

Retrieval-Augmented Machine Translation for Bohairic Coptic: A Pilot Case Study

So Miyagawa

Institute of Humanities and Social Sciences, University of Tsukuba

1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan

miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

Coptic is extremely low-resource in natural language processing despite its importance for philology, historical linguistics, and heritage-language learning. This paper reports a pilot case study of Bohairic Coptic–English translation using one passage from the *Vita Sinuthii*, one published English reference, and one retained output from each of seven deployed systems. THOTH AI, a retrieval-augmented system built with Coptological lexical, corpus, and grammatical resources, obtained 40.63 BLEU, compared with 31.40 for a separately accessed Gemini 3.0 Pro baseline. Results from chrF, ROUGE, TER, and METEOR are also reported, together with an illustrative expert reading of lexical errors and unsupported generation. These figures are descriptive, not a general benchmark or a causal retrieval ablation: the systems were accessed through different deployed interfaces, exact retrieval traces were not retained, and the evaluation contains only one passage. Three outputs contained clear passage-scale fabricated material. The study documents this bounded case and the evidence needed for a larger, reproducible evaluation.

1 Introduction

Egyptian has a documented history extending over more than four millennia, and Coptic is conventionally treated as its latest historical phase (Loprieno, 1995; Grossman and Richter, 2014). Coptic is written in an alphabet derived principally from Greek and supplemented by Demotic-derived characters. It is attested in several dialects rather than as a single uniform variety. Bohairic is the principal liturgical dialect of the Coptic Orthodox Church today, whereas Sahidic predominates in many modern grammars and digitally annotated corpora (Layton, 2000; Till, 1961). This imbalance has direct computational consequences: a model trained or retrieved primarily on Sahidic data cannot be assumed to

have adequate Bohairic lexical, orthographic, or grammatical coverage.

After the Arab conquest of Egypt, the domains of written and spoken Coptic gradually contracted as Arabic became dominant, although the chronology and regional course of this shift were not uniform (Rubenson, 1996). Coptic nevertheless remains indispensable for Egyptian historical linguistics, biblical and patristic studies, and the textual traditions of Egyptian Christianity. It also continues to be taught and used liturgically. Computational tools for Coptic thus serve several overlapping constituencies: philologists working with manuscripts and editions, linguists examining the Egyptian–Coptic continuum, and learners seeking access to Bohairic religious texts.

For natural language processing (NLP), however, Coptic is extremely low-resource. Available texts are modest in comparison with major modern languages, dialects and orthographies vary, and reliable parallel data are limited. Hagiographic translation adds culturally specific vocabulary, embedded discourse, and formulaic narrative patterns that can encourage a generative model to produce fluent but source-unsupported material. Recent work has begun to apply neural machine translation and large language models (LLMs) to Coptic, but these constraints remain substantial (Enis and Megalaa, 2024; Wannaz and Miyagawa, 2024).

Retrieval-augmented generation (RAG) offers a practical way to place specialist lexical, corpus, and grammatical evidence in a general-purpose model’s context at inference time. Yet an apparent gain from a deployed RAG application cannot automatically be ascribed to retrieval: differences in system prompts, provider interfaces, decoding defaults, and model revisions can all affect an output. The present study therefore asks a deliberately narrow observational question: what behavior is found when one deployed Coptic RAG system and

six general-purpose interfaces translate the same Bohairic passage?

The paper makes three bounded contributions. First, it documents the passage, prompt, retained-output procedure, deployed retrieval settings, interface labels, and automatic scores for a pilot case. Second, it compares THOTH AI descriptively with six separately accessed systems while making the interface confound and missing retrieval trace explicit. Third, it supplies an operational error taxonomy and an illustrative expert analysis of every retained translation, including three passage-scale cases of unsupported narrative. The study does not claim a representative benchmark, a same-code RAG ablation, a model-wide hallucination rate, or a validated annotation dataset.

