Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-assisted Egyptian hieroglyphic reading in scholarly practice: workload, usability, and acceptance in a pilot feasibility study

Ekaterina Aleksandrova
National Research University Higher School of Economics
Maksim Golyadkin
National Research University Higher School of Economics
Innokentiy Humonen
AXXX
Ilya Makarov
National Research University Higher School of Economics
Alexandre Loktionov
al621@cam.ac.uk

University of Cambridge

Article

Keywords: Egyptology, hieroglyphic texts, digital humanities, human-AI interaction, technology acceptance, usability, feasibility study

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10432519/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AI-assisted Egyptian hieroglyphic reading in scholarly practice: workload, usability, and acceptance in a pilot feasibility study

Author One[1], Author Two[1], Author Three[1], Author Four[1], Author Five[1*]

[1]Details redacted for peer review.

Abstract

Artificial-intelligence tools for reading ancient Egyptian hieroglyphic texts are advancing rapidly, but how Egyptologists actually experience such tools in scholarly practice has not been studied. This paper reports an observational pilot feasibility study of AI-assisted hieroglyphic reading with PyThoth, a human-in-the-loop assistant covering sign recognition, transliteration, and translation. Nine participants - undergraduate Egyptology students and professional researchers - read fragments of the Middle Kingdom stela BM EA567 in AI-assisted and unaided conditions; task records were combined with adapted workload (NASA-TLX), usability (SUS/UMUX-Lite), and technology-acceptance (TAM) instruments and qualitative feedback. AI assistance was associated with attempts being carried to completion rather than with more accurate work where participants were working. Perceived workload, usability, and acceptance shifted in participant-specific, bidirectional ways, descriptively aligned with the gap between pre-task expectation and direct experience; perceived usefulness and perceived ease of use diverged. The paper reports protocol-level findings for the design of future studies of AI-assisted philological work, in which partial and distributed participation should be anticipated rather than treated as attrition.

Keywords: Egyptology, hieroglyphic texts, digital humanities, human-AI interaction, technology acceptance, usability, feasibility study

1 Introduction

Recent advances in computer vision and natural language processing have produced increasingly sophisticated approaches to optical character recognition (OCR), which can now be applied to recognition, transliteration, and machine translation of hieroglyphic inscriptions (Golyadkin et al. 2025a). Ancient Egyptian hieroglyphic script presents unique challenges for automatic computer analysis: multidirectional reading, sign and spelling variants, damaged surfaces, and a writing system that encodes both phonetic and semantic information through a working repertoire on the order of a thousand sign forms.[1] Recent neural network approaches have achieved promising results on isolated tasks, namely sign classification (Franken and van Gemert 2013; Barucci et al. 2021), sequence-to-sequence transliteration, and token-level translation (De Cao et al. 2024), but the question of how such systems perform when integrated into authentic scholarly practice has not been studied yet. More critically, the epistemological implications of delegating portions of the interpretive process to algorithmic assistants remain under-theorized. This risk is not merely hypothetical: in a large field experiment, professionals using AI on a task deliberately designed to fall outside its "jagged frontier" were 19 percentage points less likely to produce correct solutions; in the same study, AI-supported participants produced recommendations that were "more coherent and persuasive", underscoring that AI "improves presentation and argumentation even when the underlying analysis is flawed" (Dell'Acqua et al. 2023). It is safe to say that fundamental questions about practical integration of technology into research and pedagogical workflows remain largely unexamined, while empirical evidence on how these tools affect the cognitive and interpretive work of Egyptologists is scarce if present.

The gap between technological capability and disciplinary integration is particularly important in educational contexts. Hieroglyphic literacy develops through intensive engagement with primary sources: learning to recognizes sign variants, parse grammatical structures, and navigate the relationship between standardized transliteration and semantic interpretation. AI assistance may affect not only speed and accuracy, but also how cognitive effort is distributed across these stages - namely, how hard a student has to try in their efforts to transliterate, translate, and parse a text. Whether AI-assisted workflows affect this developmental trajectory (by scaffolding learning or introducing dependencies inhibiting skill acquisition) is an open empirical question with significant pedagogical implications. This matters especially in educational settings, where repeated and sustained engagement with transliteration, translation and parsing is how hieroglyphic literacy is formed.

Following Vygotskian theory concerning the Zones of Proximal Development (Eun 2019), learning should be most effective if it occurs in the gap between what the student can do unaided and where extensive guidance is required: in an Egyptology setting, this would be working with text that a student is generally capable of handling by themselves, but with some limited assistance. Conventionally, such assistance would come from a course instructor, but a complementary support might be an AI tool assisting students in their independent work, before the eventual product is submitted to an instructor for feedback. At the same time, the use of such an AI tool might also help to reduce the overall workload of the student, allowing greater

attention to be paid to more intricate points of epigraphy or harder grammatical constructions instead of lengthy but relatively simple parts of the text. In the present study, this concept of workload is treated as an overarching interpretive dimension, encompassing student performance, usability of the interface, time taken to complete the task, and longer-term outcomes of adopting the technology (e.g. a change in the student's confidence level when working with Egyptian texts, or even a change in attitude towards Egyptology more broadly).

The present study is based around an observational pilot of AI-assisted hieroglyphic text interpretation framed as a human-in-the-loop workflow rather than as autonomous machine reading. In the AI-assisted condition, the system provides sequential suggestions for sign recognition, transliteration, and translation, which participants are expected to verify and correct at each stage. Importantly, it does not seek to replace the Egyptologist (or student), who retains full operational control of the final translation and can override any AI suggestions which contravene their professional judgment. The study design reflects the practical reality of current Egyptological AI tools: their scholarly value does not depend only on the quality of the translations produced with their help, but also on how easy it is for scholars to interact with them, how burdensome the process of making corrections is, and the extent to which the eventual outputs meet established disciplinary expectations.

Accordingly, the present article examines the feasibility of such an experimental protocol in a highly specialized humanities domain and explores how workload, usability, acceptance, and qualitative experience interact when AI support is introduced into Egyptian hieroglyphic reading. This is contrasted with control tasks carried out using a conventional, AI-free approach.

Our research questions are:

- **Feasibility and protocol effects**: How does a manual (AI-free) reading protocol compare to an AI-assisted one when tested in a small, distributed expert-domain sample, and what sources of protocol-induced burden emerge?
- **Workload and usability**: How do perceived workload and usability differ with and without AI assistance in the realized sample, and how do these experiences pattern with text structure, interface design, and prior expectations?
- **Acceptance and disciplinary fit**: What patterns of technology acceptance, professional hesitation, and pedagogical concern emerge when specialist readers (Egyptology students and staff) integrate AI support into established interpretive workflows?

This pilot study makes four contributions. First, it provides an exploratory observational comparison of manual and AI-assisted hieroglyphic reading workflows using authentic Ancient Egyptian materials. Second, it combines task records with workload, usability, acceptance, and qualitative feedback in order to examine AI integration as a human-factors problem rather than merely as a model benchmark. Third, it offers preliminary evidence that the perceived value of AI support is shaped not only by the overall capability of the tool, but also by ease of use, precision of individual steps within the workflow, and perhaps most significantly how its use is viewed by colleagues

in the discipline. Fourth, it develops and tests an experimental procedure and questionnaire framework that can be refined for future studies of AI-assisted philological work, both in Egyptology and beyond.

2 Methods

2.1 Study Design

We designed the study as a repeated-measures pilot with two conditions presented in intended counterbalanced order (Figure 3): manual work with conventional scholarly resources, and AI-assisted work using the same resources plus support from PyThoth, a new tool for aiding hieroglyphic translation (see below). The design was originally conceived as a within-subject comparison in which each participant would complete at least one transliteration and translation in each condition.

In practice, however, implementation in a small and geographically dispersed expert-domain population required a flexible hybrid format. For a variety of reasons, ranging from time constraints to technological issues and ideological concerns about AI use, not all participants completed every stage of every task. As a result, the data presented here combines complete protocols, partial trajectories, repeated participation, and questionnaire-only or feedback-only cases. The study should therefore be understood not as a fully standardized laboratory comparison, but as an observational pilot feasibility study with exploratory human-factors analysis.

2.2 Participants

We initially reached and received consent from 16 potential participants drawn from two populations: undergraduate Egyptology students and professional researchers across a plethora of career stages (from early-career to Professor Emeritus). Of these, eight unique participants completed the self-report questionnaire, with one researcher completing it twice for two separate text pairs. Nine participants submitted at least some task-form data, though with varying degrees of completeness[2]. Two additional invited experts contributed written feedback outside the experimental protocol: one trialled the system without completing the study, while the other declined participation on principled grounds regarding AI use in the discipline.

Participants' self-reported hieroglyphic reading proficiency ranged from intermediate to expert. Given the small and highly specialized population of active hieroglyphic readers, participation was voluntary, uncompensated, and often distributed across multiple settings. These conditions resulted in a heterogeneous but ecologically informative dataset characteristic of expert-domain pilot research. The realized composition is reported in Section 2.6 and summarized in Table 1.

Two co-authors of the present paper also served as participants. To preserve experimental naïveté, the text-selection procedure (Section 2.3) was structured so that these co-authors contributed only metadata-based evaluation at Stages 1–2 and did not read the candidate texts.

The study was approved by the authors' institutional review board. All participants provided written informed consent.

2.3 Text Selection

We selected experimental texts through a four-stage procedure designed to ensure validity while preventing contamination from prior exposure. Since two Egyptologist co-authors also participated in the experiment, the procedure was structured so that they could contribute domain expertise to selection criteria without reading the candidate texts, thereby preserving experimental naïveté.

Stage 1: Construction of lists of potential candidate texts. We compiled a list of 14 museum-held Middle Egyptian inscriptions from four collections (British Museum, Metropolitan Museum, Louvre, The Royal Museums of Art and History) with: (a) high-quality online photographs, (b) reusable licenses permitting academic reproduction (CC BY-NC-SA 4.0, Open Access, or equivalent), (c) dating and genre comparable to Advanced Middle Egyptian undergraduate coursework. For each candidate text we recorded metadata: museum, inventory number, license type, approximate dating, text type, text-to-image ratio, extent (lines/columns), preservation score (0–10), and bibliographic references to published translations.

Stage 2: Metadata-based screening (blind to content). Egyptologist team members evaluated the candidate texts using *only* metadata (period, text type, length, preservation, publication history) to estimate difficulty level by analogy with familiar curriculum texts. We explicitly excluded: (a) texts present in the training data of the sign-recognition component (3 of 14 candidates; absence from that training set was a selection criterion), (b) texts familiar to participating students or researchers from their own recent work. This "blind" screening ensures that Egyptologist co-authors remain eligible participants while guaranteeing tasks represent authentic yet novel material.

Stage 3: Content profiling (non-Egyptologist review). For shortlisted objects, non-Egyptologist team members used published translation (Lichtheim 1988) to perform qualitative content characterization: identifying formulaic sections (*ḥtp-dj-nsw*, adoration formulae), narrative passages, long strings of titles, and their approximate extent within each text. The goal was to locate optimal sub-passages of approximately five lines with coherent thematic structure and appropriate phrase boundaries, avoiding heavily formulaic opening sequences.

