Assessing Large Language Models in Translating Coptic and Ancient Greek Ostraca

Audric-Charles Wannaz
University of Basel
Petersgraben 51
4051 Basel, Switzerland
audric.wannaz@unibas.ch

So Miyagawa
University of Tsukuba
1-1-1 Tennodai
Tsukuba, Ibaraki, Japan
miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

The advent of Large Language Models (LLMs) substantially raised the quality and lowered the cost of Machine Translation (MT). Can scholars working with ancient languages draw benefits from this new technology? More specifically, can current MT facilitate multilingual digital papyrology? To answer this question, we evaluate 9 LLMs in the task of MT with 4 Coptic and 4 Ancient Greek ostraca into English using 6 NLP metrics. We argue that some models have already reached a performance that is apt to assist human experts. As can be expected from the difference in training corpus size, all models seem to perform better with Ancient Greek than with Coptic, where hallucinations are markedly more common. In the Coptic texts, the specialised Coptic Translator (CT) competes closely with Claude 3 Opus for the rank of most promising tool, while Claude 3 Opus and GPT-4o compete for the same position in the Ancient Greek texts. We argue that MT now substantially increases the incentive to work on multilingual corpora. This could have a positive and long-lasting effect on Classics and Egyptology and help reduce the historical bias in translation availability. In closing, we reflect upon the need to meet AI-generated translations with an adequate critical stance.

1 Introduction

Translations have been the cornerstone of scholarly activity in the fields of Classics and Egyptology since their inception, serving both academic and public dissemination purposes (Balmer, 2009; Westerfeld, 2016). The initial preference for Latin as the target language for translations reflects its status as the scholarly *lingua franca* during the early phase of these disciplines (Lockwood, 1918; Burke, 2017) . Over the centuries, there has been a steady transition to vernacular languages in order to make scientific content more accessible for an attempt to partly reverse this transition; See Merisalo 2015 for the example of the Italian language). In contemporary practice, English has emerged as the preferred lingua franca, broadening the accessibility and scope of translated texts (Nørgaard, 1958). This article focuses specifically on English translations in a field tangent to both Classics and Egyptology, namely digital documentary Papyrology[1]. In this specific area of study, interdisciplinary communication between Classics and Egyptology have improved slightly in last decades (van Minnen, 1993, 14). However, English translations of primary sources are not yet widely available on the Web, as the next section shows.

Currently, virtually all available translations of ancient Greek and Coptic texts have been made by human experts. The coverage of those translations reveals major disparities. On Papyri.info[2] as of 5 May 2024, 59,955 Greek texts with transliteration are available, but only 5,678 are accompanied by translations in English, and 628 in other languages, i.e. around one tenth of the total corpus. The situation is even more critical for Coptic, where out of 2,099 texts, 2,049 are untranslated, and only 50 texts are available in English, French or German, that is less than one per cent of the total corpus. If other translations exist, they are mainly printed and are not easily accessible online, making them unsuitable for research in digital papyrology.

Meanwhile, the landscape of AI-generated translations has evolved considerably, from simple rule-

[1] "Digital Papyrology can be defined as the whole set of electronic resources and methodologies aimed at creating, storing, accessing, processing, and publishing information pertaining to research and study in the various fields of interest of the papyrological discipline." (Reggiani, 2017, 8).

[2] "Papyri.info has two primary components. The Papyrological Navigator (PN) supports searching, browsing, and aggregation of ancient papyrological documents and related materials; the Papyrological Editor (PE) enables multiauthor, version controlled, peer reviewed scholarly curation of papyrological texts, translations, commentary, scholarly metadata, institutional catalog records, bibliography, and images." https://papyri.info/ [Accessed: 24/05/2024].

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 463–471
November 16, 2024.

based systems to sophisticated machine learning models. Early efforts in computer-aided translation were fundamental, but limited in terms of accuracy and scope. In recent years, the adoption of machine learning models has significantly improved the quality of translations. Since its launch in 2006, Google Translate (Wu et al., 2016) has long been the benchmark for machine translation tools, despite the initial lack of support for languages such as Coptic or ancient Greek.[3] More recently, DeepL, introduced in 2017, has set new standards for the accuracy of machine translations (although it does not include the languages in question). The most advanced development in this area concerns large generative multimodal language models (LLMs), which are serious contenders for complex translation tasks (Yang et al., 2024; Gaspari, 2024).

