# Echoes of Antiquity: Towards Understanding History through Human and LLM-Based Classical Text Translations

Phillip Benjamin Ströbel[1], and Felix Klaus Maier[1]

[1] Department of History, University of Zurich, Zurich, Switzerland

**Abstract**

This paper presents a systematic, data-driven comparison of human and large language model (LLM) translations of Ancient Greek texts, focusing on historiography and epic poetry. We assemble a parallel corpus of Greek source texts and multiple English translations, including both established human translations and new outputs from state-of-the-art LLMs such as GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. Using a unified computational workflow, we evaluate translations via lexical diversity, part-of-speech distributions, collocation profiles, and automatic translation quality metrics (BLEU, ROUGE, METEOR, chrF++). Our results reveal clear genre differences and demonstrate that, while LLMs can approach the fidelity of certain human translators and often reproduce dominant translation patterns, human versions retain greater interpretive diversity and linguistic nuance, especially in poetry. Collocation analyses further show that LLM outputs tend to converge on frequent patterns found in their training data, whereas human translators exhibit both shared conventions and unique, creative solutions. This work-in-progress study highlights both the potential and the present limitations of LLM-driven translation for classical scholarship, providing publicly available materials and quantitative benchmarks for future research in digital classics and translation studies.

**Keywords:** translation studies, Ancient Greek, large language models, digital humanities, digital history

## 1 Introduction

Modern historian Włodzimierz Lengauer controversially concludes his essay by advocating *not* translating ancient texts at all. He contends that "translation gives no idea of the original" because it is impossible to reconstruct the meaning of texts written in languages and cultures that are now "dead" [19, pp. 26–27]. While this claim has sparked debate among scholars, it is rooted in the genuine complexities involved in rendering Ancient Greek into modern languages.[1]

Translators face several interconnected challenges: **(1) Poetic Language:** Ancient Greek poetry, as found in Homer's epics, uses intricate metre and formulaic metaphors (e.g., ῥοδοδάκτυλος Ἠώς, *rosy-fingered dawn*). Conveying such poetic qualities while maintaining accuracy is non-trivial. **(2) Cultural Context:** Concepts like ἀρετή, spanning *excellence*, *virtue*, and *valour*, resist direct equivalence, requiring translators to bridge cultural gaps and provide adequate interpretation. **(3) Historical References:** Frequent allusions require contextual knowledge (e.g., understanding

Phillip Benjamin Ströbel, and Felix Klaus Maier. "Echoes of Antiquity: Towards Understanding History through Human and LLM-Based Classical Text Translations." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 33–49. https://doi.org/10.63744/XcjZ0MxpjIPj.



[1] While Lengauer's examples focus on Polish, similar obstacles arise in translations from Ancient Greek to any modern language. This study focuses on Greek-to-English translation.

| Original | Cawley | Hammond | Mynott | Warner |
|---|---|---|---|---|
| τὰ γὰρ πρὸ αὐτῶν καὶ τὰ ἔτι παλαίτερα σαφῶς μὲν εὑρεῖν διὰ χρόνου πλῆθος ἀδύνατα ἦν, ἐκ δὲ τεκμηρίων ὧν ἐπὶ μακρότατον σκοποῦντί μοι πιστεῦσαι ξυμβαίνει οὐ μεγάλα νομίζω γενέσθαι οὔτε κατὰ τοὺς πολέμους οὔτε ἐς τὰ ἄλλα. | For though the events of remote antiquity, and even those that more immediately precede the war, could not from lapse of time be clearly ascertained, yet the evidences which an inquiry carried as far back as was practicable leads me to trust, all point to the conclusion that there was nothing on a great scale, either in war or in other matters. | Accurate research into earlier or yet more ancient history was impossible given the great gap of time, but I have enquired as far into the past as I can, and on the evidence which I can trust I think there was nothing then on a large scale, either in wars or in anything else. | In respect of the preceding period and the still remoter past, the length of time that has elapsed made it impossible to ascertain clearly what happened; but from the evidence I find I can trust in pushing my enquiries back as far as possible, I judge that earlier events were not on the same scale, either as regards their wars or in other respects. | For though I have found it impossible, because of its remoteness in time, to acquire a really precise knowledge of the distant past or even of the history preceding our own period, yet, after looking back into it as far as I can, all the evidence leads me to conclude that these periods were not great periods either in warfare or in anything else. |