Stage 4: Final selection and segmentation. Combining ease of legal and digital access, formal properties of the artefact, content profiling, and estimated difficulty of the text, we selected British Museum Stela EA567 (CC BY-NC-SA 4.0, Middle Kingdom, ca. 1850 BCE, 27 lines, preservation 9/10). This text was additionally favoured because an outline copy was available - a format more suitable for the current PyThoth sign recognition pipeline. The text was divided into four consecutive fragments by a volunteer student of the authors' institution's Egyptology program who did not participate in the experiment. Text fragment boundaries were reviewed and adjusted by non-Egyptologist co-authors to ensure balanced length and avoid splitting syntactic units (the non-Egyptologist co-authors consulted already existing translations to enable this). Fragments were grouped into two pairs for counterbalanced assignment: Pair A (fragments 1 and 3) and Pair B (fragments 2 and 4).

2.4 AI System: PyThoth

In the AI-assisted condition, we used PyThoth, an academic assistant for the study of ancient Egyptian texts. It combines built-in reference materials (hieroglyphic palette) with a deep learning pipeline supporting interactive model calls. PyThoth supports close reading of the text and its digitization, enabling its further inclusion in text corpora used for computational linguistics, information retrieval, and state-of-the-art NLP methods. In prior work, the application was introduced in Golyadkin et al. (2025c), and a formative study (Humonen et al. 2026) suggested that it may improve the quality of transliterations and translations produced with its support. However, little is known about how Egyptologists themselves experience working with such a system in practice. In this paper, we address this gap.

PyThoth is organised as a web-based interface connected to a backend that generates model suggestions. It implements an OCR-transliteration-translation pipeline that mirrors the manual workflow of Egyptologists working with texts. The division of labor between the user and the system is intentionally structured around expert oversight: rather than relying on a fully automatic end-to-end result, users can inspect and refine intermediate outputs before invoking the next step of the pipeline. This design reflects the current limitations of deep learning for hieroglyphic texts (Golyadkin et al. 2025b), where accumulated errors can substantially reduce the quality of the final result.

The interface is designed to support both close reading of the source material and the creation of structured digital output (Figure 1). Users can inspect the hieroglyphic image, consult reference materials (hieroglyphic palette), manually enter or revise signs, and invoke AI assistance at different stages of the workflow. The system also supports manual lookup in a hieroglyph library and allows users to refine intermediate results, helping them maintain control over the interpretation process while using the tool for targeted support on difficult passages.

On the backend, PyThoth uses a multi-step pipeline that follows the conventional scholarly workflow (Figure 2). First, an OCR module detects and recognizess hieroglyphic signs and maps them to Gardiner codes. Next, a transliteration module converts these codes into transliterated text, and a translation module produces an English rendering. The OCR component follows the automatic text interpretation pipeline introduced in Golyadkin et al. (2025a) and is trained on annotations from the MuMMy dataset (Golyadkin et al. 2025b). The transliteration and translation components are based on models trained on the TLA-derived dataset (Richter et al. 2018) described in Humonen et al. (2025).

2.5 Pilot Self-Report Measures and Qualitative Feedback

The study combined automatically logged task metrics with adapted self-report instruments (provided in full as Supplementary Data S1) and open-ended feedback. Since the sample was small, heterogeneous, and typical of expert-domain humanities research, these measures were used descriptively rather than for local psychometric validation or strong inferential claims. Their function in the present study was to characterise perceived workload, usability, and acceptance profiles, to trace directional shifts between

expectation and experience, and to triangulate numerical ratings with completion patterns and qualitative comments.

2.5.1 Pre-experiment questionnaire

The pre-experiment questionnaire combined background variables with baseline attitude measures. Background items recorded self-reported hieroglyphic reading proficiency (1-5), frequency of work with hieroglyphic texts, prior familiarity with OCR/AI tools for hieroglyphic reading, frequency of LLM use in research, primary working device, preferred reference resources, and self-assessed translation speed. Participants were also invited to provide a brief open expectation regarding AI assistance.

In addition, the pre-experiment questionnaire included 5-point Likert items adapted from Technology Acceptance Model (TAM) research (Ibrahim et al. 2025) in order to capture initial participant expectations regarding perceived ease of use, perceived usefulness, and subjective norm. These were supplemented with domain-specific items on anticipated learning value, self-efficacy, and the perceived disciplinary impact of AI-assisted hieroglyphic reading (Ueno et al. 2022; Li et al. 2024). In the present pilot study, these items were used to establish baseline attitude profiles rather than to estimate stable latent constructs.

2.5.2 Post-experiment questionnaires

Post-experiment self-reports combined workload assessment, interface evaluation, post-experience acceptance measures, and open-ended feedback.

Perceived workload (NASA-TLX).

Perceived task load was assessed after each condition using an adapted integer-response version of the NASA Task Load Index (NASA-TLX) (Devos et al. 2020; Hutson et al. 2024), scored from 1 to 20 on six dimensions:

1. Mental demand
2. Physical demand
3. Temporal demand
4. Performance
5. Effort
6. Frustration

To reduce procedural burden, we used the unweighted Raw TLX variant. When composite scores were calculated, the performance item was reverse-coded so that higher values consistently indicated higher perceived workload. In view of the small sample, NASA-TLX results were interpreted descriptively at both the global and subscale level. In particular, the frustration subscale was treated as informative for interface- and procedure-related burden in addition to task-related difficulty.

Usability of the AI-assisted workflow.

Perceived usability was assessed only for the AI-assisted condition, since the manual condition did not involve an experimental interface. This block combined:

1. a 2-item UMUX-Lite (Lewis et al. 2013) component (7-point response format), targeting perceived capability fit and ease of use.
2. a 9-item all-positive SUS-type component (5-point response format) (Bangor et al. 2008; Lewis 2018b), targeting clarity, consistency, learnability, confidence, and willingness to use the system.

Because these two components employed different response scales (Lewis 2018a), they were analysed separately and normalised to a common 0–100 range only for rough cross-measure comparison. Given the pilot character of the study, interpretation focused on relative profiles, convergences with open-ended feedback, and broad patterns of favourable, mixed, or unfavourable usability rather than on strict benchmark classification.

Post-experience technology acceptance.

After task completion, participants completed a post-experiment acceptance block using 5-point Likert items adapted from TAM constructs. These items assessed:

1. perceived usefulness,
2. perceived ease of use,
3. attitude toward use,
4. behavioural intention,
5. selected domain-specific perceptions of professional relevance and learning.

The post-experiment acceptance block was used in comparison with the pre-experiment baseline in order to describe expectation-to-experience shifts in participants' attitudes toward AI-assisted reading of ancient Egyptian hieroglyphic texts. In the present study, these shifts were interpreted directionally (e.g., more positive, less positive, mixed, or unchanged), not as evidence of stable population-level parameters.

Open-ended feedback.

All participants were invited to answer three open-ended questions:

1. Which aspects of the AI suggestions were most useful?
2. What hindered, confused, or slowed you down during the process?
3. What errors or failures did you notice in the AI output?

These responses were used to contextualize the numerical self-reports and to identify sources of extraneous burden not fully captured by rating scales, including interface friction, procedural complexity, and discipline-specific concerns about reliance on AI.

2.6 Procedure: Planned Design and Actual Implementation

The experiment was originally designed as a structured cross-over protocol in which each participant would complete at least two text fragments, one in the AI-assisted

condition and one in the manual condition, using a shared participant ID across the questionnaire, task submission form, and the PyThoth interface. For participants willing to continue, the design also allowed completion of all four fragments in a recommended counterbalanced order. The initial materials therefore assumed a relatively coherent session structure with pre-task instructions, task execution, intermediate questionnaires after each condition, and a final post-experiment questionnaire.

In practice, however, the study had to be adapted to the realities of a small, geographically dispersed expert population. Since participants were based in different countries and often had limited availability, the experiment was conducted in a flexible hybrid format combining an initial in-person session with subsequent remote and self-paced participation. Only four students took part simultaneously in the on-site phase, and even across an approximately 90-minute in-person session they completed only one text each before later returning to the experiment individually. As a result, the actual implementation diverged substantially from the original single-session plan.

Rather than constituting a uniform laboratory session, participation took the form of distributed experimental engagement over variable time spans. Some participants completed a single pair of texts, some completed additional fragments later, and only two participants completed the full four-text protocol. The temporal structure of participation also varied substantially: in the final dataset, intervals between submissions ranged from minutes within a single sitting to delays of several days or even weeks between fragments. Partial completion, non-completion, and qualitative-only participation were therefore treated not as mere data loss, but rather as part of the empirical picture of how such a protocol functions in a specialized humanities domain and how comprehensively (if at all) users respond to it.

For this reason, the study is best understood not as a strictly synchronized lab experiment, but as a structured pilot with a common task design and common instruments, implemented under participant-paced conditions. The original counterbalancing logic remained in the design, but actual condition order was reconstructed from submission timestamps and, where necessary, was treated analytically as an observed rather than a fully enforced variable.

2.6.1 Task instructions

All participants received the same core instruction:

"Your task is to produce a transliteration and translation of the hieroglyphic text passage shown in the image. Work as you normally would, using whatever resources you find helpful. We are not expecting a perfect result. We are interested in your working process. When you consider your work complete, submit the result through the task form."

In the AI-assisted condition, participants were additionally instructed to use PyThoth across three stages (sign recognition, transliteration, and translation) and to verify, correct, accept, or reject system suggestions at each stage. The instructional materials specified that the same participant ID should be used throughout the questionnaire, response forms, and the PyThoth project workspace, so that data from different components could later be matched.

2.6.2 Data logging and protocol adaptation

Since participation frequently extended beyond a single observed session, the principal record of task progress was the timestamp structure of submitted forms rather than researcher-side session logging. For each task-form entry we reconstructed participant ID, text fragment number, experimental condition, submission timestamp, relative ordering of conditions, and completion status (complete, partial, abandoned, or qualitative feedback-only). Partial datasets were retained when they contained interpretable task output, timestamp information, or qualitative feedback. In a small-scale pilot with domain experts, incomplete trajectories are themselves informative about the feasibility of the protocol and about the interaction and form burden it imposes. They are therefore treated as findings in their own right, rather than as data loss (see Section 3.1).

2.7 Measures and Data Analysis

The dataset combined timestamp-based task records, adapted self-report measures, and qualitative feedback. Since participation often extended across multiple sittings, task order and progression were reconstructed from timestamped submissions rather than from continuous researcher observation. Partial datasets were retained when they contained interpretable timing, task, or self-report information.