1.1 Research Question

1.1.1 General Research Problem

Given the obvious gaps in translation in the fields of Classics and Egyptology, especially with respect to Digital Papyrology, this study will investigate whether MT can effectively fill these gaps today. The central question concerns the ability of modern AI-driven tools to provide accurate and reliable translations for ancient documents that remain largely untranslated or not digitised.

1.1.2 Specific Research Objectives

The objective of this study is to evaluate the effectiveness of MT systems in facilitating multilingual digital papyrology. This includes a comprehensive examination of the performance of these technologies in translating Coptic and Ancient Greek, two common languages in this field (Vierros and Henriksson, 2017; Dahlgren, 2018). To this end, our methodology is structured as follows:

Evaluation of the MT of Coptic texts (Section 2): We first introduce four Coptic texts (2.1). Then, we introduce 6 NLP metrics to evaluate the performance of 9 LLMs: the Coptic Translator, a LLM specialised on the task of Coptic-English as well as 8 generic LLMs (2.2). Evaluation of Ancient Greek texts (Section 3): Similarly, four Ancient Greek texts are presented as close equivalents to the four Coptic texts in form and content (3.1). In a second step, we describe how the same LLMs used to translate Coptic texts fare in the same task in this other ancient language (3.2). Comparative analysis (Section 4): on the basis of all produced results of MT for Coptic and Ancient Greek, we discuss the overall impact of AI-generated translations on the field of digital papyrology. Future directions (Section 5): The study closes with a discussion of the potential future implications of integrating MT into academic research and public dissemination.

In sum, the aim of this pilot study is to provide some empirical information on the current practical capabilities of AI in translating ancient texts and to stimulate debate on its strategic integration in the fields of Classics and Egyptology.

2 Evaluating Coptic-English MT

2.1 Four Coptic Texts

To evaluate the performance of Machine Translation (MT) on Coptic texts, we selected four relatively well-preserved documentary letters written on ostraca from the IFAO (Institut français d'archéologie orientale) collection: TM 874362, 874363, 874364, and 874365, which are unlikely to be used in the training of the existing LLMs.[4]

- TM 874362/ IFAO OC 252 (C 1906): 11 x 11 cm. VII CE, Western Thebes. Late Roman Amphora 7, Letter from Petros concerning a church vessel of Apa Menas in Ape (Luxor).

- TM 874363/ IFAO Inv. OC 275 (C 1917): 16 x 10 x 1.2 cm. VII-VIII CE, Theban region. Late Roman Amphora 7, Letter from the sick Antonios to Petros, asking for money, possibly to buy medicine.

- TM 874364/ IFAO Inv. OC 104 (C 1916): 10 x 9.5 cm. VII CE, Theban region. Letter from Psmoei to a deacon announcing the repayment of a tremissis and requesting lentils. Fragment of a red Pseudo-Aswanese Late Roman Amphora.

- TM 874365/ IFAO Inv. OC 270 (C 1879): 17 x 11 cm. VII-VIII CE, Thebes(?). Pseudo-Aswanese pottery. Letter concerning exchange of crops, vegetables, dates, arax (legume), and oil between several individuals.

[3] Generally, the production of textual corpora, which can be training data for machine translation, has been much less in Coptic than in Greek; cf. Clackson (2004).

[4] Later, more comprehensive studies will benefit from using a larger sample size. This preliminary study chooses to limit its scope to a few short texts that are rather homogeneous in content.

These texts were chosen to cover a range of preservation states (TM 874362 and TM 874365 are well-preserved, TM 874364 is sufficiently well-preserved, TM 874363 is partially preserved) and standardized character lengths (averaging about 225 characters). Ground truth reference translations were produced by Coptic scholars under the supervision of two eminent experts (Anne Boud'hors and Esther Garel; see Boud'hors and Garel, 2019).

