Can AI Read Your Hieroglyphs? Evaluating LLM Performance across Ancient Egyptian Script Traditions, Transliteration Practices, and Low-Resource Ancient Languages

Eva Maria Hemauer
HU Berlin/LMU Munich

Generative large language models (LLMs) seem to be well-suited for philological micro-tasks such as translating short passages, identifying morphology, normalizing transliterations, and generating grammatical explanations. But how reliable is this assistance for genuinely low-resource ancient languages, where training data is sparse, scholarly conventions diverge, and graphical representation varies substantially? This paper presents a systematic evaluation of LLM performance on ancient Egyptian, with Hieroglyphic Luwian as a cross-linguistic comparator, and examines the conditions under which AI assistance can be responsibly integrated into philological workflows.

Three questions are addressed: (1) Do different diachronic stages of Egyptian show systematic differences in LLM performance? (2) How do transliteration conventions and graphical representations affect model outputs for identical underlying inputs? (3) How do different LLM architectures perform on low-resource ancient languages, and how do the results differ from those of a RAG instance?

We test four stages of Egyptian—Old, Middle, Late Egyptian, and Demotic—modeled as a stratified continuum of resource availability. Middle Egyptian, highly standardized and didactically dominant, represents the high-resource end. Demotic, despite a large documentary record, remains functionally low-resource: its cursive script, graphical variability, and incomplete digitization limit its computational visibility. Both Demotic and Hieroglyphic Luwian are characterized by extremely small expert communities and limited teaching infrastructure, constraining opportunities for guided acquisition and sustained feedback.

Each stage is evaluated using a compact suite of diagnostic sentences probing core morphology and syntax. Inputs are presented in multiple transliteration systems (e.g. Unicode with diacritics vs. simplified conventions) as well as hieroglyphic encodings ranging from normalized ASCII-style systems to more variable graphical forms. Transliteration is treated as a central variable rather than a neutral preprocessing step: it embeds prior philological interpretation, meaning LLMs process pre-interpreted data rather than primary textual evidence. It can also obscure distinctions among homographic strings—forms identical in romanization but belonging to different sign groups, lexemes, or morphemes—a challenge non trivial even in expert lemmatization.

We compare multiple LLMs across three task types: translation, grammatical analysis, and uncertainty statements. In addition, we evaluate THOTH AI (Miyagawa, University of Tsukuba), a retrieval-augmented system incorporating standard reference works such as Faulkner's Concise

Dictionary of Middle Egyptian, against base models under low resource conditions. Outputs are assessed using an expert rubric measuring accuracy, internal consistency, and calibration, with particular attention to representation-sensitive error types such as silently normalizing non-standard inputs and homograph misinterpretation.

Hieroglyphic Luwian serves as a cross-linguistic probe: an Indo-European logo-syllabic language with a limited corpus and restricted research tradition, allowing us to test whether LLM performance primarily tracks data availability rather than linguistic structure. To mirror the Egyptian setup, we construct a domain-grounded chatbot based on Payne's Hieroglyphic Luwian, enabling direct comparison between base and retrieval-augmented performance under low-resource conditions.

Beyond benchmarking technical performance, the paper examines how AI assistance shapes philological and educational practice—where such systems can productively support textual analysis and language acquisition, and where they risk fostering dependence or obscuring the interpretive complexity that philological training is meant to cultivate.

The paper contributes (1) a representation-sensitive evaluation framework for LLM performance on ancient languages, (2) a comparative analysis of base and retrieval-augmented models across two writing systems, (3) an error taxonomy tailored to philological workflows, and (4) practical recommendations for the responsible integration of AI assistance into teaching and research.