2.7.1 Measures

The current article reports the following measures:

1. **Task timing and completion patterns:** text fragment number, condition, submission timestamp, and completion status (complete, partial, abandoned, or qualitative-only), used to reconstruct task order, spacing between tasks, and approximate completion times.
2. **Perceived workload:** Raw NASA-TLX ratings (1–20) for six dimensions (mental demand, physical demand, temporal demand, performance, effort, frustration). When composite scores were calculated, the performance item was reverse-coded.
3. **Usability of the AI-assisted workflow:** UMUX-Lite (2 items, 7-point scale) and an all-positive SUS-type block (9 items, 5-point scale), administered only for the AI-assisted condition. Since these components used different response formats, they were analysed separately and normalised to a 0–100 range only for rough comparison.
4. **Technology acceptance:** adapted pre-experiment and post-experiment TAM-based items assessing perceived usefulness, perceived ease of use, attitude toward use, behavioural intention, and related domain-specific perceptions.
5. **Qualitative feedback:** open-ended responses on useful aspects of AI support, sources of hindrance or confusion, and perceived system errors.

2.7.2 Metric calculation

Adapted Raw NASA-TLX (RTLX; unweighted, 1–20 anchors).

Six subscales rated on an integer scale of 1–20 (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration). The Performance item was reverse-coded ($x_{\text{rev}} = 21 - x$) to align directionality across items, so that higher values consistently indicate higher perceived workload. RTLX is the arithmetic mean of the six reverse-aligned scores, bounded 1–20 (Devos et al. 2020; Hutson et al. 2024). For within-subject comparison:

$$\Delta\text{TLX}_i = \text{RTLX}_{\text{AI},i} - \text{RTLX}_{\text{Manual},i} \tag{1}$$

System Usability Scale (all-positive, 9-item).

Nine items rated on a 5-point Likert scale, all phrased positively. Following the all-positive variant described by Lewis (2018a):

$$\text{SUS}_9 = \frac{100}{4 \cdot 9} \sum_{i=1}^{9} (x_i - 1), \quad \text{range } [0, 100]. \tag{2}$$

UMUX-Lite (regression-adjusted).

Two items rated on a 7-point Likert scale. The raw UMUX-Lite is computed as $\text{UMUX} = (x_1 + x_2 - 2) \cdot (100/12)$, then adjusted to the SUS scale via the regression formula of Lewis et al. (2013):

$$\text{UMUX}_r = 0.65 \cdot \text{UMUX} + 22.9, \quad \text{range } [22.9, 87.9]. \tag{3}$$

TAM-informed subscales (adapted, three items per construct).

Pre- and post-experiment TAM items were aggregated as the arithmetic mean within subscale. Pre-TAM included Perceived Ease of Use (PEU, 3 items), Perceived Usefulness (PU, 3 items), and Subjective Norm (SN, 5 items). Post-TAM included PU (3 items), PEU (3 items), Attitude Toward Use (ATU, 2 items), and Behavioural Intention (BI, 3 items). All items used a 5-point Likert scale. Expectation-to-experience shifts were computed as $\Delta\text{PU} = \text{PU}_{\text{post}} - \text{PU}_{\text{pre}}$ and $\Delta\text{PEU} = \text{PEU}_{\text{post}} - \text{PEU}_{\text{pre}}$.

Since the usability block combined a 7-point component (UMUX-Lite) with a 5-point component (SUS), the two were processed separately and reported on a common 0–100 scale only for rough comparison, following Lewis (2018b). Composite scores and deltas were computed from full-precision means and rounded only for display.

Automatic metrics conventionally used for such pipelines - CER (Levenshtein 1966) for sign recognition, BLEU (Papineni et al. 2002), chrF++ (Popović 2017) and BLEURT (Sellam et al. 2020) for translation - evaluate a system against a reference. Our question is different: not how the system performs alone, but what the reader does with it. Two properties of this material make the conventional metrics unsuitable here. They operate on surface strings and would score the two Egyptological encoding conventions as substantive disagreement; and they presuppose a single reference, whereas the reference standard here shares its provenance with the model's

training data. We therefore normalise notation before comparison, use an independent published translation (Lichtheim 1988) alongside the TLA standard, and compare participants with one another as well as with the standard - a comparison that has no counterpart in reference-based evaluation, and which is where the adoption of system output becomes visible.

Transliteration responses were compared with the TLA standard after normalisation to a common consonantal skeleton: encoding system (Unicode vs Manuel de Codage), morpheme and clitic boundaries, brackets, plural marking, digits and capitalisation were neutralised before comparison, as far as the two notations allow, so that the measure reflects the reading rather than the notation. One residue of this is worth stating explicitly, because it is a general property of scoring across transliteration conventions rather than of this corpus: the two notations overlap in their Latin characters - a capital *T* is both the capitalised form of *t* and the Manuel de Codage code for *t̲* - so in a token written entirely in ASCII characters capitalisation cannot be separated from phonemic value by automatic means. Transliteration agreement is therefore reported under one specific reconciliation of the two systems; reasonable alternative reconciliations shift the reported percentages slightly without altering the direction of the contrast or the conclusions drawn from it. Sign identification is unaffected, since sign values are resolved through the official JSesh correspondence table rather than by letter case. Tokens were aligned by Needleman-Wunsch with a Levenshtein substitution cost and scored as exact matches; we report exact agreement only. We do not report a graded "near-match" category: on the short consonantal skeletons typical of this material, small edit distances do not reliably distinguish a variant spelling of one lexeme from a different lexeme (*m* and *n*, or *wr* and *ḥr*, differ by one character but are distinct words), and Gardiner sign identifiers are categorical, so that string distance between them carries no interpretable meaning. Sign responses were parsed to individual sign identifiers before comparison, including responses using JSesh layout operators; sign values entered as phonetic Manuel de Codage codes - a standard input convention in JSesh - were resolved to their canonical Gardiner identifiers through the official JSesh correspondence table, so that the measure reflects sign identification rather than input convention. Normalisation matters for comparability rather than for scoring alone: two expert participants who agree on a substantial majority of readings share only a small fraction of their raw strings, because they write in different conventions.

3 Results

Results are reported at a descriptive and case-comparative level appropriate to a pilot feasibility study. Since the realized body of data combines complete protocols, partial trajectories, repeated participation, questionnaire-only cases, and external expert feedback, the emphasis is placed on participation structure, perceived workload, usability, acceptance, and post-hoc task complexity rather than on inferential comparison of task accuracy.

3.1 Participation trajectories and completeness of the study record

Table 1 summarizes the realized participation. Two participants (P11, P13) completed all four fragments. Three (P2, P3, P4) completed a minimal pair of one AI and one manual fragment. One (P12) completed the AI condition but abandoned the manual one during the Gardiner coding stage. One (P1) completed the AI condition only. One (P5) returned a full questionnaire without task submissions. Two additional submitters (T1, T2) contributed task data without linked questionnaires. Completion was distributed and participant-paced: inter-submission intervals ranged from minutes within a single sitting to several weeks.

Counterbalancing assignment was followed by five of seven numerically identified participants. Two (P12, P13) completed conditions in the reverse order. P12 explicitly attributed motivational pressure to beginning with AI. P13's reversal may reflect a deliberate choice by a participant already familiar with the system. One form-level error was corrected: P4 marked both entries "Fragment 2". The AI-assisted entry corresponds to Fragment 1 on the basis of sign-sequence and translation content (Table 1, footnote a).

3.2 Task timing and protocol-level burden

Task durations varied substantially across text fragments and participants (Supplementary Table S2). Within Fragment 1 the Gardiner-coding stage dominated the workflow (Mdn = 23.0 min against 12.0 for transliteration), while its translation stage was the fastest of any fragment (Mdn = 3.0 min), reflecting its formulaic content. Fragment 2 showed the inverse pattern: shorter text volume but a translation stage more than twice as long (median 7.2 against 3.0 min). Fragment 4 yielded the highest total per-submission time (median 77 min) across only two non-outlier observations.

One participant (P12) abandoned the manual condition during the Gardiner-coding stage, explicitly reporting that this step was incompatible with their habitual workflow. This case indicates that part of the experienced burden derived from the procedural requirement to produce Gardiner codes, not from the text itself.

Taken together, the timing profile also carries a protocol-level reassurance: median total times for Fragments 1-3 fall within a narrow band (39-41 min per submission), even though the mix of conditions differs across the three, so the reading tasks posed a broadly comparable practical burden. The longer total for Fragment 4 (77 min) rests on two unaided submissions and cannot be separated from condition. We treat per-fragment times as a measure of protocol burden only; intrinsic text complexity is assessed against a fixed corpus baseline in Section 3.6.

Within-subject workload comparisons were available for seven observations (Table 2). The median ΔTLX of -0.83 and mean of -0.38 conceal strongly heterogeneous movement: two records showed sizeable reductions (ΔTLX = -9.0 and -1.5), while one showed a comparably large increase (ΔTLX = $+9.0$). Directionally, four of seven observations favoured the AI-assisted condition, two did not, and one was essentially unchanged. The AI-assisted workflow therefore was not associated with a

uniform workload reduction. Instead, within-session differences were heterogeneous and participant-specific.

The two most extreme cases are informative. P12's large reduction (−9.0) reflects abandonment of the manual condition at the Gardiner-coding stage, suggesting that the reduction measured workload against an interrupted baseline rather than against a comparable manual translation. P3's large increase (+9.0) indicates the opposite: the AI-assisted condition added rather than reduced demand, consistent with this participant's later usability rating ($SUS_9 = 16.67$). At least three components interact in the observed pattern: text difficulty, procedure-related extraneous load, and participant-specific interface familiarity.

3.3 Usability of the AI-assisted workflow

Usability scores for the AI-assisted workflow separated into two visibly distinct groups (Table 3). Three of nine records formed a distinctly low group (P1, P3, P5; $SUS_9 = 16.7$–22.2), while the remaining six clustered between 52 and 72. $UMUX_r$ values showed a similar pattern, although with a narrower range imposed by the regression adjustment. Because SUS_9 is a nine-item all-positive adaptation rather than the canonical ten-item instrument, conventional SUS cutoffs and adjective ratings do not transfer to it, and we intentionally refrain from classifying these scores against published norms; the two-group structure within this sample is the observation. The median SUS_9 of 63.89 and $UMUX_r$ of 55.40 obscure this structure and should not be read as a unimodal "moderate" rating.

Examining the two modes separately is more informative. The low-scoring participants (P1, P3, P5) share no single demographic feature, and all three showed large negative post-experience shifts (Section 3.5); two of the three (P1, P3) also entered with high expectations ($PU_{pre} = 3.67$ and 4.33), while P5's expectations were moderate ($PU_{pre} = 3.00$) (Section 3.5). The higher-scoring participants include both the naive users who converted upward (P12, P13) and the engaged expert (P11). This dissociation is embodied in the written feedback of one expert (Expert A), who was "impressed by the AI's ability to detect signs" and found it "much easier to edit something than to transcribe from scratch", while at the same time criticising the number of interface steps required to correct a single misidentified sign. Praise for the system's capability and criticism of its interaction cost thus coexisted in the same respondent's qualitative feedback. Expert B declined participation on principled grounds, articulating disciplinary concerns about AI-assisted reading. Favourable assessment of the underlying model and dissatisfaction with the implemented interaction therefore coexist in the present corpus.