2.2 Assessing Coptic-English MT

2.2.1 LLMs

We compared the output of 9 LLMs on the MT task: The dedicated Coptic Translator model. GPT model family:[5] GPT-4o, GPT-4, GPT-3.5. Claude model family: Claude Opus, Claude Sonnet, Claude Haiku. Gemini model family: Gemini Advanced, Gemini.[6]

The Coptic Translator (Enis and Megalaa, 2024), developed by Maxim Enis and Andrew Megalaa from Williams College Computer Science Department, is the first contextual machine translation system for the Coptic language. The authors created the system by fine-tuning pretrained multilingual transformer models on limited Coptic-English parallel data and employing techniques such as romanization, back-translation, and transfer learning, resulting in strong translation performance on religious Coptic texts. The translator provided the first-ever English translations for over 5,700 previously untranslated Coptic sentences and will be open-sourced and made freely available online to assist Coptic language learners, scholars, and those working to revive the language.

2.2.2 Metrics

To quantitatively assess translation quality, we employed 6 metrics from the field of natural language processing (NLP):

- "school": a custom metric designed to mimic a human approach to the task of MT evaluation (see Figure 1).[7]
- Levenshtein distance: character-level edit distance.
- BLEU (BiLingual Evaluation Understudy; Papineni et al. 2002): n-gram precision with a brevity penalty. To add reliability, we used the standardised version, SacreBLEU (Post, 2018).
- TER (Translation Error Rate; Snover et al., 2006): word-level edit distance.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering; Banerjee and Lavie, 2005): alignment-based metric.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation; Lin, 2004): n-gram recall.

```
import difflib

def school_metric(base_text, text,
    ↪ weights={'reused_diff': 0.5, '
    ↪ not_reused_or_present': 1}):
    words1, words2 = set(base_text.split
        ↪ ()), set(text.split())
    #SequenceMatcher algorithm
    reused_diff = sum(1 for w1 in words1
        ↪  if difflib.get_close_matches
        ↪ (w1, words2, n=1, cutoff=0.8)
        ↪  and w1 not in words2)
    not_reused_or_present = len(words1.
        ↪ symmetric_difference(words2))

    score = (weights['reused_diff'] *
        ↪ reused_diff +
            weights['
                ↪ not_reused_or_present
                ↪ '] *
                ↪ not_reused_or_present
                ↪ )
    return score
```

Figure 1: Python code of "school" scoring

These metrics capture different aspects of similarity between the MT and human reference translations. Levenshtein and TER measure the amount of editing required to transform one text into the other. BLEU, METEOR (optimised to evaluate MT) and

[5] Models within a family are listed in decreasing recency. Model size and performance are generally correlated.

[6] Since the conception of this paper, several significant LLMs have been released, including OpenAI's o1-mini and o1-preview. These newer models will be incorporated into future studies, with the goal of generating more robust quantitative results compared to the preliminary findings presented in this proof-of-concept study.

[7] Similar to a teacher correcting a test at school, this metric counts "mistakes" (words absent either from the base or target text) and "half-mistakes" (words reused in target text but slightly modified).

metric	4 Coptic Texts top3models	meanscore	4 Greek Texts top3models	meanscore
school	gemini (71.75) gemini_advanced (67.00) gpt_4o (61.25)	58.14	gpt_3.5 (45.38) gemini (40.88) gpt_4 (38.50)	37.03
levenshtein	gemini_advanced (312.50) gpt_4 (237.25) gemini (236.75)	223.78	gpt_3.5 (188.25) gpt_4 (157.25) gemini (156.75)	148.16
ter	gemini_advanced (1.27) gpt_4 (0.99) gemini (0.95)	0.92	gpt_3.5 (0.63) gemini (0.55) gpt_4 (0.53)	0.50
sacrebleu	claude_opus (20.02) claude_haiku (11.52) coptic_translator (8.43)	5.98	gpt_4o (39.63) claude_opus (37.18) claude_sonnet (33.89)	30.89
meteor	claude_opus (0.46) claude_haiku (0.35) coptic_translator (0.30)	0.23	claude_opus (0.67) gemini_advanced (0.65) claude_haiku (0.65)	0.60
rouge	claude_opus (0.44) claude_haiku (0.37) coptic_translator (0.34)	0.25	claude_opus (0.65) gpt_4o (0.64) claude_haiku (0.61)	0.59

Table 1: Raw metric results

ROUGE (optimised to evaluate machine summarization) evaluate the degree of word and phrase overlap. Together, they provide a multifaceted view of translation quality. While more recent and sophisticated metrics like METEOR and ROUGE may be better markers, metrics on the other side of the spectrum behave in a more straightforward way and thus represent a bridge from human qualitative evaluation to more complex metrics.