2 Related Work

2.1 Digital Resources for Coptic

Coptic digital infrastructure includes Coptic SCRIPTORIUM, which provides linguistically annotated corpora and an NLP pipeline (Schroeder and Zeldes, 2016; Zeldes and Schroeder, 2016); the Coptic Dictionary Online (Feder et al., 2018); and the Comprehensive Coptic Lexicon (CCL), which integrates Egyptian-Coptic lexical material with Greek loanword data (Burns et al., 2020). The Corpus dei Manoscritti Copti Letterari represents an earlier major application of digital methods to Coptic manuscript research (Orlandi, 1990). Community resources such as Tasbeha.org additionally make liturgical texts searchable for church users and learners (Tasbeha.org, 2026).

These resources are complementary rather than interchangeable. A manuscript catalogue, an annotated corpus, a lexicon, a reference grammar, and a liturgical website differ in structure, evidential status, and intended task. Their dialect balance also matters: substantial Sahidic resources do not automatically resolve Bohairic forms. This motivates a retrieval design that can present dialect-sensitive lexical evidence together with grammatical and textual context. It does not, by itself, demonstrate that retrieval improves translation; that question requires a controlled comparison beyond the observational case reported here.

2.2 Coptic and Low-Resource Machine Translation

For Coptic machine translation, Enis and Megalaa (2024) investigate a specialized low-resource neu-

ral pipeline using normalization, romanization, backtranslation, and domain restrictions. Wannaz and Miyagawa (2024) evaluate LLMs on Coptic and Ancient Greek ostraca and identify source-grounding problems in Coptic outputs. More broadly, massively multilingual translation work shows both the promise and persistent limitations of transfer to low-resource languages (Costa-jussà et al., 2022). The present paper instead examines a Bohairic hagiographic passage and a retrieval-based deployed system.

The setting differs from ordinary modern-language machine translation in several respects. The source language is historical and dialectally diverse; spelling and editorial practice vary; parallel data are small; and a published literary translation may be more interpretive than a sentence-aligned benchmark reference. Consequently, surface-overlap metrics can penalize legitimate philological choices while still rewarding outputs that reproduce the wording of one translator. These properties motivate combining automatic scores with source-based expert reading.

2.3 Retrieval-Augmented Generation

RAG conditions generation on documents retrieved from an external collection (Lewis et al., 2020). THOTH AI previously applied this general approach to Middle Egyptian translation (Miyagawa, 2025). The system evaluated here is not a new retrieval algorithm: it is a domain composition of standard vector retrieval and a foundation model, implemented in Dify (Dify, n.d.), with Coptological resources selected and configured for translation. The contribution is the documented Bohairic case and failure analysis, not an architectural innovation over standard RAG.

3 System and Comparison Design

3.1 THOTH AI and Author Role

THOTH AI is developed and maintained by the author, who selected its sources, configured the deployed system, collected the comparison outputs, and performed the qualitative analysis. This dual role creates a potential conflict of interest and increases the importance of independently reproducible evaluation. No THOTH AI output was manually post-edited before scoring.

The evaluated deployment used the interface label *Gemini 3.0 Pro Preview* as its generator. This is the label preserved in the study record, not a claim

Setting	Recorded value
Chunk structure	General
Index method	High quality
Embedding model	text-embedding-3-large
Retrieval	Vector search
Reranker	None configured
Top- k	3
Score threshold	0

Table 1: Recorded Dify settings for the CCL knowledge base.

that the provider still exposes an identical archival snapshot. THOTH AI adds retrieved material to the model context before generation. Its principal Coptic sources were:

- a system-internal representation based on CCL v1.2, containing Egyptian-Coptic lexical material and Greek loanwords (Burns et al., 2020);
- Coptic SCRIPTORIUM corpus data and linguistic annotation (Schroeder and Zeldes, 2016; Zeldes and Schroeder, 2016); and
- digitized reference grammars, including Layton (2000), Reintges (2018), and the dialectal description of Till (1961).

This evidence mix is intended to address Bohairic lexical ambiguity and the uneven dialectal coverage of individual resources. It should not be understood as a homogeneous training corpus: the sources differ in format, dialect, and scholarly purpose, and are used as retrieval-time context rather than as newly fine-tuned model parameters.

3.2 Recorded Retrieval Configuration

The deployment was built in Dify. Table 1 reports the settings preserved for the CCL knowledge base. Dify’s general chunk structure was used; parent-child and structured question-answer indexing were not used for this resource.