3.4 Technology acceptance and attitude shift

Pre/post technology acceptance data were available for nine records, including two post-task measurements for P13 (Table 4). The sample-level pattern is a divergence between the two main TAM dimensions: mean PU was essentially unchanged ($\Delta PU = +0.04$), while mean PEU declined ($\Delta PEU = -0.44$). Post-task ATU ($M = 3.56$) and BI ($M = 3.15$) remained in the moderately positive range. Perceived usefulness

and perceived ease of use are distinct dimensions rather than a single acceptance construct (Ibrahim et al. 2025); that they diverge in response to concrete interaction experience is a within-session pattern in our own repeated-measures data, for which a cross-sectional adopter survey provides no comparison.

Aggregate means conceal substantial bidirectional movement: of the nine records, five showed negative ΔPU and four positive, and the same five-to-four split holds for ΔPEU - but two records (P2, P11) change sides between the two dimensions, reproducing the sample-level usefulness/ease dissociation at the individual level. These two side-changing shifts are also the smallest in the table (±0.33, a single item moving one scale point). The direction of the shift does not correlate straightforwardly with demographic variables but does align with the relation between pre-task expectation and the implemented interface. We elaborate this pattern in Section 3.5.

Separating the two components of task performance is more informative than a single accuracy figure (Figure 4). Every AI-assisted attempt was carried through all three layers (twelve of twelve), whereas five of the nine unaided attempts were. Incompleteness in the unaided condition was heterogeneous and layer-specific rather than a uniform tendency to break off: one attempt left the transliteration at 36% coverage; one supplied a near-complete transliteration but almost no sign layer; one completed the sign and transliteration layers (100% and 98%) and broke off only in translation; and one fell short across layers (53-55%). Once attempts carried through all layers are compared, the coverage difference between conditions disappears: sign coverage is 100.0% under AI assistance against 99.5% unaided, and transliteration coverage 96.5% against 97.5% - marginally higher unaided. Exact-match sign identification is effectively indistinguishable between conditions among completed attempts (83.4% under AI assistance against 85.1% unaided; cluster-bootstrap 95% CI for the difference −8.5 to +3.6 percentage points), and the paired within-participant contrast likewise includes zero (−3.8 pp, 95% CI −33.6 to +26.0); the unaided side of this comparison rests on five completed attempts from three participants, four of them contributed by the two co-author participants (Figures 5 and 6). Exact-match transliteration agreement remains higher under AI assistance among completed attempts (60.2% against 43.2%), but this is the layer most directly exposed to agreement with the editorial conventions on which the system was trained (see Limitations). AI assistance in this corpus is therefore associated with attempts being carried to completion rather than with demonstrably more accurate work where participants were working; given that participation was voluntary and uncompensated, we report this as an observed association and not as an effect of the tool.

The degree to which participants' responses converge on a common source can be read directly from the task record (Figure 7). In the AI-assisted condition, pairs of participants working on the same fragment share runs of up to 28 consecutive identical transliteration tokens - including spacing, plural marking and diacritics - and up to 44 consecutive identical words of translation; in the unaided condition the longest shared runs are 6 tokens and 5 words respectively. Four of the six AI-assisted participants on Fragment 1 open their translation with the same sentence carrying the same typographical error (*togehter*), which is unlikely to have arisen independently. In the transliteration layer, four of the six - an overlapping but not

identical set - segment one word into two (*sḏm* rendered as *sḏ m*), whereas the single unaided participant on the same fragment renders it as one word. Such convergence is not uniform: one participant produced a transliteration and translation that share no such runs, and one expert participant re-encoded his responses in his own notational convention throughout. Because the system's output was not itself recorded, these observations establish that the responses share a common textual source, consistent with adoption of the system's output, but they do not license claims about what the system proposed on any given trial or about whether a given participant inspected it.

An unanticipated observation concerns notation itself. Egyptological transliteration admits several encoding conventions; the TLA-derived models underlying the system emit Unicode transliteration, whereas participants working unaided used Manuel de Codage. In this corpus the proportion of Unicode-encoded tokens was zero in all nine unaided attempts, and ranged from 40% to 55% in ten of twelve AI-assisted attempts; the two exceptions are both attempts by a single expert participant who normalised his responses into his own convention. Where the same participant worked in both conditions, the shift is categorical rather than gradual (from pure Manuel de Codage unaided to 43-50% Unicode when assisted, in six of seven cases). Two further markers - uppercase plural marking and the inverted-breve diacritic - pattern with the Unicode rate rather than independently of it, as expected if all three are features of a single inherited notation. We report this as a descriptive property of the present corpus and not as a validated indicator of AI use: the threshold was chosen and evaluated on the same twenty-one attempts, fragment identity alone already classifies eighteen of them, and there are no within-participant comparisons holding fragment constant. Whether a notational trace of this kind generalises to other texts, participants and versions of the tool is an open question.

3.5 Exploratory trajectory profiles

The aggregate TAM summary (Section 3.4) conceals a heterogeneous pattern of individual movement. To characterise this heterogeneity, we examined per-record trajectories (nine records from eight respondents) of Perceived Ease of Use (PEU) and Perceived Usefulness (PU) between the pre- and post-experiment measurements (Table 4), jointly with demographic and prior-use variables from the background questionnaire and with the within-subject workload and usability indices (Tables 2, 3). Three broadly distinct profiles emerge. We describe them as within-session expectation-experience trajectories specific to this sample; they are not the adopter categories of diffusion research, which classify populations over calendar time (Ibrahim et al. 2025), and a sample of nine records from eight respondents could not support such classification.

Profile A: high prior expectations followed by disappointment

Three undergraduate participants (P1, P3, P4: self-reported proficiency 3-4/5, prior LLM use "sometimes" to "frequently") entered the study with high pre-task expectations ($PU_{pre} = 4.33$ for two of the three and 3.67 for P1, PEU_{pre}= 3.33-4.00). All three showed pronounced negative shifts in PEU ($\Delta PEU = -1.33$ to -2.33) and negative

shifts in PU (ΔPU = -0.67 to -1.00). Their usability ratings fell in the low-to-middle range (SUS_9 = 16.67-52.78). This pattern suggests that their baseline familiarity with conversational-AI tools might have set up an expectation the PyThoth interface did not meet: direct manipulation, Gardiner-code disambiguation, and multi-stage confirmation introduced interaction costs unfamiliar from their habitual LLM use. The term "experimentation phase" is borrowed descriptively from Ibrahim et al. (2025), who use it for initial enthusiasm revised on direct interaction; we do not map these participants onto adopter categories, which a within-session design cannot support.

Profile B: low prior expectations followed by upward revision

Two participants, P12 (no prior AI familiarity, never uses LLMs) and P13 (senior researcher, "has heard of" AI, never uses LLMs), entered with markedly lower or equivocal baseline expectations (PEU_{pre} = 1.00 for P13 round 1, = 3.00 for P12 and P13 round 2, PU_{pre} = 3.00-3.33). Both showed the largest positive shifts in the sample (ΔPU up to +2.00, ΔPEU up to +1.67) and the highest usability ratings (SUS_9 = 63.89-72.22, $UMUX_r$ up to 77.07). Notably, P12 produced these scores despite abandoning the manual condition partway through, indicating that the positive shift was attached to the AI-assisted workflow specifically, not to the experimental protocol overall. For both, the upward revision followed direct, task-embedded contact with the tool; we describe this as a within-session trajectory and do not map it onto adopter categories.

Profile C: calibrated expectations, stable engagement

One participant (P11, senior researcher, has tried AI previously, frequent LLM user) exhibited the most stable trajectory: moderate pre-task expectations (PEU_{pre} = 3.67, PU_{pre} = 3.00) and near-zero post-experience shifts (ΔPEU = +0.33, ΔPU = -0.33). P11 was also the only participant to complete the full four-fragment protocol in a single session (approximately 1.5 h), with $RTLX_{AI}$ = 10.5 and SUS_9 = 63.89. The combination of calibrated expectations, low extraneous load from the experimental framework, and sustained engagement made this the most stable trajectory in the sample. This stability, however, admits a narrower reading: P11 is a participant co-author who designed much of the protocol and had used the tool in earlier teaching projects, so the familiarity at work here extends beyond general AI experience (P11 reported both prior trial of AI-based tools for hieroglyphic work and frequent LLM use, the only such combination in the sample) to this specific instrument and task format. The calibrated-expectations pattern should therefore not be generalised from a single insider case.

A boundary case sets off what "calibrated" does and does not mean here: P2 reached a similar near-zero net shift from a high prior (PU_{pre} = 4.33, PEU_{pre} = 4.00, the level of Profile A), through the usefulness/ease dissociation that characterises the sample-level means, and with selective effort in the unaided condition (the Gardiner stage was passed over in three minutes at 4% sign coverage, while transliteration and translation were carried through). Stability of outcome is not calibration of expectations; we treat P2 as the boundary of this profile rather than a member.

Non-completion after engagement

One senior researcher (P5) fits none of the three profiles above. P5 submitted no task-form data, returned free-text NASA-TLX responses instead of numeric ratings, and returned the lowest usability profile in the sample ($SUS_9 = 16.67$, tied with P3, and the lowest $UMUX_r = 39.15$) with uniformly negative TAM shifts ($\Delta PEU = -2.00$, $\Delta PU = -1.67$, post-task $PU_{post} = 1.33$). P5 is better understood not as resistance to the technology but as a distinct mode of exit, which we interpret as a digital-workflow threshold: P5 reports having engaged with the system and carried out the work, yet no task-form data were submitted, and the NASA-TLX block was returned as free text rather than in the required format. The attribution to the workflow is our reading of this mismatch between reported engagement and received records, not P5's own account. This is distinct from the principled disciplinary objection expressed in the written expert feedback (Section 3.7), which came from a different contributor. Non-completion in this small sample in fact took several distinct forms - principled refusal, engagement outside the protocol, a hardware-access barrier, and a digital-workflow threshold - and we treat it as part of the feasibility finding rather than as simple attrition.

Interpretation

Taken together, these profiles suggest that in a specialist humanities context, neither age nor professional seniority straightforwardly predicts AI acceptance trajectory. As a post-hoc interpretive hypothesis, post-experience attitude in the present corpus patterns with the *gap* between pre-task expectation and direct experience: participants with calibrated or low baseline expectations moved upward, while those with high expectations drawn from general-purpose conversational AI moved downward. Part of this pattern is expected on arithmetical grounds alone - a change score is negatively coupled to its own baseline, and in these nine records baseline expectation correlates strongly with the shift ($r = -.59$) but only weakly with the post-experience level itself ($r = -.11$) - so the gap is a descriptive organising device, subject to regression to the mean, rather than a discriminating variable. This expectation–experience gap is a within-session pattern observed directly in our repeated-measures data, rather than the population-level dynamic of a cross-sectional adopter survey (Ibrahim et al. 2025). It also reinforces the usefulness/ease-of-use divergence reported in the main TAM analysis (Section 3.4): participants who converted upward did so primarily on the PU axis, while those who disengaged did so primarily on the PEU axis.