2.2.3 Results (Coptic Texts)[8]

The results of comparing nine preprocessed[9] translations made by different LLMs across four Coptic texts are illustrated in Table 1 and Figure 2. Each graph represents one of the Coptic texts, with the x-axis showing different evaluation metrics and the y-axis displaying scaled, directionally normalised values of these metrics.

For texts TM 874362 and TM 874363, the specialized Coptic Translator model and the general-purpose Claude Opus performed consistently well, achieving scores near the top across most metrics. Claude Haiku also showed relatively high performance but lagged slightly behind the top performers. The Gemini and Gemini Advanced models exhibited lower performance, with scores dropping significantly in certain metrics, particularly TER and ROUGE. Other models, such as GPT variants and Claude Sonnet, displayed mixed results, performing well in some metrics and poorly in others. A similar trend was observed for TM 874364, where Claude Opus, Claude Haiku and the Coptic Translator emerged as strong performers. However, TM 874365, which is in a poor preservation state, posed challenges for all models, leading to generally lower scores and greater variability. The Gemini models and some GPT variants, in particular, struggled significantly with this text, indicating difficulty in handling degraded source material. The specialized Coptic Translator model and the general-purpose Claude Opus and Haiku achieved

[8]The code used to obtain these results can be found at https://github.com/somiyagawa/GreekCopticMTEval.

[9]Given the low number of texts involved in this pilot study, we opted for a semi-manual normalisation and preprocessing of the strings involved to make the metrics more meaningful. In addition to common steps like lowercase punctuation removal, we also made more case specific choices. For example, some archaic expressions were modernised ("thou art"→ "you are") and the spelling of names was uniformized (*-os*/*-us* ending). The complete steps involved will be made available together with all other results in a jupyter notebook upon acceptance of this paper.

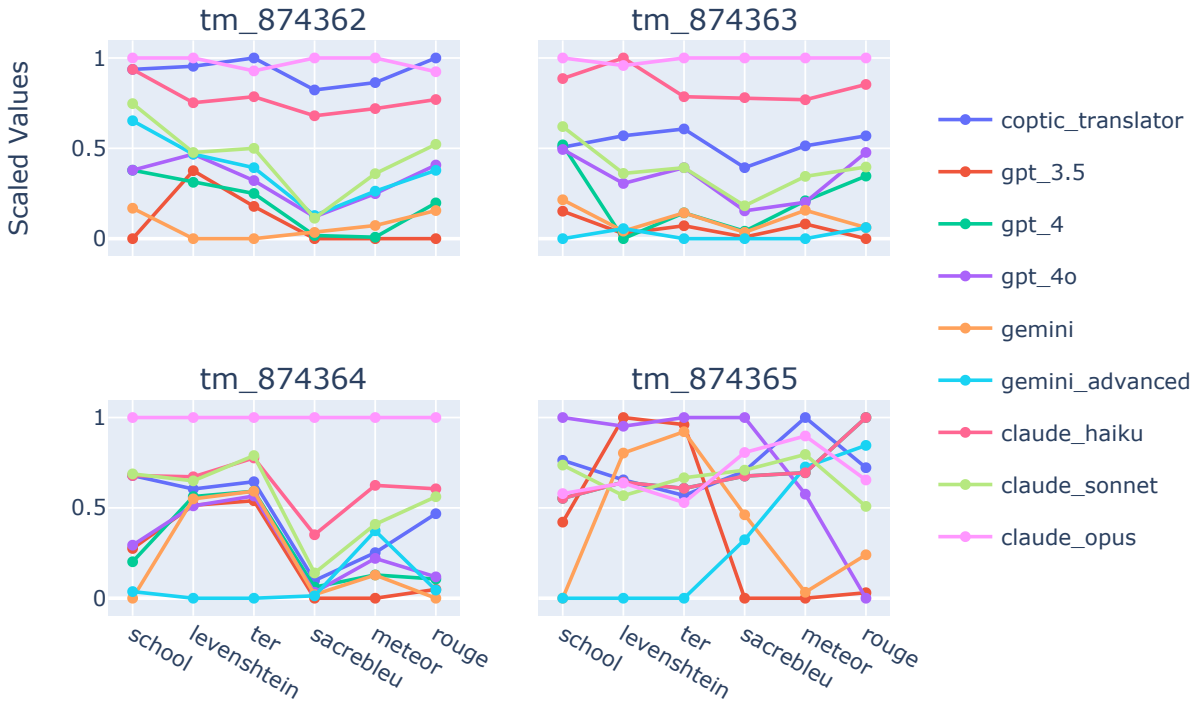

Figure 2: Scaled metrics evaluating LLMs' MT (Coptic texts)