The exact retrieved chunks, source identifiers, ranks, similarity scores, and complete context passed for the evaluated query were not archived. The test passage and Bell’s translation were not intentionally loaded as evaluation-specific records. Nevertheless, without the retrieval trace, incidental overlap with indexed materials cannot be categorically excluded. The present paper therefore makes neither a passage-level no-leakage claim nor a causal attribution to particular retrieved evidence.

The settings in Table 1 describe the deployment, but they are not a reconstruction of the missing query trace.

3.3 Nature of the Comparison

THOTH AI and the nominally corresponding Gemini baseline were accessed through different deployed interfaces. System prompts, provider-side preprocessing, hidden defaults, and model revisions may therefore differ. This is a comparison of deployed systems, not a same-code RAG-on/RAG-off experiment. A causal ablation would require a shared API model identifier, identical decoding and system prompts, fixed seeds where available, saved retrieval traces, and repeated generations.

4 Experimental Design

4.1 Evaluation Passage

The source is a Bohairic passage from the *Vita Sinuthii*, following *Sinuthii archimandritae vita et opera omnia*, volume I, §88 (Leipoldt, 1906). The English reference is Bell’s translation of §88 (pp. 67–68; Bell, 1983). Because the exact page or line range in the 1906 edition has not been rechecked against a physical or page-image copy, the paragraph identifier rather than an unverified page number is reported for the Coptic source.

The scored unit is one passage-level segment. By whitespace-delimited surface counting, the preserved source contains 53 Coptic surface tokens in four sentence-like units; the reference contains 102 English tokens in five sentence-like units. These are descriptive counts for the exact strings supplied to the metric implementation, not claims about morphologically segmented Coptic words. No sentence-level score distribution can therefore be inferred from the table.

The passage concerns the curse and punishment of Kesios/Gesios and contains eschatological vocabulary, embedded speech, and polysemous body-part terminology. It was selected as a philologically difficult case, not sampled to represent Coptic genres or Bohairic generally.

4.2 Prompt and Collection

Each interface received the same user-facing request:

Please translate and analyze this passage: [Bohairic Coptic source passage].

No in-context examples or model-specific hints were added. For scoring, the English translation portion was manually separated from any surrounding analysis. One retained generation per model was scored; outputs were collected manually through the deployed interfaces in January 2026. The preserved record does not contain stable API model IDs, exact per-output timestamps, provider system prompts, temperature, top- p , maximum-token settings, random seeds, or a complete retry history. Provider defaults applied where controls were not exposed. These omissions prevent exact regeneration and are treated as limitations rather than reconstructed retrospectively.

4.3 Evaluated Systems

The labels below reproduce those exposed by the interfaces at collection time; commercial preview labels and aliases are not archival identifiers. The evaluated systems were THOTH AI (Gemini 3.0 Pro Preview), Gemini 3.0 Pro, Claude 4.5 Opus, Grok 4.1, DeepSeek V3.2, Kimi K2, and Qwen3 Max. Provider documentation is cited for the corresponding model families and interfaces (Google, n.d.; Anthropic, n.d.; xAI, n.d.; DeepSeek, n.d.; Moonshot AI, n.d.; Alibaba Cloud, n.d.). The manuscript does not imply that the providers expose reproducible snapshots under these labels. An eighth condition labelled “GPT 5.2 Pro” in an earlier draft was excluded from the camera-ready analysis because its original output string could not be recovered and its preserved scores could therefore not be independently audited. No output from a later GPT model was substituted.