Because the sample is small, heterogeneous in role, and includes two co-author participants, these profiles are offered as hypotheses for future, better-powered work rather than as established findings. The trajectory analysis nonetheless indicates that a scale-up of this protocol should stratify recruitment by prior AI experience and collect personality or AI-mindset measures, as integrated in the extended TAM of Ibrahim et al. (2025), to allow proper cluster-level modelling.

3.6 Post-hoc estimation of task complexity

To characterise the intrinsic complexity of the four passages independently of the experiment, we scored each fragment against a fixed, reproducible corpus baseline rather than against participants' behaviour: the versioned TLA corpus of Earlier Egyptian (v18; 12,773 clauses), in which 93% of the stela's lemma types are attested. Each fragment was profiled on orthographic density (Gardiner signs per lexical word), lexical rarity (share of content lemmas with corpus frequency ≤ 3, computed over normalised lemma names), orthographic irregularity (variant or erroneous spellings flagged in the edition), and syntactic embedding (subordinate verb forms per finite verb). Supplementary Table S3 reports the profile.

The four passages differ along axes that no single dimension orders together. The two formulaic passages (Fragments 1 and 2) sit near the middle of the corpus frequency distribution (50th and 61st percentile of mean clause lemma frequency), whereas the two biographical passages fall lower (33rd and 37th). Fragment 3 is the hardest lexically and orthographically dense: 21% of its content lemmas are rare and two are absent from the corpus altogether (e.g. *hmhm.t*, "war cry"). Fragment 4 combines comparatively common vocabulary with the deepest syntactic embedding (4.0 subordinate forms per finite verb). Fragment 1 is orthographically the least regular but syntactically the simplest; Fragment 2 is the most name-dense, a litany of priestly titles and festival names that a novice must decode sign by sign but an experienced reader recognises as familiar units. Because the load imposed by a passage depends on the reader's existing schemata, the same fragment can be differentially demanding across participants - an instance of the expertise reversal effect, in which the value of support and the burden of complexity are relative to the reader's level rather than absolute (Kalyuga et al. 2003; Kalyuga 2007). Reading, on this evidence, is not one difficulty but several - sign decoding, lexical retrieval, and syntactic parsing vary independently across the fragments and are differently supported by an OCR-transliteration-translation pipeline - which provides a textual basis for the heterogeneous, task-dependent effect of AI assistance observed above. Completion times cannot serve this purpose in the present design: fragment and condition were confounded (Fragment 3 was attempted only with AI assistance, Fragment 4 only unaided), and time normalised by attempted volume conflates difficulty with completion.

3.7 Expert written feedback outside the experimental protocol

Two invited experts contributed written responses without completing the experimental protocol. Expert A engaged with PyThoth for approximately one hour and returned detailed observations: strong positive impressions of sign recognition, coupled with criticism of the number of interface steps needed to correct a single sign, and one case in which translation output appeared correct despite incorrect intermediate Gardiner codes. Expert A's observation - a correct translation despite an uncorrected sign misidentification (the string left as I9-S29-D21-D36 for *ḏsr*) - is instructive rather than anomalous. It suggests that partial graphemic information, combined with the strong textual priors of a formulaic passage, can be sufficient for the system to recover

a known translation without a correct intermediate analysis. This is the central epistemic risk of the workflow: a fluent, confident output can mask an incorrect path to it - the same presentation-accuracy gap documented at scale by Dell'Acqua et al. (2023), whose AI-supported participants produced recommendations that were "more coherent and persuasive" even where the underlying analysis was flawed.

Expert B declined the experimental protocol on principled grounds, articulating concerns that systematic AI assistance risks undermining the formation of hieroglyphic reading competence in students through progressive exposure to primary sources, and even strongly implying that this could have long-term detrimental implications for the survival of professional Egyptologists as a distinct occupation.

These responses are not included in the quantitative analysis. They are reported here because they sharpen interpretation of the usability and acceptance data: favourable assessment of model capability can coexist with dissatisfaction with interface design (Expert A), and disciplinary concern about AI-assisted reading is not captured by instruments designed to measure acceptance of an already-adopted tool (Expert B).

4 Discussion

Two substantive patterns emerge from the pilot. First, AI assistance was not associated with a uniform workload reduction or a uniform acceptance shift. Workload changes under PyThoth were bidirectional (Table 2). Usability ratings were sharply divided (Table 3). TAM trajectories ranged from large upward conversions to large disappointments (Section 3.5). The pattern most consistent with these observations is that AI assistance interacts with three factors simultaneously: intrinsic text complexity, interface-related extraneous load, and the match between a participant's pre-task expectation and the implemented interaction. These three factors are not independent but appear to be organised by the participant's position on a competence trajectory from student to professional. The expertise reversal effect predicts that support which scaffolds a novice reader becomes redundant - adding extraneous rather than germane load - for a reader who already holds the relevant sign schemata (Kalyuga et al. 2003; Kalyuga 2007), which would link our complexity analysis (Section 3.6) to our trajectory profiles (Section 3.5). Our data bear on this only weakly and unevenly. The one participant who re-encoded the assisted output into his own notational convention throughout is also the most experienced; but convergence on a common source is not otherwise ordered by experience - a second co-author of comparable standing reproduced the shared opening sentence together with its typographical error, and the participant whose responses diverge most from the shared text is not among the most senior. We therefore advance the trajectory reading as an interpretive hypothesis rather than a finding: the comparison was identified after inspecting the data, participants worked on different fragments, and two of the relevant participants are co-authors of this study.

Read in this light, Expert B's principled refusal is best understood not as an empirical objection to the tool but as a signal on the axis of disciplinary acceptance and

legitimacy: the concern is that assistance of this kind targets precisely the competences students must acquire for themselves. The concern identifies a real pedagogical boundary, yet it is answerable within the same framework that motivates our design. The expertise-reversal literature prescribes not the removal of support but its dynamic tailoring and fading as competence grows (Kalyuga 2007); and the zone-of-proximal-development framing requires that scaffolding be withdrawn so that the developing function is ultimately performed independently (Eun 2019). The risk Expert B identifies is therefore not AI-assisted reading as such, but a design that fixes support at a novice level instead of fading it.

Second, the usefulness/ease-of-use dissociation in the TAM data is substantively important for a research tool. Perceived usefulness remained stable on average, while perceived ease of use declined. On the individual level, the participants who converted upward did so mainly on the PU axis, while those who disengaged did so on the PEU axis. This configuration suggests that intellectual value and interaction cost can be evaluated independently: for PyThoth specifically, this means that improvements to the interaction layer (Gardiner-code disambiguation flow, correction affordances, step granularity) should be distinguished from improvements to the underlying model.

The feasibility of the protocol itself is a methodological result: in a small, internationally dispersed expert population, partial completion, delayed returns, and qualitative-only engagement are not exceptional but recurrent, and should be anticipated in the design of future studies rather than treated as attrition.

4.1 Limitations

Several limitations constrain the analyses reported above. First, the unit of observation is smaller than the number of attempts suggests: twenty-one attempts were contributed by nine participants with analysable outcomes, and attempts by the same participant, and pairs formed from them, are not independent observations. Two of the nine were co-authors of this study and were not blind to its hypotheses; four of the five completed unaided attempts were theirs, so that excluding co-authors leaves a single completed unaided attempt by one participant. On the self-report side, the nine questionnaire records come from eight respondents (one expert completed the instrument twice, for two successive fragment pairs), and one record reflects engagement with the system without any task-form submission; sample-level questionnaire means therefore weight one participant twice and are reported for records, not persons. Self-reported completion-time fields were incomplete and are not used; all timing analyses draw on the task log, with twenty-one submitted attempts, all of which enter the primary valid stratum: the single attempt logged against the wrong fragment was reassigned to the fragment its content matches (Table 1, footnote a), and this correction is applied consistently across all analyses. Second, the AI-assisted attempts are not independent of one another: several attempts on the same fragment share long verbatim runs, so the effective number of independent observations in that condition is substantially smaller than the number of attempts. Third, condition is confounded with fragment and with participant: two of the four fragments were attempted in a single condition only, no participant attempted the same fragment in both conditions, and assignment was not randomised. Fourth, the stela used here was excluded from the

training data of the sign-recognition component by design: absence from that training set was one of the criteria applied when candidate texts were selected, before any participant saw them. The transliteration and translation components, however, were trained on an aggregated TLA-derived dataset that does not preserve text-level provenance; text-level holdout therefore cannot be verified for those components in either direction. Independently of this, agreement with the TLA standard partly measures agreement with the editorial conventions on which those components were trained, and is not a neutral criterion across conditions. Fifth, the translation layer is compared here only by surface overlap with two reference translations, which penalises correct paraphrase and cannot speak to semantic adequacy; one unaided response was written in a language other than English and is not comparable on this measure at all. Finally, our use of the expertise reversal effect is interpretive: expertise here is indexed by professional trajectory and self-report, whereas the effect is defined over experimentally measured schema formation and element interactivity, neither of which we measured (Kalyuga et al. 2003). For these reasons all comparisons between conditions are reported as descriptive and case-comparative; no inferential or causal comparison of conditions is claimed, and the study is best read as establishing feasibility and generating hypotheses for a balanced randomised design.

5 Conclusion

Three conclusions follow from the pilot. AI assistance in specialist hieroglyphic reading is not a single-valued variable. Its effect varied with text complexity, interface cost, and pre-task expectation in ways this design cannot disentangle. Perceived usefulness and perceived ease of use should be measured and reported separately in this domain, since they respond differently to direct interaction with a tool such as PyThoth. Finally, experimental protocols for computational Egyptology should be designed from the outset to accommodate distributed, participant-paced completion rather than treating such patterns as attrition.