the strongest results across the four texts. Claude Opus performed best, obtaining the highest scores on two out of four texts. In contrast, the GPT and Gemini model families struggled to produce viable translations, often generating largely irrelevant or incoherent text. When comparing across the four texts, all models found TM 874365 the most difficult, likely due to its poor preservation state. The texts in good condition, TM 874362 and TM 874363, yielded the best translation quality overall. This highlights the significant impact that the physical deterioration of source material can have on the MT process.

Examining the different metrics, we observe reasonable agreement in model rankings between "school", Levenshtein, TER, and METEOR. However, ROUGE scores exhibit more variability, suggesting that n-gram recall may be less reliable for ancient languages. In summary, the Coptic Translator, Claude Haiku, and Claude Opus demonstrate the potential for usable MT of Coptic texts, although challenges remain with heavily damaged ostraca. The GPT and Gemini models appear unsuitable for this domain based on their inability to generate meaningful translations. GPT and Gemini tend to output Biblical quotations triggered by a proper name in the text. In the next section, we turn our attention to Ancient Greek to explore whether these findings generalize to another historically significant language.

3 Assessing Ancient Greek-English MT

3.1 Four Greek Texts

To compare the MT metrics obtained with the four coptic letters on ostraca, we selected four Greek texts of similar length that are also letters on ostraca: TM 817896, 89219, 89224 and 42504. For each of them, an openly available English translation made by human experts is provided on Papyri.info.

- TM 817897/ Pap.Lugd.Bat. 23 S. 7: 9 x 11.3 cm. II CE, Thebes. It is clearly a business letter, but since it is broken away at the right, the exact transaction between sender and addressee is not completely clear.

- TM 89219/ O.Ber. II 193: 9.3 x 8.7 cm. I CE, Berenike. is also a fragmentary business letter. Its end is missing, the opening and the first requests are preserved.

- TM 89224/ O.Ber. II 198: 11.5 x 9.5 cm. I CE, Berenike. From the same historical context as TM 89219, it seems to be complete. It also discusses one business-like matter together with more social elements.

- TM 42504/ O.Mich. I 91: 12.9 x 12.7 cm. III CE, Arsinoites. It is arguably in the best preservation state. This letter seems to have been sent mainly in order to obtain the favor of using borrowed oxen for an extended period of time.

3.2 Results (Ancient Greek Texts)

Notably, there seem to exist no equivalent of the CT for Ancient Greek-English translation yet. Table 1 and Figure 3 shows the metrics of the translations by the other eight LLMs. Looking at the scaled metrics across the four chosen texts, we note the following: Claude Opus, the best generic LLM in Coptic-English translation, seems to remain competitive in Greek-English translation too. Notably, the two other Claude models, Sonnet and Haiku, also perform well here. The latest model GPT-4o makes the most visible relative improvement and distances itself visibly from its predecessors GPT-3.5 (the worst performing model) and GPT-4, except in TM 817896. Upon inspection of the raw values, this text has been translated comparatively similarly across all models. We suspect this is due to its rather basic vocabulary and syntax. While Gemini Advanced scores better overall than its base model, the difference is not as pronounced, similar to the three Claude models but with slightly worse results.

From a Qualitative point of view, all models have performed better than expected, the amount of hallucinations differed radically from the one observed in translations of Coptic. In the next section, we explore whether unscaled quantitative metrics confirm this impression.

4 Comparing Results (Coptic and Greek)

Figure 4 shows the mean performance all surveyed LLMs achieved on one given metric at text level. It reveals that all MTs of Ancient Greek texts obtained better scores than their Coptic equivalents on those metrics. We note that the gap is even bigger in more complex, possibly more meanigful metrics like SacreBLEU and METEOR.

