parallel p_{\text{norm}}) + D_{\text{KL}}(p_{\text{norm}} \parallel p_{\text{orig}})\Big], \quad (2)$$

where gradients are detached on the teacher side to prevent representational collapse.

The total loss is:

$$\mathcal{L}_{\text{total}} = W_{\text{MLM}} \left(\frac{1}{\sigma^2_{\text{MLM}}} \mathcal{L}_{\text{MLM}} + \log \sigma^2_{\text{MLM}} \right) + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{consistency}}, \quad (3)$$

with λ_{KL} as a fixed hyperparameter (`--consistency_lambda 0.5`).

This regularization encourages the encoder to learn view-invariant representations across orthographic variants. While this strategy is applicable to tasks beyond MLM (e.g., translation or POS tagging), our experiments restrict it to MLM due to computational resource limits.

(B) Early Fusion via Embedding Mixture. We integrate the original and normalized views at the embedding level by constructing a learnable weighted mixture, enabling soft alignment without requiring multiple forward passes.

Given token embeddings E_{orig} and E_{norm}, we compute a fused representation:

$$E_{\text{fused}} = \alpha \cdot E_{\text{orig}} + (1 - \alpha) \cdot E_{\text{norm}}, \quad (4)$$

where $\alpha \in (0, 1)$ is derived from a learnable scalar via sigmoid activation.

The fused embedding is passed to the encoder to compute the MLM loss as usual. The total loss is:

$$\mathcal{L}_{\text{total}} = W_{\text{MLM}} \left(\frac{1}{\sigma^2_{\text{MLM}}} \mathcal{L}_{\text{MLM}} + \log \sigma^2_{\text{MLM}} \right). \quad (5)$$

Unlike KL divergence, this method requires only one forward pass per example and introduces no auxiliary loss term, making it computationally efficient. The model learns to balance both views during training: when $\alpha \to 1.0$, it favors the original form; when $\alpha \to 0.0$, it prefers the normalized variant. This soft fusion lets the model learn representations that incorporate phonological regularity while retaining language-specific features under a single, global learned weighting parameter.

Strategy Comparison. These strategies address complementary aspects of multi-representation learning. KL consistency provides output-level regularization via bidirectional distillation, whereas early fusion performs input-level soft integration. Together, they allow us to empirically evaluate whether early fusion outperforms late-stage consistency in capturing script-invariant semantics under low-resource constraints.

It is important to note an implementation detail regarding both strategies: our current formulations of KL consistency and early fusion operate via strict position-wise alignment across padded subword sequences. We analyze the structural effects of this design choice in Section 6.3.

4. Experiments

4.1. Dataset Preparation

We standardized lacuna markers (e.g., '—', '...', '<gap>') into one or more `[gap]` tokens. Sentences with gaps were kept. To retain essential morphological and syntactic information, we preserved key linguistic markers, which include the period (`.`), separating stems from gender and number endings, and the suffix sign (⸗) in pre-Coptic Egyptian, and the hyphen (–).

Table 1 shows corpus statistics after preparation. To preserve the natural distribution of the attested corpora, we did not apply any sentence-level resampling during tokenizer or model training to mitigate the imbalance across Egyptian languages.

Lang	Sentences	Sent%	Tokens	Tok%
H	145,060	65%	2,705,136	56%
D	30,488	14%	646,327	13%
S	25,902	12%	800,198	17%
B	20,538	9%	693,097	14%
TOTAL	**221,988**	**100%**	**4,844,758**	**100%**

Table 1: Corpus size and distribution. Lang = language stage (H = Hieroglyphic, D = Demotic, S = Sahidic, B = Bohairic). Token counts are computed using the shared byte-level BPE tokenizer (excluding special tokens).

Each entry consists of the original transcription, its English translation, and a language identifier. No normalization is applied during preprocessing;

instead, normalization is introduced during training, as described in Section 3.4.

4.2. Training Setup

We trained a shared byte-level BPE tokenizer (vocab size: 32,000, min frequency: 2) on the combined corpus of all four Egyptian language stages and their English translations. The vocabulary also includes task-specific tokens (e.g., `[CLS]`, `[MASK]`), language tags (e.g., `<hiero>`), and the gap marker `[gap]`. To isolate the effect of cross-lingual alignment mechanisms, English data were strictly limited to in-domain parallel texts.

We intentionally excluded large-scale monolingual English corpora for three main reasons: (i) introducing massive English data would lead to severe over-representation, causing it to disproportionately dominate the shared vocabulary and model capacity; (ii) modern English corpora introduce significant domain shift, as contemporary vocabulary diverges from the specific historical and religious contexts of Ancient Egyptian, thereby injecting noise into the shared semantic space; and (iii) under our computational constraints, a restricted, in-domain setup maximizes training efficiency without wasting resources on irrelevant data.

The tokenizer was trained exclusively on the original historical scripts and the English translations, without exposure to the normalized Latin and IPA views. While this strictly grounds the vocabulary in the attested corpora, it forces the tokenizer to over-segment the unseen normalized inputs into much shorter subwords (borrowed from English) or raw character fallbacks, leading to higher sequence fragmentation.

The dataset is split into training/validation/test sets in an 8:1:1 ratio.

All models are trained for 10 epochs using the encoder-decoder architecture described in Section 3, with 6 encoder and 6 decoder layers, hidden size 768, 12 attention heads, and max sequence length 768. The encoder and decoder share embeddings. Training was performed on a single NVIDIA RTX 5090 GPU with a batch size of 16, learning rate of $5 \cdot 10^{-5}$, bfloat16 mixed precision, cosine decay with 500 warm-up steps, and gradient accumulation set to 2.

4.3. Ablation Design

We perform ablation studies along two independent dimensions, using an MLM-only model trained on the original texts (without normalization) as our unified baseline.