4.4 Metrics

The retained translation strings were compared with the single Bell reference using BLEU (Papineni et al., 2002), chrF (Popović, 2015), ROUGE-1/2/L (Lin, 2004), TER (Snover et al., 2006), and METEOR (Banerjee and Lavie, 2005), through Hugging Face Evaluate.¹ In Evaluate, the `chrF` and `ter` modules are backed by `sacreBLEU`, whereas `bleu` and `sacrebleu` are separate modules. The scoring script, Evaluate version, and module revisions were not archived, so the exact BLEU module and configuration used for the retained values cannot now be verified. The score is therefore reported simply as BLEU, following the caution about implementation-sensitive BLEU reporting

in Post (2018). BLEU emphasizes exact word n-gram precision; chrF is based on character n-grams; ROUGE-1/2 measure unigram and bigram overlap; ROUGE-L uses longest-common-subsequence overlap; TER estimates the edits needed to match the reference; and METEOR allows limited English stemming and synonym matching. Scores are presented on a 0–100 scale; lower TER is better, and TER can exceed 100 when the required edits outnumber reference tokens.

The metrics have different formal emphases, but they are all computed from the same retained hypothesis and the same single reference. They are therefore correlated descriptive views rather than independent confirmations of translation quality. No aggregate average is calculated, and the analysis does not treat a small numerical difference as a reliable system ranking.

5 Results

Table 2 gives the preserved scores. In this single case, THOTH AI obtained 40.63 BLEU and the separately accessed Gemini baseline obtained 31.40, a difference of 9.23 points. THOTH AI also has the highest chrF, ROUGE, and METEOR values and the lowest TER in the table. These observations describe the one retained set of strings; Section 3 explains why they cannot be interpreted as an isolated retrieval effect. Qwen3 Max’s TER of 102.94 means that the edit count exceeded the reference length, not that the metric calculation failed.

The seven metrics move broadly together in this case, which is unsurprising because they are computed from the same strings and reference. For example, the differences between Gemini and Claude are much smaller than those between either of them and the three outputs with passage-scale unsupported narrative. Even this contrast is passage-specific: the table should not be used to infer population-level rankings or provider-wide quality. Stable comparison would require more texts, multiple retained generations, controlled inference conditions, and human assessment.

6 Illustrative Expert Error Analysis

6.1 Definition and Annotation Scope

An *unsupported addition* is defined here as a target-language proposition for which the annotator could identify no plausible corresponding source span and which was not a reasonable explication required for intelligible English. A *genre-driven fab-*

¹<https://huggingface.co/docs/evaluate>

System	BLEU	chrF	R-1	R-2	R-L	TER↓	METEOR
THOTH AI (Gemini 3.0 Pro Preview)	40.63	63.84	76.85	48.76	69.95	44.12	70.35
Gemini 3.0 Pro	31.40	60.20	70.87	42.16	66.02	50.00	66.78
Claude 4.5 Opus	29.97	56.91	70.14	40.19	63.51	50.98	64.70
Grok 4.1	16.32	49.36	59.33	29.95	53.59	64.71	48.33
Kimi K2	6.64	40.88	48.87	15.53	32.58	90.20	36.66
DeepSeek V3.2	4.76	33.69	39.79	10.58	27.23	83.33	24.08
Qwen3 Max	3.38	30.55	37.44	3.69	19.18	102.94	22.20

Table 2: Descriptive scores for one passage-level segment and one retained output per system. R-1/2/L are ROUGE F1 scores.

rication is a sequence of unsupported propositions that together form a coherent hagiographic episode. These phenomena differ from (i) mistranslation, where an identifiable source item is rendered incorrectly; (ii) omission, where source content has no target counterpart; and (iii) interpretive variation, where the target remains plausibly licensed by the source but selects one philological or cultural interpretation.

The sole author, a researcher in Coptology and Egyptian linguistics, inspected all seven retained translation strings against the Coptic source and reference. There was no second annotator, adjudication, or inter-annotator-agreement calculation. Accordingly, this section is an illustrative expert reading, not a validated annotation study. At output level, three of seven retained translations—DeepSeek V3.2, Kimi K2, and Qwen3 Max—contained at least one clear passage-scale genre-driven fabrication. Category frequencies below the output level are not reported because no pre-registered span annotation or independent labels were available.

Table 3 records the codebook used for this reading. The categories are analytical distinctions, not mutually exclusive labels for an entire output: a single clause may contain, for example, both a lexical error and a structural divergence. The examples are drawn from this passage and are not proposed as an exhaustive taxonomy for Coptic machine translation.