We explain this perceived difficulty of current LLMs to translate Coptic texts at the level of Ancient Greek texts with the likely very large gap in available training corpus for each corpus. This cannot be said in certainty for LLMs that are not open source, but a recent digital contribution estimated the overall size of Coptic digital papyrology to 102,080 words across 1,973 texts. This represents about two percent of the Greek equivalent, 4,926,263 words across 58,975 texts (Riaño Rufilanchas, 2024). Otherwise, no likely secondary factors have been found that contribute to the difference in MT performance. The length or completeness of the text appears unlikely to play a role in the eight chosen texts. (Cf. Riaño Rufilanchas, 2024)

5 Future Directions

This study examines Coptic/Greek-English MT by LLMs in only eight texts and could be scaled and improved in multiple directions.

- Latin and other ancient languages could be included in the evaluation process. Doing so could reveal nuances in translation performance beyond mere training corpus size.
- While human translations were used as ground truth, we acknowledge that experts vary in their translaring styles and preferences. Future studies might benefit from gathering translations from the same human expert en masse.
- The MTs were retrieved by accessing the Chat UIs respective to each model family and inputting a basic prompt (for example: “translate this text” + target string). Developing a more complex prompting pipeline might improve performance.
- Similarly, accessing the LLMs via API rather than via a Chat UI would allow to fine-tune the models and could alter the overall performance. This would also allow to collect large samples of translations for one individual ancient text and to better study variance within a given model.
- Experiments conducted during research for this paper suggest that the directionality of the translation greatly affects performance. English-Ancient Greek translation beared comparatively worse results than its counterpart. Many LLMs refused to perform English-Coptic translation altogether, or produced strong hallucinations with close to no grounding in the original prompt. This anecdotical experience suggests that back translation is a promising task to evaluate in the future, albeit current models will liekly yield poor results in this task.

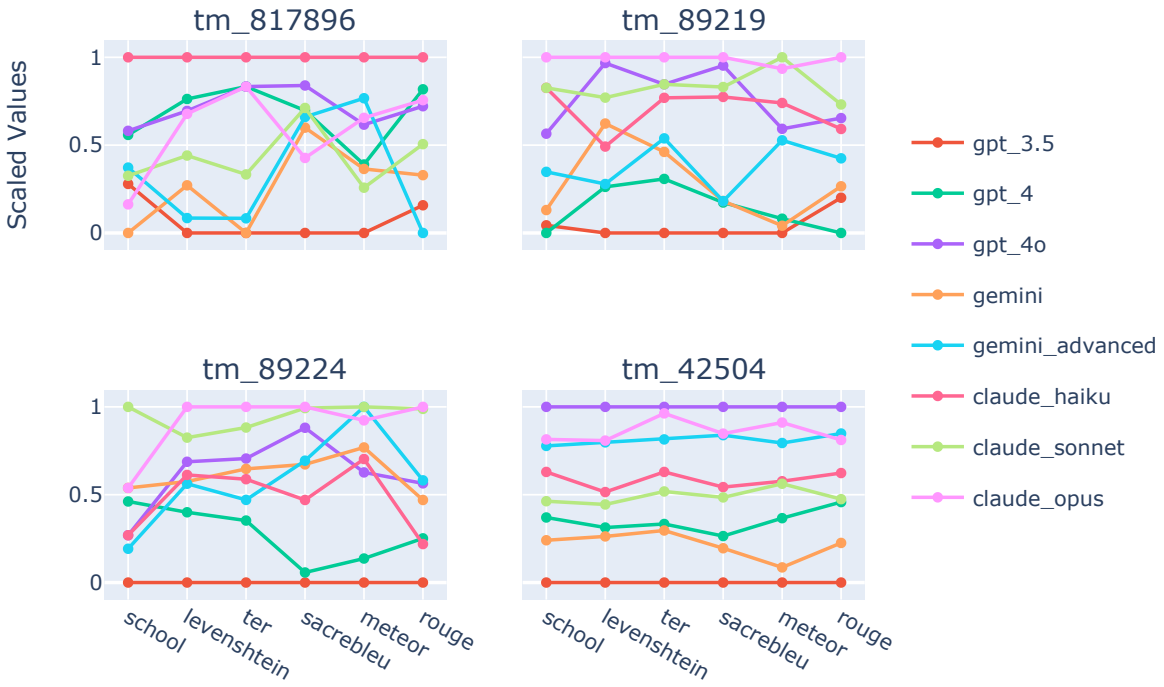

Figure 3: Scaled metrics evaluating LLMs' MT (Ancient Greek texts)

In sum, many steps could be taken to improve either the MTs themselves or their critical evaluation. However, we note that performance related considerations are not the sole concern of the scholarly assessment of MT.