Task-level ablation. To evaluate the contribution of each auxiliary task, we selectively activate or deactivate them during training with the original scripts, without normalization. When a task is active, its fixed prior weight (W_i, as defined in Section 3.3) is set to a predefined value ($W_{\text{MLM}} = 1.0$, $W_{\text{Trans}} = 1.0$, $W_{\text{TLM}} = 1.0$, and $W_{\text{POS}} = 0.5$). To ablate a task, its weight is set to zero. The MLM objective is always retained ($W_{\text{MLM}} = 1.0$).

Representation-level ablation. To assess the impact of normalization, we evaluate three input formats under the MLM-only setup: (1) raw orthography, (2) Latin transliteration, and (3) IPA-based phonemic representation. Each normalization is paired with either KL consistency (fixed $\lambda_{\text{KL}} = 0.5$) or early fusion (trainable α, initialized to 0.5) to evaluate integration strategies.

5. Evaluation

Our evaluation focuses on cross-lingual semantic alignment at the word level across four historical stages of Ancient Egyptian. As our objective is not sequence generation or syntactic tagging, we omit BLEU and POS F1 as primary metrics.

5.1. Visualization of Clustering Geometry

We begin by visualizing the structure of the multilingual embedding space. This subsection illustrates why nearest-neighbor search may be unreliable in our setting. Instead of relying on top-k lexical retrieval, which assumes a well-aligned semantic space, we use relative distance-based metrics (see Subsection 5.3).

For each word, we compute its embedding by mean-pooling subword vectors over all occurrences in the validation and test sets. We visualize 500 (or 400, when including English) word-level embeddings per language, sampled with a frequency of $\geq$3, using t-SNE (perplexity=30, init=PCA, n_iter=2000) after L2 normalization. All embeddings are produced using the shared byte-level BPE tokenizer.

Fig. 2 shows the embedding space for the four Egyptian languages. The MLM-only baseline yields four distinct language clusters. The languages with their respective branches naturally cluster together. Since internal coherence is a prerequisite for alignment, the left panel suggests that while auxiliary supervision in multitask training (MLM+TLM+Translation+POS) might improve alignment within the Coptic branch, it weakens the coherence of pre-Coptic Egyptian scripts. Consequently, overall alignment remains limited. Most word embeddings stay concentrated within their respective language clusters, with minimal cross-lingual overlap.

Fig. 3 includes English as a fifth language through the MLM-only baseline. English acts as a semantic pivot, inducing radial clustering and hubness. While this improves global comparability, it distorts local neighborhoods among Egyptian variants. Incorporating Latin normalization via KL consistency increases the global spatial separation of the Egyptian clusters instead of merging them. We attribute this paradox to normalization acting as a strong regularizer that strips away surface noise while harmonizing surface orthography. Consequently, the MLM objective is forced to cluster tokens by underlying language-specific morphosyntactic structures, isolating each language into a distinct subspace.

Nevertheless, it is important to note that t-SNE visualizations primarily illustrate macro-level topological clustering and do not strictly quantify word-level semantic alignment precision. Because the embeddings are globally segregated by language, direct top-k retrieval fails to provide a faithful measure of cross-lingual semantic overlap under current settings and should only be used after successful normalization or alignment strategies are applied.

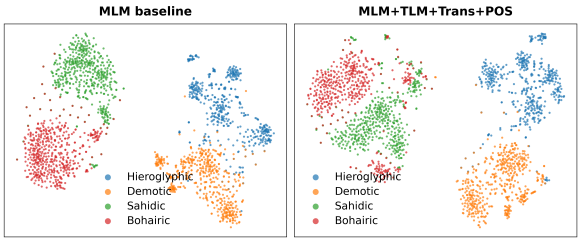

Figure 2: t-SNE visualization of word embeddings across four Egyptian language stages through the MLM-only baseline and the full multitask model (MLM+TLM+Translation+POS).

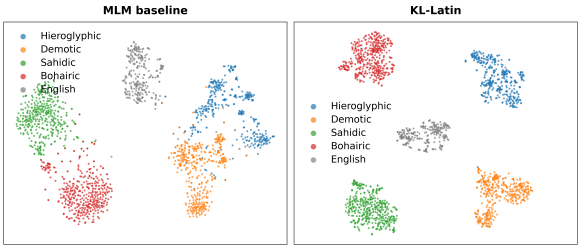

Figure 3: t-SNE visualization including English as a pivot language through the MLM-only baseline with or without Latin normalization and KL consistency training.

5.2. Evaluation Setup and Datasets

We assess semantic alignment using alignment-agnostic metrics, namely triplet accuracy and ROC-AUC, computed over curated positive and negative word pairs. These metrics assume only that true cognates should be closer than unrelated pairs on average. Word embeddings are computed by averaging subword vectors from the shared BPE tokenizer.

This setup is robust to global embedding shifts and script-based offsets, making it suitable for low-resource, cross-script scenarios. Unlike top-k retrieval, which assumes the existence of well-aligned vector spaces, these relative distance-based metrics evaluate relative semantic proximity without requiring absolute alignment. Due to the genre divergence across stages (e.g., indigenous pre-Coptic Egyptian religious vs. Coptic biblical texts), shared contexts are rare. We thus compute word embeddings by averaging over all occurrences and fix the random seed for consistency.

In this framework, we evaluate semantic alignment at two levels:

Egyptian–English. We collect dictionary pairs from TLA and Coptic SCRIPTORIUM. These test whether the model can associate Egyptian words with their corresponding English translations. To assess generalization, we split the paired examples into *Seen* (where the Egyptian word and its English equivalent co-occurred within the same sentence-translation pair during training) and *Unseen*. A strong model should perform well even on *Unseen* pairs.