References

Bangor A, Kortum PT, Miller JT (2008) An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction 24(6):574–594. https://doi.org/10.1080/10447310802205776, URL http://www.tandfonline.com/doi/abs/10.1080/10447310802205776

Barucci A, Cucci C, Franci M, et al (2021) A deep learning approach to ancient egyptian hieroglyphs classification. IEEE Access 9:123438–123447. https://doi.org/10.1109/ACCESS.2021.3110082

De Cao M, De Cao N, Colonna A, et al (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). Association for Computational Linguistics, Hybrid in Bangkok, Thailand and online, pp 71–86, https://doi.org/10.18653/v1/2024.ml4al-1.9

Dell'Acqua F, McFowland III E, Mollick ER, et al (2023) Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Technology & Operations Mgt. Unit Working Paper 24-013, Harvard Business School, https://doi.org/10.2139/ssrn.4573321

Devos H, Gustafson K, Ahmadnezhad P, et al (2020) Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults. Brain sciences 10(12):994. URL https://www.mdpi.com/2076-3425/10/12/994

Eun B (2019) The zone of proximal development as an overarching concept: A framework for synthesizing Vygotsky's theories. Educational Philosophy and Theory 51(1):18–30. https://doi.org/10.1080/00131857.2017.1421941

Franken M, van Gemert JC (2013) Automatic egyptian hieroglyph recognition by retrieving images as texts. In: Proceedings of the 21st ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, pp 765–768, https://doi.org/10.1145/2502081.2502199

Gardiner AH (1957) Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs, 3rd edn. Griffith Institute, Ashmolean Museum, Oxford, sign list: pp. 438–548

Golyadkin M, Humonen I, Plevokas Y, et al (2025a) Automatic interpretation of ancient egyptian texts for education and research. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters. Association for Computing Machinery, New York, NY, USA, SIGGRAPH Posters '25, https://doi.org/10.1145/3721250.3743025, URL https://doi.org/10.1145/3721250.3743025

Golyadkin M, Humonen I, Rubanova V, et al (2025b) Mummy: Multimodal dataset supporting vlm-based egyptology research assistant. In: Proceedings of the 33rd ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '25, p 12875–12881, https://doi.org/10.1145/3746027.3758231, URL https://doi.org/10.1145/3746027.3758231

Golyadkin M, Rubanova V, Utkov A, et al (2025c) Meh: A multi-style dataset and toolkit for advancing egyptian hieroglyph recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 24488–24496

Humonen I, Golyadkin M, Kalin D, et al (2025) Automatic image translation of long ancient egyptian texts for augmented reality applications. In: 2025 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp 875–876, https://doi.org/10.1109/ISMAR-Adjunct68609.2025.00234

Humonen I, Golyadkin M, Rubanova V, et al (2026) Human-in-the-loop egyptology: A system for ancient egyptian text study. In: Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI EA '26, https://doi.org/10.1145/3772363.3798955, URL https://doi.org/10.1145/3772363.3798955

Hutson JW, Franklin AE, Rogers BA, et al (2024) Psychometric Testing of NASA-TLX to Measure Learners' Cognitive Load in Individual and Group Nursing Simulations. Clinical Simulation in Nursing 95:101607. URL https://www.sciencedirect.com/science/article/pii/S1876139924000999

Ibrahim F, Münscher JC, Daseking M, et al (2025) The technology acceptance model and adopter type analysis in the context of artificial intelligence. Frontiers in Artificial Intelligence 7:1496518. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1496518/full

Kalyuga S (2007) Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychology Review 19(4):509–539. https://doi.org/10.1007/s10648-007-9054-3

Kalyuga S, Ayres P, Chandler P, et al (2003) The expertise reversal effect. Educational Psychologist 38(1):23–31. https://doi.org/10.1207/S15326985EP3801_4

Levenshtein VI (1966) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10:707

Lewis JR (2018a) Measuring Perceived Usability: The CSUQ, SUS, and UMUX. International Journal of Human–Computer Interaction 34(12):1148–1156. https://doi.org/10.1080/10447318.2017.1418805, URL https://doi.org/10.1080/10447318.2017.1418805, _eprint: https://doi.org/10.1080/10447318.2017.1418805

Lewis JR (2018b) The System Usability Scale: Past, Present, and Future. International Journal of Human–Computer Interaction 34(7):577–590. https://doi.org/10.1080/10447318.2018.1455307, URL https://www.tandfonline.com/doi/full/10.1080/10447318.2018.1455307

Lewis JR, Utesch BS, Maher DE (2013) UMUX-LITE: when there's no time for the SUS. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '13, pp 2099–2102, https://doi.org/10.1145/2470654.2481287, URL https://doi.org/10.1145/2470654.2481287

Li Y, Wu B, Huang Y, et al (2024) Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. Frontiers in psychology 15:1382693. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1382693/full

Lichtheim M (1988) Ancient Egyptian Autobiographies Chiefly of the Middle Kingdom: A Study and an Anthology. Universitätsverlag / Vandenhoeck & Ruprecht, Freiburg, Switzerland / Göttingen, Germany

Papineni K, Roukos S, Ward T, et al (2002) Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D (eds) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318, https://doi.org/10.3115/1073083.1073135, URL https://aclanthology.org/P02-1040/

Popović M (2017) chrF++: words helping character n-grams. In: Bojar O, Buck C, Chatterjee R, et al (eds) Proceedings of the Second Conference on Machine Translation. Association for Computational Linguistics, Copenhagen, Denmark, pp 612–618, https://doi.org/10.18653/v1/W17-4770, URL https://aclanthology.org/W17-4770/

Richter TS, Hafemann I, Fischer-Elfert HW, et al (2018) Teilauszug der Datenbank des Vorhabens "Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache" vom Januar 2018. Thesaurus Linguae Aegyptiae – Data Sets, No. 1. Akademienvorhaben Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache, Berlin-Brandenburgische Akademie der Wissenschaften

Richter TS, Werning DA, Fischer-Elfert HW, et al (2024) Thesaurus linguae aegyptiae. Online Resource, URL https://thesaurus-linguae-aegyptiae.de, corpus issue 19, Web app version 2.2.0, 11/5/2024, ed. by Tonio Sebastian Richter & Daniel A. Werning on behalf of the Berlin-Brandenburgische Akademie der Wissenschaften and Hans-Werner Fischer-Elfert & Peter Dils on behalf of the Sächsische Akademie der Wissenschaften zu Leipzig

Sellam T, Das D, Parikh AP (2020) Bleurt: Learning robust metrics for text generation. In: Proceedings of ACL

The British Museum (n.d.) Stela BM EA 567, collection online record. URL https://www.britishmuseum.org/collection/object/Y_EA567, accessed: october 2025

The Unicode Consortium (2009) The Unicode Standard, Version 5.2.0. The Unicode Consortium, Mountain View, CA, URL https://www.unicode.org/versions/Unicode5.2.0/, egyptian Hieroglyphs block U+13000–U+1342F, 1,071 characters; released 1 October 2009

The Unicode Consortium (2024) The Unicode Standard, Version 16.0.0. The Unicode Consortium, South San Francisco, CA, URL https://www.unicode.org/versions/Unicode16.0.0/, egyptian Hieroglyphs Extended-A block U+13460–U+143FF, 3,995 characters; released 10 September 2024

Ueno T, Sawa Y, Kim Y, et al (2022) Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. ACM, New Orleans LA USA, pp 1–7, https://doi.org/10.1145/3491101.3519772, URL https://dl.acm.org/doi/10.1145/3491101.3519772

Notes

[1]The count is conventional and depends on what is included. Gardiner's sign list (Gardiner 1957), the standard reference for Middle Egyptian, catalogues 763 signs in common use and deliberately omits rarer forms; the first encoding of the script in the Unicode Standard (The Unicode Consortium 2009) fixed 1,071 characters (Egyptian Hieroglyphs, U+13000-U+1342F), largely following Gardiner. The attested repertoire is substantially larger, particularly in Ptolemaic and Roman-period texts: Unicode 16.0 (The Unicode Consortium 2024) added a further 3,995 signs (Egyptian Hieroglyphs Extended-A, U+13460-U+143FF). An order of a thousand therefore characterises the working inventory of Middle Egyptian - the register of the material used in this study - rather than the total number of attested sign forms.

[2]Participant codes P1-P13 are arbitrary identifiers; the mapping between codes and identities is known only to the research team. Two co-author participants are identified by the same coding scheme (P11, P13) and are noted as such at first mention in each section. Task submitters without linked questionnaire data are referred to as T1 and T2. Invited experts who returned written feedback without completing the protocol are referred to as Expert A and Expert B.

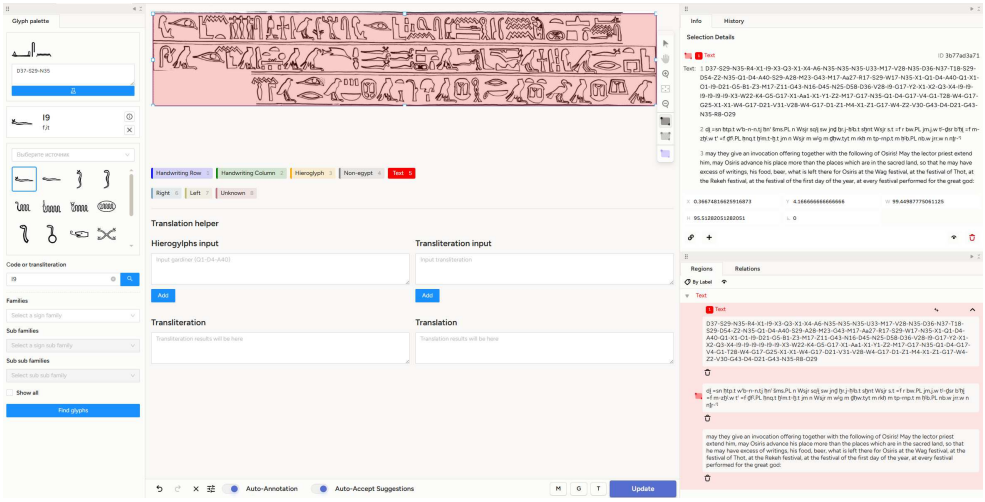


Fig. 1: PyThoth interface for AI-assisted study of Egyptian hieroglyphic texts. The system supports automatic text recognition, transliteration, and translation, as well as user refinement of intermediate results and manual lookup in a hieroglyph library, helping improve the final output. The source of the hieroglyphic image is The British Museum (n.d.).

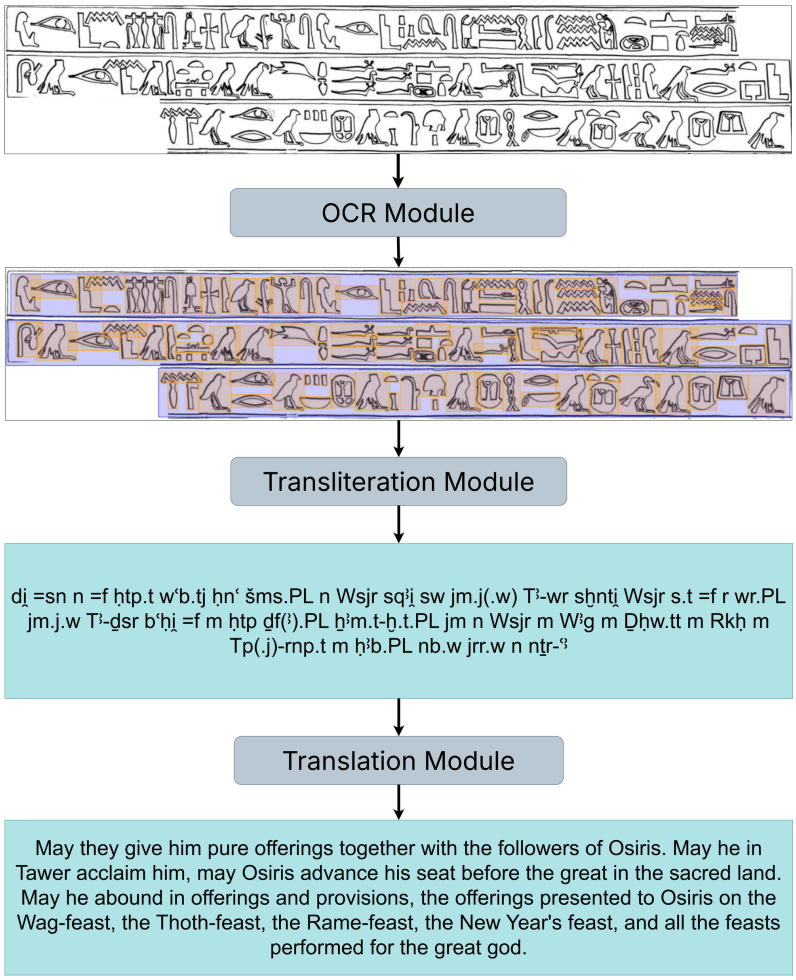


Fig. 2: Mirroring the human workflow, PyThoth maps an image of hieroglyphic text to an English translation. In the experiment, participants were likewise asked to produce these three results in the same order. The hieroglyphic text image, its transliteration, and its translation are taken from The British Museum (n.d.), Richter et al. (2024), and Lichtheim (1988), respectively.