6.2 Lexical and Interpretive Cases

***amenti* (the West/netherworld).** Crum glosses the Bohairic form *amenti* as “hades, literally western place” (Crum, 1939, 8b). The expression therefore denotes the western place or netherworld, understood as the realm of the dead. Bell renders it as “hell,” placing the expression in a Christian English interpretive frame. THOTH AI, Gemini 3.0 Pro, and Claude 4.5 Opus retained

“Amenti” or “Amente,” sometimes with an explanatory gloss, whereas Grok 4.1 used “Hades.” The three strategies—transliteration, Christianizing translation, and a classicizing substitution—are not equivalent, but neither can their relative philological adequacy be decided by single-reference overlap alone. An automatic metric will favor Bell’s wording even when transliteration is a defensible editorial choice.

***ini ... nte tef caloč* (great toe of his foot).** The relevant phrase contains two body-part lexemes. Crum glosses Bohairic *ini* as “thumb; great toe” and cites this exact *Vita* phrase in that entry (Crum, 1939, 81a); *caloč* is separately glossed “foot, knee” (Crum, 1939, 814a). The construction *e-ti-ini nte tef caloč* therefore means “to the great toe of his foot”: *ini* supplies “great toe,” while *nte tef caloč* supplies “of his foot.” Bell’s rendering specifies the “big toe.” THOTH AI likewise produced “big toe,” while Gemini 3.0 Pro and Claude 4.5 Opus used “heel.” Grok 4.1 instead produced “root of his tongue,” which does not identify the source body part correctly. This contrast is compatible with the possibility that domain retrieval supplied useful lexical evidence. It is not proof of that mechanism: because the retrieved chunks were not archived, the study cannot show that a particular lexicon entry was retrieved or that it caused THOTH AI’s selection.

***helēnos* (normalized *hellēnos*; Greek/pagan).** The source form is a Greek-derived ethnonym, within the Coptic–Greek loanword domain represented in CCL (Burns et al., 2020). More directly, the Coptic Dictionary Online glosses normalized Coptic *hellēnos* as “pagan, gentile” and relates it to *Hellēn*, for which both the ethnic sense “Greek” and the Christian religious senses “gentile, pagan” are registered (KELLIA, n.d.). Patristic Greek lexicography corroborates this religious extension (Lampe, 1961, s.v. *Hellēn*). Bell’s “pagan” is there-

Category	Operational question	Illustration in this case
Lexical mistranslation	Is an identifiable Coptic lexical item mapped to an English item not licensed in context?	Rendering <i>ini ... nte tef caloč</i> as “heel” although <i>ini</i> denotes the great toe and <i>caloč</i> the foot.
Cultural or theological interpretation	Does the output select a culturally situated interpretation while remaining plausibly source-licensed?	Transliteration “Amenti,” Bell’s “hell,” and the classicizing “Hades” represent different treatments of the netherworld term.
Morphosyntactic or structural divergence	Does the output alter predicate–argument relations, negation, tense/aspect, reported speech, or clause linkage?	Reassigning an action or embedded utterance to a participant not licensed by the source syntax.
Omission	Does a source proposition or participant lack a target counterpart?	Leaving out part of the curse or punishment sequence without replacing it elsewhere.
Unsupported addition	Does the target assert a proposition with no plausible source span and no need for English explication?	DeepSeek’s added imperative “Let them do my will.”
Genre-driven fabrication	Do multiple unsupported propositions form a fluent new hagiographic episode?	Qwen’s statement that “Christ appeared to him in the dream” and the narrative developed around it.

Table 3: Operational taxonomy used for the illustrative expert reading. Quoted fragments are evidence from retained outputs, not reference translations.

fore lexicographically and contextually well supported. The general interfaces’ “Greek” remains etymologically and semantically related, so it is not an unsupported addition, but it is less context-sensitive in this passage.

Taken together, these items show why the expert reading is not reducible to marking every deviation from Bell as an error. The relevant question is whether the English output is licensed by the Coptic source, and, if so, what interpretive commitment it makes. Multiple references or a human adequacy protocol would represent this distinction more faithfully than a single surface form.