Intra-Egyptian. The intra-Egyptian evaluation dataset is based on cognate pairs among Egyptian languages produced by TLA, which are accessed under research terms and cannot be redistributed. We remove duplicates and low-frequency terms and distinguish between Sahidic and Bohairic forms. To illustrate the data structure, we provide sample pairs of these cognate pairs in Appendix C, and include ten example groups in `/resource_eval/` in the GitHub repository (see footnote 1).

We categorize examples as:

- **Cross-branch:** e.g., Hieroglyphic–Sahidic
- **Within-branch:** e.g., Sahidic–Bohairic
- **Homographs vs. Heterographs:** Defined by whether the forms share identical spelling

This evaluation directly addresses our core research question: can models capture semantic correspondence across orthographically divergent but genealogically related stages?

5.3. Evaluation Metrics

All metrics operate at the word level, using cosine similarity $\text{sim}(\cdot,\cdot)$. For each language pair, we use up to N gold pairs $\mathcal{P} = \{(w_i, w_i^+)\}_{i=1}^N$ (detailed evaluation dataset statistics and valid pair counts are provided in Appendix D), and construct matched negatives $\mathcal{N} = \{(w_i, w_i^-)\}_{i=1}^N$ by randomly sampling non-cognates from the same target language.

Triplet Accuracy. Measures whether each true pair is more similar than its negative counterpart:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\text{sim}(w_i, w_i^+) > \text{sim}(w_i, w_i^-)\right] \quad (6)$$

The result reflects the proportion of triplets where the model assigns higher similarity to the correct pair. Since each comparison is local and independent, this metric is sensitive to the model's decision for each word. The expected value of a random model is 50%.

ROC-AUC. AUC is computed over the entire distributions of positive and negative similarity scores. It estimates the probability that a randomly chosen positive pair is more similar than a randomly chosen negative:

$$\text{AUC} = \frac{1}{|\mathcal{P}||\mathcal{N}|}\sum_{p\in\mathcal{P}}\sum_{n\in\mathcal{N}} \mathbf{1}\left[\text{sim}(p) > \text{sim}(n)\right] \quad (7)$$

This metric considers all cross-pair combinations between positive and negative sets. It is invariant to monotonic transformations and global shifts in embedding space, providing a holistic view of score separability.

Together, Triplet Accuracy and ROC-AUC offer a comprehensive view of alignment performance in noisy, cross-script scenarios where top-k recall is unreliable.

6. Results

Our main results are presented in two heatmaps detailing alignment performance across tasks and normalization strategies (Fig. 4 and Fig. 5). They report Egyptian–English results and intra-Egyptian alignment respectively, and will be analyzed separately in the following subsections.

In both figures, the left subfigure varies task settings (MLM, +TLM, +Translation, +POS), while the right subfigure varies representation strategies (Latin, IPA) and integration methods (KL consistency, early fusion). The highest AUC in each row is highlighted with a white rectangle.

Notation: C = cross-branch; I = intra-branch; Ht = heterograph; Ho = homograph; D/H/S/B = Demotic/Hieroglyphic/Sahidic/Bohairic; E = English; S/U = *Seen/Unseen*.

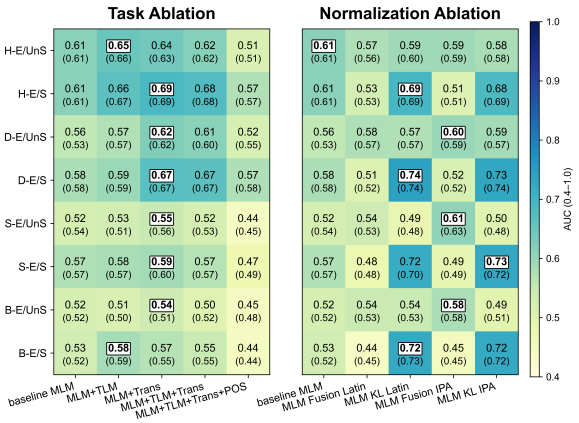

Figure 4: Alignment performance (AUC and Accuracy) between Ancient Egyptian and English under different multitask and normalization settings. Each cell displays AUC (top) and Accuracy (bottom, in parentheses), rounded to two decimal places. The colormap encodes AUC scores from 0.40 to 1.00, which serves as the primary metric; Accuracy is shown for reference. The highest AUC in each row is highlighted with a white rectangle.

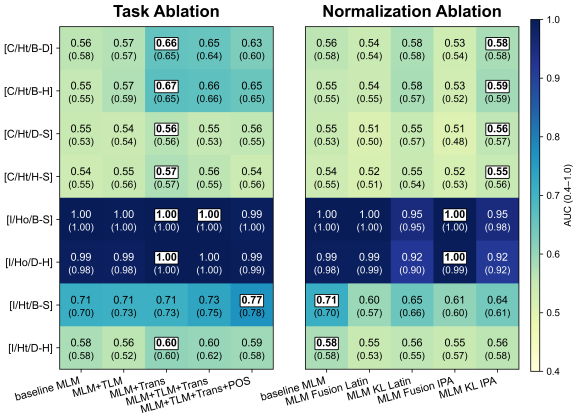

Figure 5: Intra-Egyptian alignment performance (AUC and Accuracy) across historical stages and training variants. The highest AUC in each row is highlighted with a white rectangle.

6.1. Result Analysis: Egyptian-English

Task Ablation As shown in the left panel of Fig. 4, AUC and Accuracy follow consistent trends. *Seen* pairs slightly outperform *Unseen* pairs, suggesting that training provides moderate alignment

gains. Among Egyptian varieties, Demotic and Hieroglyphic achieve better alignment with English than the Coptic languages. While Hieroglyphic benefits from a larger corpus, Demotic's performance is notable despite its smaller size. This may stem from their shared orthographic overlap with English, with the exception of a few special signs. In contrast, Coptic uses Greek-based characters, making surface alignment with English harder, a pattern further confirmed in the normalization experiments.