- In this pilot study, paid models did not overwhelmingly outperform their free equivalents. However, this may be specific to the task as we defined it. In the face of the rapid changes in the AI industry, it can not be excluded that this will change.
- We advocate for a rethinking of translations in digital datasets of ancient texts in the LLMs era. Despite their high quality, translations by human experts are limited by the availability of specialists and not systematically added to open source databases due to a lag in publishing practices.
- There is a need for a deeper reflection upon the shared and distinct goals in MT between the industry and academia. For example, output speed is a metric current models are competing over while it bears close to no significance in the context of academic MT, where translation quality is preferred over all other aspects. Altered behaviour in translating sensitive contents is another aspect where academia and industry might have unaligned wants.

6 Conclusion

The results presented in this paper serve as recommendations for leveraging currently available Large Language Models for the Machine Translation of Ancient Greek and Coptic texts. Egyptologists and Classicists seeking to focus on just one of the nine models evaluated will find the specialised Coptic Translator or the Claude model family most beneficial. Although the best-performing version, Claude Opus, requires a paid subscription, the lighter Haiku and Sonnet models produced nearly equivalent results for Ancient Greek. The success of the Coptic Translator, a smaller specialised tool fine-tuned from a larger model, stands out in the context of increasingly capable general-purpose models. GPT-4o, the newest model included in this pilot study, did not show significant improvement for Coptic but did for Ancient Greek. Aside from GPT-3.5, which consistently underperformed, the other three models form a middle tier in performance. However, scholars are not limited to selecting only one model, especially given the rapid development and deployment of new models, alongside related ethical considerations. We advocate for the joint use of multiple models to provide the best support for human experts.

Clearly, more work is needed to fully understand how scholars can benefit from LLM-powered translations. Nevertheless, it is evident that multilin-

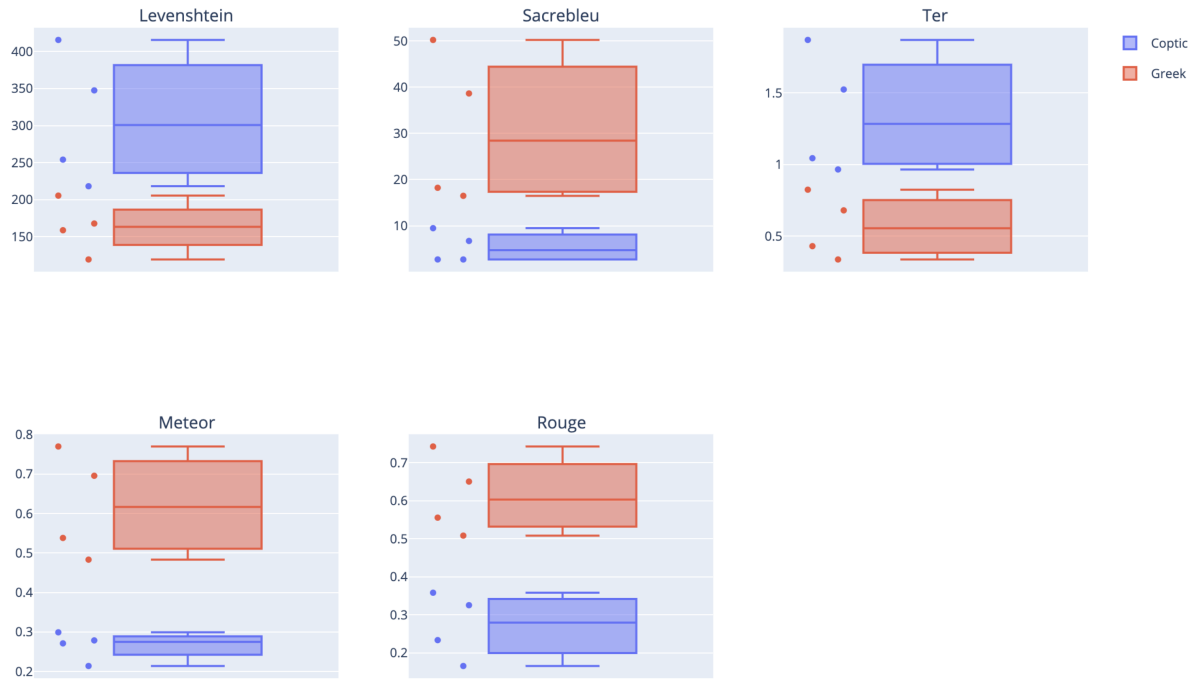

Figure 4: Distribution of mean raw metrics

gual Digital Papyrology will benefit from the semi-automatic generation of metadata (e.g., translations, summaries) enabled by these technologies. The potential to link and integrate previously monolingual datasets seems to outweigh the risks of hallucination, which can be formally addressed using NLP metrics like those employed in this study.