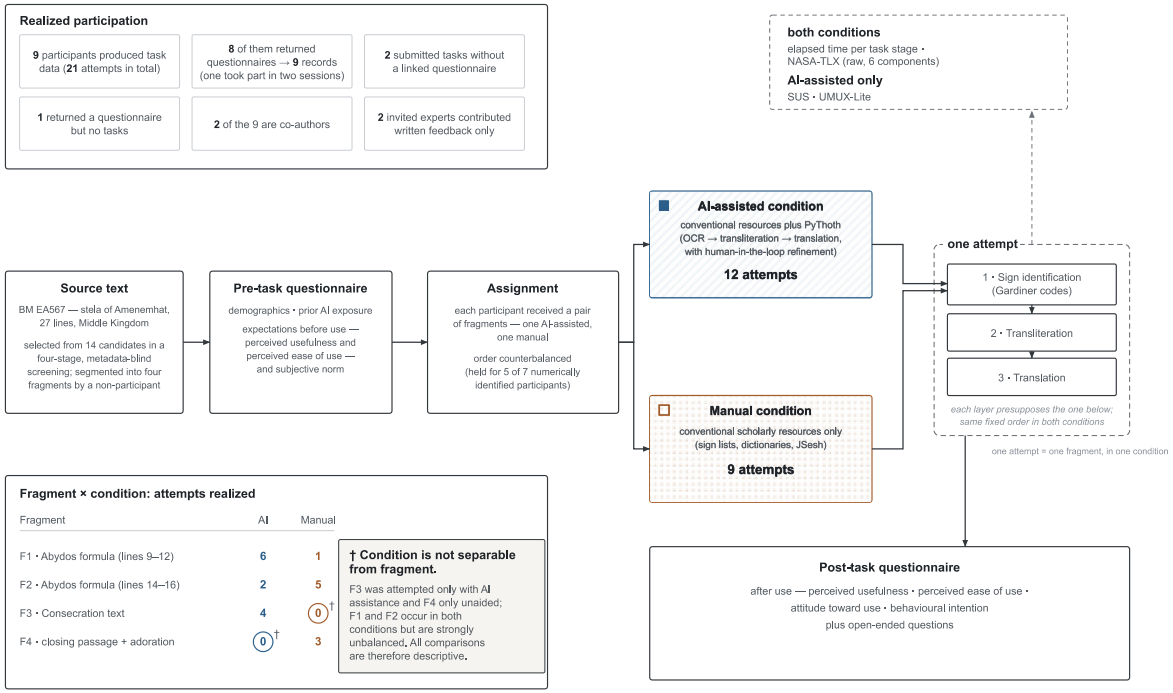


Fig. 3: Experimental design and realized participation. Four fragments of the stela BM EA567 were read under two conditions - unaided, using conventional scholarly resources, and AI-assisted, using the same resources together with PyThoth. Each participant received a pair of fragments, one per condition, in a counterbalanced order, and produced three results in a fixed sequence: Gardiner sign codes, transliteration, and translation. Twenty-one attempts were realized by nine participants (twelve AI-assisted, nine unaided); seven of the nine task participants also returned questionnaires; one further questionnaire-only participant brings the total to eight respondents and nine questionnaire records, as one respondent completed two sessions. Two of the nine are co-authors. The fragment-by-condition matrix shows that condition is not separable from fragment: Fragment 3 was attempted only with AI assistance and Fragment 4 only unaided, and the two remaining fragments are strongly unbalanced across conditions. Comparisons between conditions are therefore reported descriptively throughout.

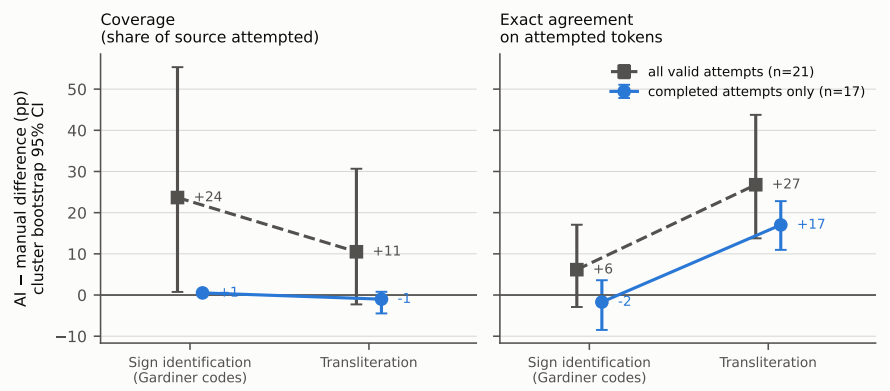


Fig. 4: AI − manual differences by layer in the primary valid stratum (cluster bootstrap 95% CIs; positive = higher under AI assistance). Left: coverage (response length relative to the source, capped at 100% - an output-length proxy for the share of the source attempted). Right: exact agreement with the TLA standard on attempted tokens, after notational normalisation. Restricting the comparison to attempts carried through all layers removes the coverage difference almost entirely, while the agreement contrast retains wide confidence intervals that include zero for sign identification; the raw between-condition gap is therefore carried primarily by attempts being carried to completion rather than by more accurate work where participants were working.

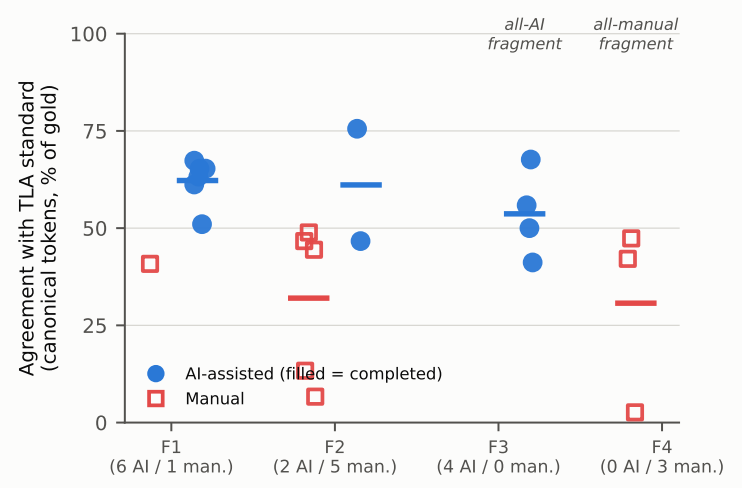


Fig. 5: Per-attempt agreement with the TLA standard (canonical tokens, % of gold) by fragment and condition, in the primary valid stratum (scoreable attempts; the attempt logged under an incorrect fragment label is counted under its corrected fragment, Table 1, footnote a). Horizontal bars mark condition means where at least two attempts are available. Condition is not separable from fragment: Fragment 3 was attempted only with AI assistance and Fragment 4 only unaided, so between-condition comparisons are reported descriptively.

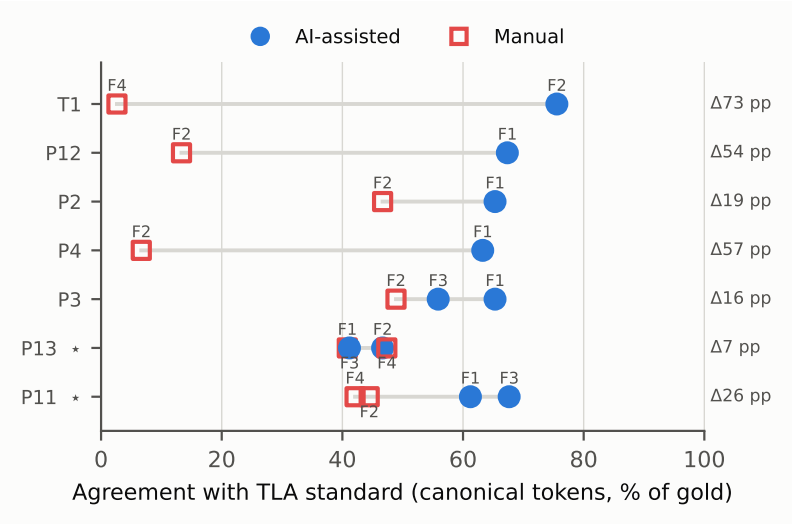


Fig. 6: Within-participant spread of agreement with the TLA standard across runs, for participants with at least two scoreable attempts. Circles = AI-assisted, squares = unaided; point labels give the fragment; ⋆ marks co-author participants. Person-by-fragment variation is of the same order as the between-condition contrast.

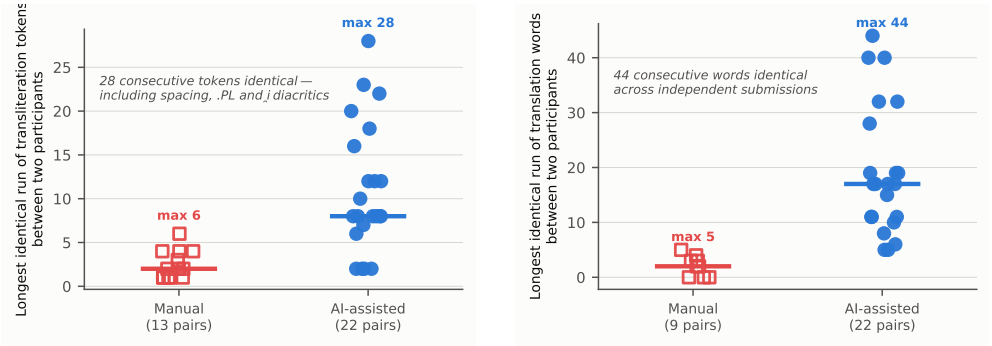


Fig. 7: Verbatim overlap between participants working on the same fragment in the same condition. Left: longest identical run of transliteration tokens (raw strings, including spacing, plural marking and diacritics). Right: longest identical run of translation words. Each point is a pair of attempts; horizontal bars mark medians. Long shared runs occur only in the AI-assisted condition (maximum 28 tokens and 44 words, against 6 and 5 unaided), consistent with convergence on a common source.