Across training objectives, two-task models generally outperform MLM-only. MLM+Translation consistently yields stronger alignment than MLM+TLM. Adding POS offers no clear improvement over two-task setups despite being a word-level task, highlighting the need for careful multitask design.

Normalization Ablation On the right panel of Fig. 4, early fusion performs poorly. In most cases, the *Seen* scores are even lower than *Unseen*, and the overall performance is similar to or below the MLM baseline. This indicates that early fusion tends to disrupt rather than enhance cross-lingual alignment. Nevertheless, when combined with IPA normalization, early fusion outperforms the other groups in the *Unseen* scenario for Demotic, Sahidic, and Bohairic. This unexpected pattern warrants further investigation in future work.

In contrast, KL-based consistency training yields a clear *Seen-Unseen* gap in favor of *Seen*, indicating that the model captures stronger alignment signals in this setting. Both IPA and Latin normalization show comparable results on *Seen* pairs and perform much better than the model without normalization. On *Unseen* pairs, the improvement is smaller, suggesting limited generalization ability.

Across the four stages, Demotic, Sahidic, and Bohairic align more closely to English than Hieroglyphic. This outcome is not fully explained by corpus size alone, since Hieroglyphic is also the largest corpus. Domain differences across sources may be a factor, but require further analysis. Overall, normalization under KL consistency improves Egyptian–English alignment, while early fusion remains ineffective and sometimes counterproductive.

6.2. Result Analysis: Intra-Egyptian

Task Ablation AUC and Accuracy in Fig. 5 show similar trends. Alignment is stronger within branches than across, likely due to shared transliteration and more identical cognate forms. In cross-branch cases, adding translation helps Bohairic align better to the pre-Coptic stages than Sahidic, despite having less data. This is likely due to Bohairic's limited corpus size and restricted semantic domain, allowing semantic alignment via English to be more easily learned.

Homographs within branches reach near-perfect alignment. Even the MLM-only baseline achieves strong results without cross-lingual supervision. This within-branch alignment is primarily driven by their shared transliteration schemes. As illustrated in Appendices A and C, Hieroglyphic and Demotic pairs are often transcribed into identical or highly similar surface forms.

For heterographs, Sahidic–Bohairic performs markedly better than Demotic–Hieroglyphic. This indicates that the Coptic branch exhibits stronger orthographic distinctiveness. Because Coptic preserves vowels, words with different meanings are more clearly differentiated from each other, allowing true cognates to remain recoverable even with minor dialectal spelling variations.

Pre-Coptic Egyptian, however, contains a much greater number of homographs due to unwritten vowels. In the original script, semantic ambiguity was resolved by iconographic features like determinatives, but these crucial visual cues are lost in the transliterations used for training. Consequently, even identical phonetic representations of consonants can map to multiple distinct meanings, severely complicating the disambiguation of homographs and the alignment of heterographs.

This observation strongly motivates adding lemma identifiers where available to disambiguate pre-Coptic Egyptian homographs, while also suggesting that the absence of lemma IDs on the Coptic side may be less damaging than initially expected.

Across multitask settings, MLM+Translation consistently outperforms other settings. Adding TLM or POS to this setup brings little further gain. This implies that, under limited word-level supervision, sequence-level translation signals are the strongest driver of intra-Egyptian alignment.

Unlike in Egyptian–English alignment, POS does not harm performance here, likely due to shared underlying syntax across Ancient Egyptian stages. Adding POS especially helps the alignment between dialectal variants of Sahidic and Bohairic. This is probably due to the syntactic stabilization during the Coptic stage, when grammar became more standardized and structured than in the pre-Coptic periods.

Normalization Ablation The vertical patterns across evaluation groups largely match the task ablation trends and are not repeated here. Focusing on normalization choices, IPA with KL consistency gives the best cross-branch alignment. A likely reason is that IPA partially restores the

vowel structure of pre-Coptic Egyptian, facilitating connections to Coptic varieties containing vowels. The gains, however, are modest, indicating that finer-grained phonological reconstruction rules are needed.

In contrast, for within-branch heterographs, normalization often reduces alignment quality. This supports the view that normalization can remove useful language-specific cues and distort originally informative spellings. When surface forms are already similar, further normalization may harm both semantic proximity and other latent signals. These findings support adopting selective normalization policies. Phonology-oriented normalization should be applied only for distant scripts and cross-branch comparisons.

6.3. Error Analysis: The Subword Fragmentation Bottleneck

While translation yields strong gains, early fusion and KL consistency performed below expectations in certain configurations (e.g., early fusion consistently underperforming the baseline). To understand this, we conducted a qualitative error analysis on the tokenized inputs (see Appendix B).

We identified a structural bottleneck rooted in subword-level positional misalignment. As noted in Section 4.2, the shared BPE tokenizer was not exposed to the normalized views during training. Consequently, Latin and IPA sequences suffer from severe over-segmentation. For example, a single word in the original script might be tokenized into two subwords, while its Latin or IPA equivalent is fragmented into four or more short subwords or fallback tokens.

Because our integration mechanisms (KL consistency and early fusion) currently rely on strict position-wise alignment across padded sequences, this length discrepancy causes a semantic mismatch. The model is frequently forced to align semantically disparate linguistic fragments (e.g., matching the root of a verb in the original script with a meaningless subword or phonetic affix in the elongated IPA sequence). This rigid token-level alignment inadvertently introduces positional noise, explaining the limited gains of our fusion strategies and highlighting a fundamental challenge in integrating word-level normalization with subword-level architectures.