6.3 Unsupported Narrative

Three retained outputs contain passage-scale material that meets the operational definition of genre-driven fabrication. DeepSeek V3.2 adds “Let them do my will,” an imperative for which no corresponding source proposition was identified. Qwen3 Max introduces an episode in which “Christ appeared to him in the dream.” Kimi K2 mixes fragments related to the passage with extended hagiographic prose that cannot be aligned to the Coptic sequence. These are not simply free translations of difficult words: each output adds propositions or an episode while maintaining the tone expected of a saint’s life.

This fluency makes the failure important for educational and scholarly use. A visibly broken translation invites checking, whereas a genre-compatible addition may be accepted by a reader who cannot consult the source. At the same time, the evidence remains deliberately limited. The study establishes

three observed output-level cases among seven retained translations of this passage; it supplies neither a per-proposition error rate nor evidence about the providers’ models across other texts. Without repeated sampling, independent annotators, and a registered span-level scheme, the cause and prevalence of the behavior remain unknown.

7 Discussion

7.1 What This Case Can Support

The observed score pattern and the *ini ... nte tef caloč* example are consistent with specialist retrieval being useful for Bohairic translation. They do not establish that retrieval alone produced the difference. THOTH AI and the general Gemini interface did not share a fully controlled inference stack, and the missing trace prevents connecting an output token to a specific retrieved record. The defensible result is therefore narrower: in one preserved deployed-system comparison, the THOTH AI output was closer to Bell’s translation across the reported metrics and avoided the passage-scale unsupported narrative observed in three other outputs.

This narrower claim remains informative as a system case study. It identifies a concrete evaluation item, exposes the configuration and provenance information that a deployed RAG system should retain, and demonstrates that an apparently straightforward comparison becomes difficult to interpret when commercial interfaces hide model and decoding details. For future ancient-language NLP experiments, logging retrieval evidence is not merely an engineering convenience. It is part of

the scholarly apparatus needed to evaluate whether a proposed translation is grounded in a lexicon, grammar, parallel passage, or potentially leaked reference.

7.2 Metrics and Philological Interpretation

The lexical cases illustrate a basic mismatch between single-reference metrics and philological translation. Bell’s “hell” and “pagan” are interpretive choices. A model that returns “Amenti” or “Greek” may diverge from the reference while remaining connected to the source. Conversely, a fluent output may share religious vocabulary with the reference while adding an unsupported event. BLEU, chrF, ROUGE, TER, and METEOR can summarize formal similarity, but none can determine whether a theological or cultural interpretation is appropriate.

For this reason, the metrics are best read at coarse resolution in the present paper. They make the large surface differences among these seven strings visible; they do not establish a definitive quality order or replace source-based assessment. A stronger protocol would combine multiple translations where feasible with expert ratings of semantic adequacy, fluency, named entities, grammatical relations, and source support. Span-level links between an English proposition, the Coptic source, and any retrieved evidence would make the analysis auditable.

7.3 Educational and Revitalization Context

Modern efforts to extend the learning and use of Coptic have a history well preceding AI. Nineteenth- and twentieth-century initiatives associated with Coptic church education and figures such as Claudius Labib produced pedagogical materials and new vocabulary, while modern digital corpora and dictionaries have broadened access to scholarly resources (Meinardus, 2002; Miyokawa, 2017). Retrieval-based tools belong to this longer history of mediation; they do not by themselves constitute language revitalization.

Their plausible educational value is narrower and practical. A system may help a learner locate lexical alternatives, compare a draft translation with grammatical evidence, or formulate questions for further study. That value depends on showing provenance and uncertainty. Ordinary mis-translation may be noticed by an advanced learner, whereas fluent religious narrative can appear authoritative. A responsible interface should there-

fore separate source text, retrieved evidence, generated translation, and commentary, and should encourage consultation of editions, dictionaries, and qualified readers rather than presenting generated prose as an authoritative ecclesiastical or scholarly translation.

8 Limitations and Reproducibility

The evaluation has one passage-level segment, one dialect, one hagiographic work, one English reference, and one retained generation per system. It provides no variance or uncertainty estimate and cannot establish population-level rankings. The passage was chosen for philological difficulty rather than randomly or stratifiably sampled, so its vocabulary and narrative structure may disproportionately affect the observed contrast.