Limitations

One of the main limitations to scaling up the use of large language models (LLMs) or their fine-tuning for specific tasks is the considerable cost associated with these processes. The financial burden includes the expense of the high-performance GPUs required for learning and inference, as well as the cost of access to state-of-the-art models, often hidden behind paywalls. In addition, access to the user interfaces of several advanced LLMs usually entails additional costs, making comprehensive evaluations across multiple models a financial challenge. These barriers can prevent researchers and smaller institutions from fully exploiting the possibilities offered by LLMs, potentially limiting the diversity and breadth of research in this field.

Ethics Statement

While our discussion focuses on improvements in Machine Translation (MT), we emphasize the importance of supporting human-human collaboration in scientific undertakings. The development of MT should not overshadow the important role of human translators and experts in the translation process. We advocate not only a "human-in-the-loop" approach, where human oversight and collaboration are essential to ensure the accuracy and reliability of translation. We also stress that a responsible and ethical MT technology must be human-centered.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Josephine Balmer. 2009. Jumping their bones: Translating, transgressing and creating. *Living Classics: Greece and Rome in Contemporary Poetry in English*, pages 43–64.

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, Michigan. Association for Computational Linguistics.

Anne Boud'hors and Esther Garel. 2019. Ten Coptic Ostraca at the IFAO. *Bulletin de l'Institut français d'archéologie orientale (BIFAO)*, (119):51–77.

Peter Burke. 2017. *Popular culture in early modern Europe*. Routledge.

Sarah J Clackson. 2004. Papyrology and the utilization of coptic sources. In *Papyrology and the history of early Islamic Egypt*, pages 21–44. Brill.

Sonja Dahlgren. 2018. Outcome of language contact: Transfer of Egyptian phonological features onto Greek in Graeco-Roman Egypt. *Journal of Greek Linguistics*, 18(1):155–165. Publisher: Brill.

Maxim Enis and Andrew Megalaa. 2024. Ancient voices, modern technology:low-resource neural machine translation for coptic texts.

Federico Gaspari. 2024. The History of Translation Technologies. In *The Routledge Handbook of the History of Translation Studies*. Routledge. Num Pages: 15.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dean P. Lockwood. 1918. Two Thousand Years of Latin Translation from the Greek. *Transactions and Proceedings of the American Philological Association*, 49:115–129. Publisher: [Johns Hopkins University Press, American Philological Association].

Outi Merisalo. 2015. Translating the Classics into the vernacular in sixteenth-century Italy. *Renaissance Studies*, 29(1).

Holger Nørgaard. 1958. Translations of the Classics into English before 1600. *The Review of English Studies*, 9(34):164–172. Publisher: Oxford University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

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, Belgium, Brussels. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Nicola Reggiani. 2017. *Digital Papyrology I*. De Gruyter, Berlin, Boston.

Daniel Riaño Rufilanchas. 2024. Counting the number of words in greek and latin papyri.

Matthew Snover, Bonnie Dorr, Richard 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.

Peter van Minnen. 1993. The Century of Papyrology (1892-1992). *The Bulletin of the American Society of Papyrologists*, 30(1/2):5–18. Publisher: American Society of Papyrologists.

Marja Vierros and Erik Henriksson. 2017. Preprocessing Greek Papyri for linguistic annotation. *Journal of Data Mining & Digital Humanities*, (Towards a Digital Ecosystem: NLP. Corpus infrastructure. Methods for Retrieving Texts and Computing Text Similarities). Publisher: Episciences. org.

Jennifer Westerfeld. 2016. Decipherment and Translation: An Egyptological Perspective. *CR: The New Centennial Review*, 16(1):29–36. Publisher: Michigan State University Press.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.