7. Conclusion and Future Work

This study investigates cross-lingual alignment across four historical stages of Ancient Egyptian. We find that sequence-level multitask learning remains effective under low-resource, high-noise conditions. In particular, translation serves as a strong supervision signal for word-level alignment, outperforming TLM and POS. Normalization combined with KL consistency further improves alignment with English, though it may weaken internal consistency within the four stages of Ancient Egyptian, revealing a trade-off between external comparability and internal coherence.

Aligning historical languages is inherently challenging. The languages differ not only in vocabulary and syntax, but also in scripts and orthographic conventions. These surface differences obscure genealogical links and complicate semantic alignment. Despite these challenges, our results offer a solid foundation for computational research on historical language families and highlight where alignment methods succeed and where they fail. This can support future work on underrepresented and endangered languages with similar structural diversity.

Looking ahead, several directions remain open. The current normalization scheme can be refined with more linguistically informed rules. Lemma information, especially for pre-Coptic Egyptian, should be incorporated to reduce ambiguity from homographs. Finally, this alignment framework can support future studies of semantic change by tracing cognates across time in a shared embedding space.

8. Limitations

Normalization. While normalization reduces surface divergence, it introduces several risks:

- **Information loss.** Distinct consonants (e.g., ḥ, ḫ and ẖ → *kh*) can be conflated into the same grapheme.
- **False similarity.** Over-aggressive mappings can make unrelated forms appear similar, misleading alignment and evaluation.
- **Oversimplification.** The actual principles governing vowel realization in pre-Coptic Egyptian are more complex than the approximations used in this experiment.

Mismatch between Normalization and Token-Level Alignment. As detailed in our error analysis (Section 6.3), integrating word-level normalization with subword-level architectures introduces positional noise due to sequence length divergence. To resolve this structural mismatch, future work should investigate boundary-preserving normalization. Instead of independently tokenizing the normalized text, we could preserve the token boundaries generated from the original Egyptian script and directly map their normalized forms into the vocabulary.

Although this risks disrupting context-dependent phonological rules at token boundaries, it guarantees equal input sequence lengths across views, allowing token-level KL divergence and early fusion to operate reliably without requiring complex soft-alignment algorithms like dynamic time warping (DTW).

Granularity of Early Fusion. In our experiments, the early embedding fusion strategy consistently underperformed compared to late-stage KL consistency. We hypothesize that a core limitation of our current implementation is the reliance on a single, global learnable scalar (α) to control the fusion ratio for all tokens across the entire corpus. While computationally efficient, this global parameter forces a uniform trade-off between original orthography and normalized phonology. Given the heterogeneity across different Egyptian scripts and the varying degrees of normalization opacity, this single-scalar approach lacks flexibility. Future work should investigate fine-grained integration, such as token-level gating mechanisms or language-pair-specific fusion weights, which would allow the model to dynamically balance original and normalized representations on a word-by-word basis.

Lack of lemmatization for pre-Coptic Egyptian. We did not incorporate lemma annotations for pre-Coptic Egyptian, omitting a strong source of supervision that could disambiguate homographs and stabilize word-level alignment. This likely contributes to weaker Egyptian-English alignment performance.

Dependence on English Pivot. Our current architecture relies heavily on English as a semantic pivot during training. While our intra-Egyptian evaluation demonstrates that the resulting word embeddings do achieve certain direct semantic alignment (as English is entirely absent during this evaluation phase), it remains challenging to strictly isolate how much of this alignment was learned independently of English supervision. Future ablation studies completely removing the English pivot are necessary to quantify direct cross-branch transfer capabilities.

9. Data and Code Availability

All code, data samples, and evaluation scripts used in this study are available in our GitHub repository (https://github.com/Merythuthor/Egyptian-alignment).

Regarding data availability, the resources vary by language branch. For Coptic, we utilize all available Sahidic and Bohairic corpora from Coptic SCRIPTORIUM (Caroline T. Schroeder, Amir Zeldes, et al., 2013-2025) (https://github.com/CopticScriptorium/corpora), which is currently the largest annotated open-access Coptic resource online.

For pre-Coptic Egyptian, we use a version of the dataset from Thesaurus Linguae Aegyptiae (TLA) for Hieroglyphic and Demotic texts (AV Altägyptisches Wörterbuch, AV Wortschatz der ägyptischen Sprache and others, 2025b,a). These data are accessed under research terms and cannot be redistributed in accordance with TLA's policies. However, a version of the dataset sharing the similar annotation format is publicly available at https://huggingface.co/thesaurus-linguae-aegyptiae, allowing readers to examine the data structure. To ensure maximum reproducibility under these constraints, we provide our comprehensive data processing script (`data/clean_corpora.py`) in the GitHub repository of this paper, enabling researchers to recreate our exact experimental setup once they obtain official TLA access.

10. Acknowledgements

I would like to sincerely thank Thesaurus Linguae Aegyptiae (TLA) for providing the necessary data access via their official API and Hugging Face and for maintaining such a vital resource for the research community.

I also extend my sincere thanks to Dr. Caroline T. Schroeder and Dr. Amir Zeldes from the Coptic SCRIPTORIUM for their continuous efforts in maintaining this essential open-access dataset, as well as for their valuable discussions and feedback that greatly benefited the current study.

I am deeply grateful to my doctoral supervisor, Dr. Friedhelm Hoffmann, for his invaluable instruction in the various historical stages of Ancient Egyptian and for inspiring my interest in language alignment. Furthermore, I sincerely appreciate his continuous support and encouragement in my pursuit of new technologies and interdisciplinary approaches beyond the traditional boundaries of Egyptology.