The deployed-system comparison confounds retrieval with interface, hidden system instructions, provider-side preprocessing, decoding defaults, and model revision. Exact API snapshot identifiers and exact per-output timestamps were not preserved. Temperature, top- p , maximum-token settings, seeds, and a complete retry history are also unavailable. The interface labels are therefore historical labels from the collection record, not reproducible model identifiers.

The retrieval evidence needed for a causal or leakage analysis was not archived. Complete chunks, source identifiers, ranks, scores, and the context actually passed to the generator are missing. Although the passage and Bell’s translation were not intentionally ingested as evaluation-specific items, this does not prove that related or overlapping text was absent from the indexed resources or the foundation model’s training data. The paper consequently cannot distinguish assistance from lexical evidence, incidental textual overlap, hidden interface behavior, or general model competence.

The qualitative analysis was conducted by the author, who also develops THOTH AI. Domain expertise permits close source reading, but involvement creates a risk of expectation bias. There was no independent annotator, adjudication, or inter-annotator-agreement estimate, and no pre-registered span-level error inventory. The 3/7 observation is therefore an output-level description, not a general hallucination frequency. Automatic metrics add a separate limitation: all use one published English translation and cannot determine philological adequacy for culturally loaded renderings.

A reproducible follow-up should use a fixed API model snapshot with RAG toggled within the same code path; archive all prompts, parameters, output attempts, and retrieval traces; run multiple generations; and evaluate a planned sample of Bohairic and Sahidic texts across genres. At least two independent Coptologists should annotate source support and translation adequacy using a fixed codebook, with disagreements adjudicated and agreement reported. Multiple references or acceptable-variant annotations should be added where copy-right and scholarly practice permit. Specialized Coptic MT systems should also be included alongside general LLM interfaces.

9 Conclusion

This paper reports a pilot case study of seven deployed systems translating one Bohairic passage. In the preserved comparison, THOTH AI produced the strongest single-reference automatic scores, while three other retained outputs contained clear passage-scale unsupported narrative. Detailed lexical cases also show that disagreement with one reference may reflect a defensible philological choice rather than simple error.

Because the sample is minimal, the interfaces are confounded, and the retrieval trace is missing, the findings are not a general benchmark or a causal demonstration of RAG effectiveness. The contribution is instead a transparent account of what was observed, an operational taxonomy for distinguishing mistranslation from unsupported narrative, and a specification of the controls and provenance needed for a defensible larger study of Coptic machine translation.

Conflict of Interest and Author Role

The author develops and maintains THOTH AI and configured and evaluated the system reported here. This constitutes a potential conflict of interest. Independent replication with archived retrieval traces and separately annotated outputs is needed.

Ethical Considerations

Machine translation of historical religious texts can mislead scholarly, educational, or faith communities when fluent output is mistaken for a source-grounded translation. The systems studied here should be used as aids, not substitutes for philological review. Reporting unsupported generation

and the limits of the evidence is therefore central to responsible use.

References

- Alibaba Cloud. n.d. [Model Pricing: Qwen3 Max](#). Official Model Studio documentation. Accessed January 2026.
- Anthropic. n.d. [Claude Opus 4.5 System Card](#). System card. Accessed January 2026.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI. Association for Computational Linguistics.
- David N. Bell. 1983. *The Life of Shenoute*. Number 73 in Cistercian Studies Series. Cistercian Publications, Kalamazoo, MI.
- Dylan Michael Burns, Frank Feder, Katrin John, and Maxim Kupreyev. 2020. [Comprehensive Coptic lexicon: Including loanwords from ancient Greek v 1.2](#). Dataset, Freie Universität Berlin.
- Marta R. Costa-jussà et al. 2022. [No language left behind: Scaling human-centered machine translation](#). arXiv preprint arXiv:2207.04672.
- Walter E. Crum. 1939. *A Coptic Dictionary*. Clarendon Press, Oxford.
- DeepSeek. n.d. [DeepSeek API Updates: DeepSeek-V3.2 Release](#). Official API documentation. Accessed January 2026.
- Dify. n.d. [Knowledge Retrieval](#). Online documentation. Accessed January 2026.
- Maxim Enis and Andrew Megalaa. 2024. [Ancient voices, modern technology: Low-resource neural machine translation for Coptic texts](#). Technical report, Coptic Translator. 15 pages; accessed August 2026.
- Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T. Schroeder, and Amir Zeldes. 2018. [A linked Coptic dictionary online](#). In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2018)*, pages 12–21, Santa Fe, NM. Association for Computational Linguistics.
- Google. n.d. [Gemini 3 Pro Preview](#). Online documentation. Accessed January 2026.
- Eitan Grossman and Tonio Sebastian Richter. 2014. [The egyptian-coptic language: Its setting in space, time and culture](#). In Eitan Grossman, Martin Haspelmath, and Tonio Sebastian Richter, editors, *Egyptian-Coptic Linguistics in Typological Perspective*, volume 55 of *Empirical Approaches to Language Typology*, pages 69–102. De Gruyter Mouton.