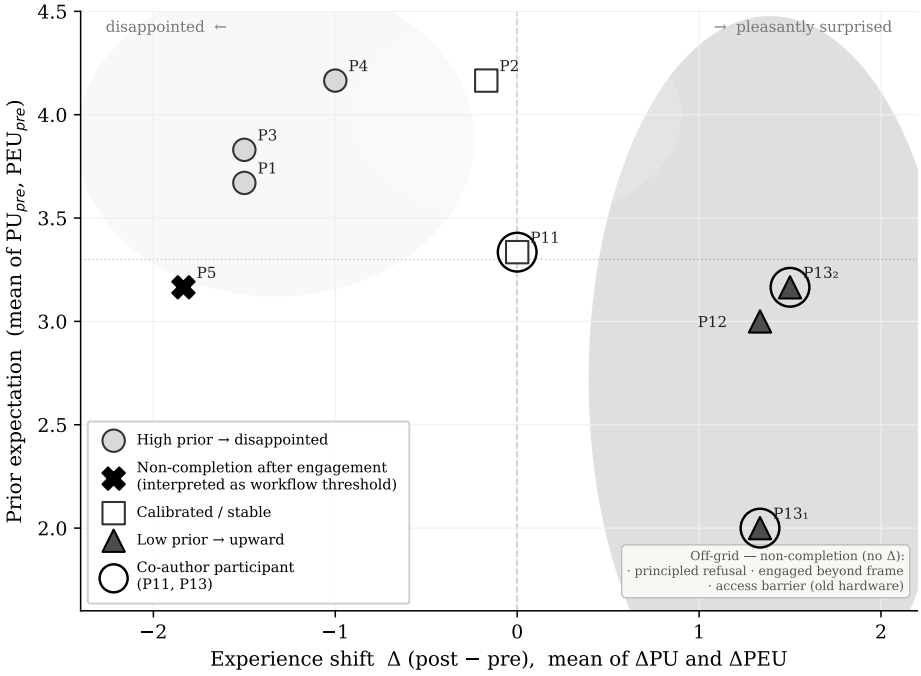


Fig. 8: Expectation-experience trajectories of the nine self-report records. Horizontal axis: post-experience shift (mean of ΔPU and ΔPEU); vertical axis: pre-task expectation (mean of $\mathrm{PU}_{\mathrm{pre}}$ and $\mathrm{PEU}_{\mathrm{pre}}$). Marker classes are within-session trajectory descriptors specific to this sample, not diffusion-of-innovations adopter types: high prior followed by disappointment (P1, P3, P4); low prior followed by upward revision (P12, $\mathrm{P13}_1$, $\mathrm{P13}_2$); near-zero mean shift (P2, P11 - P2's aggregate masks an internal usefulness/ease dissociation, see text); and non-completion after engagement, which we interpret as a digital-workflow threshold (P5). Ringed markers denote co-author participants (P11 and P13; P13 contributed two records). The off-grid inset lists non-completion modes for which no Δ was measured. Axes are arithmetically coupled (the horizontal axis is the change from the vertical baseline), so a negative baseline-shift association is partly expected by construction; the panel is a descriptive map of cases, not an inferential display.

Table 1: Participation matrix. Dataset composition across questionnaire, task submission, and qualitative contribution modes. Asterisk (*) indicates co-author participants whose dual role is discussed in Section 2.2. "AI/M" denotes AI-assisted and manual conditions respectively.

ID	Role	Task Fragments		Total	Complete	4-text	Self-	NASA-TLX
		AI	Manual	*N*	Pair	Protocol	Report	Pairs
P1	BA student	1	—	1	—	—	✓	AI only
P2	BA student	1	2	2	✓	—	✓	✓
P3	BA student	1, 3	2	3	✓	—	✓	✓
P4	BA student	1[a]	2	2	✓	—	✓	✓
P5	Senior res.	—	—	0	—	—	✓[b]	invalid
P11*	Senior res.	1, 3	2, 4	4	✓	✓	✓	✓
P12	BA student	1	2[c]	2	partial	—	✓	✓
P13*	Senior res.	2, 3	1, 4	4	✓	✓	✓(×2)	×2
T1	Student	2	4	2	✓	—	—	—
T2	Student	3	—	1	—	—	—	—
E1	Expert (letter)	—	—	0	—	—	written feedback	—
E2	Expert (letter)	—	—	0	—	—	written feedback	—

[a]Fragment number corrected from form-submitted "Fragment 2" to Fragment 1 on the basis of sign-sequence and translation content (ritual passage). [b]NASA-TLX ratings submitted as free-text commentary rather than numeric values, treated as qualitative data. [c]Manual condition abandoned during the Gardiner-coding stage. Transliteration incomplete, translation invalid.

Table 2: Perceived workload (adapted Raw NASA-TLX, scale 1–20) by participant and condition. ΔTLX < 0 indicates lower perceived workload in the AI-assisted condition. "Observed order" refers to the actual sequence of conditions reconstructed from submission timestamps. Two participants (P12, P13) reversed the prescribed counterbalancing assignment.

ID	RTLX Manual	RTLX AI	ΔTLX	Observed order
P2	8.83	8.00	−0.83	Manual first
P3	4.83	13.83	+9.00	AI first
P4	15.00	13.67	−1.33	Manual first
P11*	9.00	10.50	+1.50	AI first
P12	14.33	5.33	−9.00	AI first (reversed)
P13*(r1)	16.17	15.67	−0.50	Manual first (reversed)
P13*(r2)	13.83	12.33	−1.50	Manual first
Median	13.83	12.33	−0.83	
Mean	11.71	11.33	−0.38	

*Co-author participant. P13 completed the questionnaire twice, corresponding to two successive pairs of fragments.

Table 3: Usability scores for the AI-assisted workflow. SUS_9 is the 9-item all-positive System Usability Scale variant (Lewis 2018b). $UMUX_r$ is the regression-adjusted UMUX-Lite score (Lewis et al. 2013), bounded 22.9-87.9. Both are reported on the 0-100 scale.

ID	SUS_9	$UMUX_r$
P1	22.22	55.40
P2	69.44	71.65
P3	16.67	49.98
P4	52.78	66.23
P5	16.67	39.15
P11*	63.89	66.23
P12	63.89	55.40
P13*(r1)	63.89	49.98
P13*(r2)	72.22	77.07
Median	63.89	55.40
Range	16.67-72.22	39.15-77.07

*Co-author participant.

Table 4: Technology acceptance scores before and after direct experience with PyThoth. All scores on a 5-point Likert scale. PEU = Perceived Ease of Use, PU = Perceived Usefulness, ATU = Attitude Toward Use, BI = Behavioural Intention. Δ = post − pre. Positive Δ indicates favourable shift.

ID	PEU_{pre}	PEU_{post}	ΔPEU	PU_{pre}	PU_{post}	ΔPU
P1	3.67	1.33	−2.33	3.67	3.00	−0.67
P2	4.00	3.33	−0.67	4.33	4.67	+0.33
P3	3.33	1.33	−2.00	4.33	3.33	−1.00
P4	4.00	2.67	−1.33	4.33	3.67	−0.67
P5	3.33	1.33	−2.00	3.00	1.33	−1.67
P11*	3.67	4.00	+0.33	3.00	2.67	−0.33
P12	3.00	3.67	+0.67	3.00	5.00	+2.00
P13*(r1)	1.00	2.67	+1.67	3.00	4.00	+1.00
P13*(r2)	3.00	4.67	+1.67	3.33	4.67	+1.33
Mean	3.22	2.78	−0.44	3.56	3.59	+0.04

*Co-author participant. P13 completed the questionnaire twice, corresponding to two successive pairs of fragments. P5 completed the questionnaire after engaging with the system but submitted no task-form data. The mean ΔPU is near zero despite substantial individual movement, indicating bidirectional rather than uniform shift (see Figure 8 or Section 3.5).

Data availability. The de-identified data underlying every result reported in this article are provided with the paper: Supplementary Table S1 (scale-level descriptive statistics and internal consistency), Supplementary Table S2 (per-record self-report measures), Supplementary Data 1 (per-attempt task metrics for all 18 analysed attempts), Supplementary Data 2 (per-record self-report measures in machine-readable form) and Supplementary Data 3 (item-level questionnaire responses). The questionnaire instrument in both administered versions (Supplementary Files S2 and S3), the participant protocol and task instructions (Supplementary File S4) and the variable codebook (Supplementary File S5) are also provided. Raw questionnaire exports, the raw task log and scanned paper forms are not published: active readers of hieroglyphic texts form a small and internationally recognisable specialist community, and record-level free-text responses together with demographic detail would permit re-identification of individual participants. In line with the conditions communicated to participants at consent, these materials are held by the corresponding author, were made available to the editors and referees for the purposes of peer review, and are available to other researchers on official request with the approval of the HSE University ethics committee.

Acknowledgements. We thank the volunteer student who segmented the stela text, the two invited experts for their detailed written feedback, and all participants for their time and effort.

Competing interests. The authors declare no competing interests.

Ethics approval. and consent to participate: The study was reviewed and approved by the authors' institutional review board. All questionnaire participants provided written informed consent. Two further volunteers received the full study information package, including the informed-consent form, and voluntarily submitted task responses without returning a signed form or the questionnaire; their submission was treated as consent, only their anonymised task submissions are analysed, and no demographic or self-report data were collected from them.

Consent for publication. Participants consented to the use of their data in anonymised and aggregated form; no individual-level data tables are published, and anonymised examples of task responses are used as permitted by the consent form. The two participant co-authors additionally consented to the disclosure of their participant status.

Funding. This work was supported by a research grant

Materials availability. Not applicable.

Code availability. The PyThoth system is available from the authors on reasonable request. The analysis code used to compute the task metrics and figures reported here is available from the corresponding author on request.

Author contribution. Author 1 conceived the study. Author 1, Author 4, and Author 5 designed the experimental protocol and the adapted questionnaire instruments. Author 2 and Author 3 developed the PyThoth system. Author 1, Author 2, Author 3, and Author 5 collected the data. Author 1 curated and verified the

data. Author 1 and Author 2 performed the analysis. Author 1 and Author 2 wrote the original draft with input from Author 5. All authors reviewed and edited the manuscript. Two co-authors participated in the experimental tasks. The experimental design included a blinded text-selection procedure (Section 2.3) to preserve task naïveté despite this dual role.

Supplementary information. Supplementary Information accompanies this paper: Supplementary Table S1 (scale-level descriptive statistics and internal consistency of the self-report measures), Supplementary Tables S2 and S3 (task timing by fragment; corpus-based complexity profile of the four fragments), and Supplementary Data S1 (the full pre- and post-experiment questionnaire instrument, English version; a parallel version in participants' native language was also administered and is available on request).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFilesandParticipantQuestionnaires.zip
- CodingandPythonData.zip