I would also like to sincerely thank Dr. Daniel A. Werning, Dr. Peter Dils, Dr. Simon D. Schweitzer, Dr. Eliese-Sophia Lincke, Dr. Tonio Sebastian Richter, Dr. Frank Feder, Yihong Liu, Yunting Xie, and Du Cheng for their insightful suggestions and valuable discussions.

11. Bibliographical References

V. S. D. S. Mahesh Akavarapu, Hrishikesh Terdalkar, Pramit Bhattacharyya, Shubhangi Agarwal, Vishakha Deulgaonkar, Pralay Manna, Chaitali Dangarikar, and Arnab Bhattacharya. 2025. A case study of cross-lingual zero-shot generalization for classical languages in llms.

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*.

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.

Athanasios Georgakopoulos and Stéphane Polis. 2021. Lexical diachronic semantic maps. mapping the evolution of time-related lexemes. *Journal of Historical Linguistics*, 11(3).

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of CVPR*.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.

So Miyagawa. 2025. RAG-enhanced neural machine translation of Ancient Egyptian text: A case study of THOTH AI. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 33–40, Albuquerque, USA. Association for Computational Linguistics.

Ricardo Muñoz Sánchez. 2024. When hieroglyphs meet technology: A linguistic journey through Ancient Egypt using natural language processing. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 156–169, Torino, Italia. ELRA and ICCL.

Serge Rosmorduc. 2020. Automated Transliteration of Late Egyptian Using Neural Networks. *Lingua Aegyptia - Journal of Egyptian Language Studies*, 28:233–257.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Aleksi Sahala and Eliese-Sophia Lincke. 2024. Neural lemmatization and POS-tagging models for Coptic, demotic and earlier Egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 87–97, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Caroline T. Schroeder and Amir Zeldes. 2020. A collaborative ecosystem for digital coptic studies. *Journal of Data Mining and Digital Humanities*. Special issue on Collecting, Preserving, and Disseminating Endangered Cultural Heritage.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency treebank. In *Proceedings of the Second Workshop on Universal De-*

pendencies (UDW 2018), pages 192–201, Brussels, Belgium. Association for Computational Linguistics.

Amir Zeldes and Caroline T. Schroeder. 2015. Computational methods for coptic: Developing and using part-of-speech tagging for digital scholarship in the humanities. *Digital Scholarship in the Humanities*, 30(suppl_1):i164–i176.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for Coptic. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.

Amir Zeldes, Nina Speransky, Nicholas E. Wagner, and Caroline T. Schroeder. 2025. A UD treebank for bohairic Coptic. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 59–69, Ljubljana, Slovenia. Association for Computational Linguistics.

12. Language Resource References

AV Altägyptisches Wörterbuch, AV Wortschatz der ägyptischen Sprache and others. 2025a. *Thesaurus Linguae Aegyptiae: Demotic Texts*. Berlin-Brandenburgische Akademie der Wissenschaften and Sächsische Akademie der Wissenschaften zu Leipzig. PID https://thesaurus-linguae-aegyptiae.de/. Full author list: AV Altägyptisches Wörterbuch, AV Wortschatz der ägyptischen Sprache, Günter Vittmann, Marcel Moser, Joachim Friedrich Quack, Martin Stadler, R. Dominik Blöse, Jakob Höper, Simon D. Schweitzer, Daniel A. Werning. Editors: Tonio Sebastian Richter & Daniel A. Werning on behalf of BBAW, and Hans-Werner Fischer-Elfert & Peter Dils on behalf of SAW Leipzig.

AV Altägyptisches Wörterbuch, AV Wortschatz der ägyptischen Sprache and others. 2025b. *Thesaurus Linguae Aegyptiae: Hieroglyphic and Hieratic Texts*. Berlin-Brandenburgische Akademie der Wissenschaften and Sächsische Akademie der Wissenschaften zu Leipzig. PID https://thesaurus-linguae-aegyptiae.de/. Full author list: AV Altägyptisches Wörterbuch, AV Wortschatz der ägyptischen Sprache, Stefan Baumann, Ralph Birk, Anke Blöbaum, R. Dominik Blöse, Marc Brose, Adelheid Burkhardt, Svenja Damm, Sophie Diepold, Peter Dils, Marlies Elebaut, Frank Feder, Erhart Graefe, Silke Grallert, Sebastian Hoedt, Jakob Höper, Kay Christine Klinger, Ines Köhler, Emilia Mammola, Aurélie Paulet, Lutz Popko, René Preys, Andreas Pries, Daniel von Recklinghausen, Alexa Rickert, Josefine Bar Sagi, Jakob Schneider, Simon D. Schweitzer, Lisa Seelau, Gunnar Sperveslage, Jan Tattko, Doris Topmann, Jonas Treptow, Susanne Töpfer, Anja Weber, Daniel A. Werning, Annik Wüthrich. Editors: Tonio Sebastian Richter & Daniel A. Werning on behalf of BBAW, and Hans-Werner Fischer-Elfert & Peter Dils on behalf of SAW Leipzig.

Caroline T. Schroeder, Amir Zeldes, et al. 2013-2025. *Coptic SCRIPTORIUM*. 6.0.0. PID http://copticscriptorium.org/. Full author list: Caroline T. Schroeder, Amir Zeldes, et al.

Appendices

Appendix A: Sample Sentences across Egyptian Language Stages

To provide context for the orthographic diversity of texts within our dataset, we present short illustrative examples of the cleaned text from each of the four language stages used in our training corpus.

Notice the visual and structural differences: pre-Coptic Egyptian (Hieroglyphic and Demotic) relies on a shared transliteration scheme from the original signs without vowels, whereas the Coptic dialects (Sahidic and Bohairic) use Greek-derived alphabets that also record vowels.