- KELLIA. n.d. [Coptic Dictionary Online: TLA Lemma C11084 \(hellēnos\)](#). Online lexicon entry. Accessed August 11, 2026.
- G. W. H. Lampe, editor. 1961. *A Patristic Greek Lexicon*. Clarendon Press, Oxford.
- Bentley Layton. 2000. *A Coptic Grammar: With Chrestomathy and Glossary. Sahidic Dialect*. Harrassowitz, Wiesbaden.
- Johannes Leipoldt, editor. 1906. *Sinuthii archimandritae vita et opera omnia, Vol. I: Sinuthii vita Bohairice*. Number 41 in CSCO. E Typographeo Reipublicae, Paris. Adiuvante W. E. Crum.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona. Association for Computational Linguistics.
- Antonio Loprieno. 1995. *Ancient Egyptian: A Linguistic Introduction*. Cambridge University Press, Cambridge.
- Otto F. A. Meinardus. 2002. *Two Thousand Years of Coptic Christianity*. American University in Cairo Press, Cairo.
- So Miyagawa. 2025. [RAG-enhanced neural machine translation of Ancient Egyptian text: A case study of THOTH AI](#). In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 33–40, Albuquerque, USA. Association for Computational Linguistics.
- Hiroko Miyokawa. 2017. [The revival of the Coptic language and the formation of Coptic ethnoreligious identity in modern Egypt](#). In Nelly van Doorn-Harder, editor, *Copts in Context: Negotiating Identity, Tradition, and Modernity*, pages 151–156. University of South Carolina Press, Columbia, SC.
- Moonshot AI. n.d. [Kimi K2](#). Official GitHub repository. Accessed January 2026.
- Tito Orlandi. 1990. The Corpus dei Manoscritti Copti Letterari. *Computers and the Humanities*, 24:397–405.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: Character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT 2015)*, pages 392–395, Lisbon. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.
- Chris H. Reintges. 2018. *Coptic Egyptian (Sahidic Dialect): A Learner’s Grammar*, 2nd edition. Rüdiger Köppe, Cologne.
- Samuel Rubenson. 1996. [The transition from Coptic to Arabic](#). *Égypte/Monde arabe*, 27–28:77–92.
- Caroline T. Schroeder and Amir Zeldes. 2016. [Raiders of the lost corpus](#). *Digital Humanities Quarterly*, 10(2).
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231. Association for Machine Translation in the Americas.
- Tasbeha.org. 2026. [Coptic church liturgical resources](#). Website. Accessed January 2026.
- Walter C. Till. 1961. *Koptische Dialektgrammatik: Mit Lesestücken und Wörterbuch*, 2nd edition. C. H. Beck, Munich.
- Audric-Charles Wannaz and So Miyagawa. 2024. [Assessing large language models in translating Coptic and Ancient Greek ostraca](#). In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 463–471, Miami, USA. Association for Computational Linguistics.
- xAI. n.d. [Grok 4.1](#). Official announcement. Accessed January 2026.
- Amir Zeldes and Caroline T. Schroeder. 2016. [An NLP pipeline for Coptic](#). In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, pages 146–155, Berlin. Association for Computational Linguistics.