- **Hieroglyphic (Transliteration):** ḏi ⸗s n ⸗k t̲ɜw r fnḏ ⸗k ꜥnḥ ⸗k
  English Translation: "She'll give you air to your nose to make you live."
- **Demotic (Transliteration):** ḏd ⸗y n ⸗f r ỉr wꜥ t̲k ỉ.ỉr-ḥr ⸗tn
  English Translation: "I told him to make a letter to you."
- **Sahidic (Coptic Script):** ϩⲛ ⲡⲉ ϩⲟⲟⲩ ⲉⲛⲧ ⲁ ⲡ ϫⲟⲉⲓⲥ ⲱⲁϫⲉ ⲙⲛ ⲙⲱⲩⲥⲏⲥ ϩⲣⲁⲓ ϩⲛ ⲕⲏⲙⲉ
  English Translation: "in the day in which the Lord spoke to Moses in the land of Egypt."
- **Bohairic (Coptic Script):** ⳇ ⲟⲓ ⲛ ϩⲟϯ ⲟⲩⲟϩ ⳇ ⲟⲩⲟⲛϩ ⲉⲃⲟⲗ ⲡⲉⳇ ϩⲁⲡ ⲉⳇⲉ ⳍⲱⲡⲓ ⲉⲃⲟⲗ ⲛϧⲏⲧ ⳇ
  English Translation: "He is terrible and famous; his judgment shall proceed of himself, and his dignity shall come out of himself."

Appendix B: Tokenization and Multi-Task Input Formats

The following example illustrates the multitask input formats and the subword fragmentation caused by normalization.

Note: The examples below display the raw internal string representations of the byte-level BPE tokens. As shown, while normalization is performed on the word level during preprocessing, the BPE tokenizer segments the text at the subword (token) level. This inherently leads to severe sequence length mismatch across views (e.g., expanding from 37 tokens to 68 tokens), visually illustrating the structural bottleneck discussed in Section 6.3.

```
[Sample: Sahidic Coptic]
--- STEP 1 & 2: Text & Normalization ---
Original : ⲱ ⲡ ⲡⲁⲣⲑⲉⲛⲟⲥ ⲉⲧ ⲛϩⲟⲧ ⲁ ⲥ ϣⲱⲡⲉ ⲇⲉ ϩⲙ ⲡⲉ ⲟⲩⲟⲉⲓϣ ⲛ
ⲥⲟⲗⲟⲙⲱⲛ
English : O faithful virgin. Now it came to pass in
the time of Solomon
LATIN : oo p parthenos et nhot a s shoope de hm pe
ouoejsh n solomoon
IPA : oː p partʰənos ət nhot a s ʃoːpə də hm pə uoiʃ n solomoːn

--- STEP 3: Tokens & Model Inputs ---
 Original View (Length: 37 tokens)
Tokens : ['[CLS]', '_', '<sah>', '_â²±', '_â²¡',
'_â²¡â²ǵâ²£â²ɟjâ²īâ²Ł̂â²Łâ²¥',
        '_â²īâ²Ⴝ', '_â²Ł̈©â²Łâ²Ⴝ', '_â²ǵ', '_â²¥',
'_Ɪ̈£â²±â²¡â²ī', '_â²īâ²ī',
        '_Ɪ̈©â²Ł̣', '_â²¡â²ī', '_â²Łâ²©â²Łâ²īâ²ɟÏ£',
'_â²Ł̂', '_â²¥â²Łâ²Ĺâ²Łâ²Ł̣â²±â²Ł̂',
        '_', '[SEP]', '_', '<eng>', '_O',
'_faithful', '_virgin', '.', '_Now',
        '_it', '_came', '_to', '_pass', '_in',
'_the', '_time', '_of',
        '_Solomon', '_', '[SEP]']
IDs : [2, 154, 8, 787, 233, 5949, 379, 23686, 256,
278, 1063, 520,
        718, 338, 2916, 208, 11756, 154, 3, 154, 9,
500, 7311, 6282,
        23, 5016, 364, 836, 257, 1164, 263, 198,
1097, 215, 11443,
        154, 3]

 LATIN View (Length: 48 tokens)
Tokens : ['[CLS]', '_', '<sah>', '_o', 'o', '_p',
'_part', 'hen', 'os', '_et',
        '_nh', 'ot', '_a', '_s', '_sho', 'ope',
'_de', '_hm', '_pe', '_o',
        'u', 'oe', 'j', 'sh', '_n', '_sol', 'om',
'oon', '_', '[SEP]', '_',
        '<eng>', '_O', '_faithful', '_virgin', '.',
'_Now', '_it', '_came',
        '_to', '_pass', '_in', '_the', '_time',
'_of', '_Solomon', '_',
        '[SEP]']
IDs : [2, 154, 8, 202, 87, 223, 1527, 530, 529,
5257, 2597, 295, 203,
        204, 17228, 2680, 660, 10372, 758, 202, 93,
2828, 82, 1229,
        207, 3167, 288, 3321, 154, 3, 154, 9, 500,
7311, 6282, 23,
        5016, 364, 836, 257, 1164, 263, 198, 1097,
215, 11443, 154, 3]

 IPA View (Length: 68 tokens)
Tokens : ['[CLS]', '_', '<sah>', '_o', '[UNK]', 'Ɯ',
'_p', '_p', 'arth',
        '[UNK]', 'Ł̣', 'n', 'os', '_', '[UNK]', 'Ł̣',
't', '_nh', 'ot',
        '_a', '_s', '_', 'Ê', 'ĥ', 'o', '[UNK]',
'Ɯ', 'p', '[UNK]', 'Ł̣',
        '_d', '[UNK]', 'Ł̣', '_hm', '_p', '[UNK]',
'Ł̣', '_u', 'oi', 'Ê',
        'ĥ', '_n', '_sol', 'om', 'o', '[UNK]', 'Ɯ',
'n', '_', '[SEP]',
        '_', '<eng>', '_O', '_faithful', '_virgin',
'.', '_Now', '_it',
        '_came', '_to', '_pass', '_in', '_the',
'_time', '_of',
        '_Solomon', '_', '[SEP]']
IDs : [2, 154, 8, 202, 1, 171, 223, 223, 4397, 1,
180, 86, 529, 154,
        1, 180, 92, 2597, 295, 203, 204, 154, 137,
158, 87, 1, 171,
        88, 1, 180, 252, 1, 180, 10372, 223, 1,
180, 354, 4773, 137,
        158, 207, 3167, 288, 87, 1, 171, 86, 154,
3, 154, 9, 500,
        7311, 6282, 23, 5016, 364, 836, 257, 1164,
263, 198, 1097,
        215, 11443, 154, 3]

--- STEP 4: Multi-Task Formats ---
  Task (MLM & TLM)
MLM Input:
[CLS] <sah> ⲱ [MASK] [MASK] ⲉⲧ ⲛϩⲟⲧ ⲁ ⲥ [MASK] ⲇⲉ ϩⲙ ⲡⲉ
ⲟⲩⲟⲉⲓϣ ⲛ [MASK] [SEP]

TLM Input:
[CLS] <sah> ⲱ [MASK] [MASK] ⲉⲧ ⲛϩⲟⲧ ⲁ ⲥ [MASK] ⲇⲉ ϩⲙ ⲡⲉ
ⲟⲩⲟⲉⲓϣ ⲛ [MASK] [SEP] <eng> [MASK] faithful virgin.
[MASK] it [MASK] to pass in the time of Solomon
[SEP]
```

Appendix C: Examples of Intra-Egyptian Cognate Pairs

To evaluate semantic alignment without relying on English, we produced an intra-Egyptian cognate dataset, which is based on TLA's summarization of cognates. Below is a subset of the JSON-formatted data demonstrating the structural format of our evaluation pairs.

Dataset Construction Notes:

1. **Pairwise Mappings:** Cognate pairs are extracted in a strictly pairwise format (e.g., Demotic-to-Sahidic, Sahidic-to-Bohairic). This is because not all historical concepts survived across all four language stages (e.g., Concept `0014` is only attested in Demotic and Sahidic).

2. **Spelling Variants:** A single etymological concept (e.g., Concept `0005`) often generates numerous paired entries. This occurs because the same word may have multiple attested spelling variants within a single language stage. All possible cross-stage cognate pairs are included to ensure robust evaluation.


```
{"lang1": "demotic", "word1": "ꜥbyqe",
"lang2": "sahidic", "word2": "ⲁϥⲟⲕ",
"source_concept_id": "0005"}
{"lang1": "demotic", "word1": "ꜥbyqe",
"lang2": "sahidic", "word2": "ⲁⲃⲱⲕ",
"source_concept_id": "0005"}
{"lang1": "demotic", "word1": "ꜥbq",
"lang2": "sahidic", "word2": "ⲁⲃⲟⲟⲕⲉ",
"source_concept_id": "0005"}
{"lang1": "demotic", "word1": "ꜥbqy",
"lang2": "sahidic", "word2": "ⲁⲃⲟⲕ",
"source_concept_id": "0005"}
{"lang1": "demotic", "word1": "ꜥbyqe",
"lang2": "bohairic", "word2": "ⲁⲃⲟⲕⲓ",
"source_concept_id": "0005"}
{"lang1": "demotic", "word1": "ꜥbq",
"lang2": "bohairic", "word2": "ⲁⲃⲟⲕ",
"source_concept_id": "0005"}
{"lang1": "sahidic", "word1": "ⲁϥⲟⲕ",
"lang2": "bohairic", "word2": "ⲁⲃⲟⲕⲓ",
"source_concept_id": "0005"}
```

```
{"lang1": "sahidic", "word1": "ⲁⲃⲱⲕ",
"lang2": "bohairic", "word2": "ⲁⲃⲟⲕ",
"source_concept_id": "0005"}
{"lang1": "hieroglyphic", "word1": "ˁqy",
"lang2": "demotic", "word2": "ˁyq",
"source_concept_id": "0010"}
{"lang1": "hieroglyphic", "word1": "ˁqy",
"lang2": "sahidic", "word2": "ⲁϵⲓⲕ",
"source_concept_id": "0010"}
{"lang1": "demotic", "word1": "ˁlwɜ",
"lang2": "sahidic", "word2": "ⲁⲗ",
"source_concept_id": "0014"}
```


Appendix D: Evaluation Dataset Statistics

Table 2 details the number of valid word pairs (N) used for the semantic alignment evaluation across all language pairs and conditions. A word pair is considered valid if both constituent words appear in the held-out context pool and possess contextualized embeddings for cosine similarity computation. The numbers reflect the final filtered and deduplicated evaluation sets used in all reported experiments.

Evaluation Task	Category	Language Pair	Valid Pairs (N)
Intra-Egyptian	Cross-Branch (Heterograph)	Bohairic–Demotic	396
		Bohairic–Hieroglyphic	294
		Demotic–Sahidic	708
		Hieroglyphic–Sahidic	644
	Within-Branch (Heterograph)	Bohairic–Sahidic	318
		Demotic–Hieroglyphic	318
	Within-Branch (Homograph)	Bohairic–Sahidic	82
		Demotic–Hieroglyphic	181
Egyptian–English	Hieroglyphic–English	Seen	6,883
		Unseen	1,825
	Demotic–English	Seen	1,778
		Unseen	468
	Sahidic–English	Seen	1,967
		Unseen	355
	Bohairic–English	Seen	1,598
		Unseen	354

Table 2: Number of valid word pairs (N) utilized for each specific evaluation group.