**Table 1:** Original and four translations of Thuc. Hist. I, 1,3.

| Original | GPT-4o | Gemini 1.5 Pro | Claude 3.5 Sonnet |
|---|---|---|---|
| τὰ γὰρ πρὸ αὐτῶν καὶ τὰ ἔτι παλαίτερα σαφῶς μὲν εὑρεῖν διὰ χρόνου πλῆθος ἀδύνατα ἦν, ἐκ δὲ τεκμηρίων ὧν ἐπὶ μακρότατον σκοποῦντί μοι πιστεῦσαι ξυμβαίνει οὐ μεγάλα νομίζω γενέσθαι οὔτε κατὰ τοὺς πολέμους οὔτε ἐς τὰ ἄλλα. | For it was impossible to find clearly the events before them and those even older due to the length of time, but from the evidence that I have examined as far back as possible, I conclude that they were not significant either in terms of wars or in other respects. | For the events of an earlier date, and those still more remote, the lapse of time makes it impossible to know with certainty; but from such evidence as I have been able to obtain by careful inquiry, I think that they were not very important, either in the wars themselves, or otherwise. | For though the events of remote antiquity, and even those that more immediately preceded the war, could not from lapse of time be clearly ascertained, yet the evidences which an inquiry carried as far back as was practicable leads me to trust, all point to the conclusion that there was nothing on a great scale, either in war or in other matters. |

**Table 2:** Translations of Thuc. Hist. I, 1,3 by GPT-4o, Gemini 1.5 Pro and Claude 3.5 Sonnet.

the Peloponnesian War in Thucydides). **(4) Ambiguity and Wordplay:** Ambiguities in the source often elicit diverse interpretive choices, challenging translation fidelity.

Lengauer's "non-translation" solution is radical and, although intellectually rigorous, impractical, given the loss of access for most readers. We illustrate persistent translation challenges by a single sentence from Thucydides' *Historiae* (Table 1), where translators diverge on authorial voice and temporal phrases and none produces an identical outcome. This underscores that every translation involves choice and interpretation.

Such diversity is central in translation theory: Benjamin [5] argues that translations should let the original "echo" in the target language, yet achieving perfect equivalence is impossible. Derrida [7] extends this view, emphasising transformation and creative agency, while Jakobson [17] frames translation as interpretive recoding.

Despite these challenges, new translations continue to appear. Homer's *Odyssey* alone has over 70 English versions [8], none of which are identical, highlighting how translators' backgrounds and exposure to previous versions influence the outcome [1]. Translation thus becomes "re-telling."

Large Language Models (LLMs) are increasingly used for translation. Their outputs, shaped by training data and non-determinism, differ widely; e.g., Claude's output matches Cawley's, while GPT-4o and Gemini blend features from various translations (see Table 2). This research examines such variants across historiography and epic, probing how LLMs might replicate, recombine, or reinterpret ancient texts. At the end of this research (not this article), we aim to answer the question: What translation best captures the ancient mind?

### 1.1 Our contribution

This article presents work-in-progress for systematically comparing and aligning multiple translations of Ancient Greek, combining human editions with outputs from state-of-the-art LLMs. Our aim is (1) to assemble a parallel corpus that enables like-for-like analysis across genres (historiography and epic) and translator traditions, and (2) to develop a measurement toolkit (see Section 4.3) to quantify where renderings converge or diverge.

Conceptually, we treat translation as a measurable act of historical interpretation. Differences between renderings are not noise but signals, i.e., operational traces of reception that reveal how

ancient conceptual categories are preserved, shifted, or re-weighted in modern language. Hence, we approximate what we call the "echo" of the ancient mind by the set of semantic and stylistic constraints that persist across translators and systems. The quantitative results, therefore, do not replace close reading; they prioritise it by flagging loci of interpretive pressure where qualitative analysis is most informative.

This paper contributes three elements toward that broader programme: **(i)** a reproducible corpus and processing pipeline; **(ii)** a comparative battery of metrics sensitive to genre and style; and **(iii)** preliminary evidence that LLM translations cluster around specific human exemplars, suggesting training data imprint or stylistic convergence.

Looking ahead, we will extend the corpus (languages, genres), test memorisation versus generalisation, and formalise an "interpretive distance" index that integrates semantic similarity with controlled concept inventories (e.g., ritual, polity, kinship). The goal is not to declare a single best translation, but to characterise how interpretive space is shaped. Following Martindale's reception-oriented view, we treat translation as a historically situated act of interpretation [21].

## 2 Related Work

Sun and Li [28] comprehensively chart the transformation of literary translation research via digital humanities (DH) methodologies. They argue that the integration of corpus-based and computational approaches replaced earlier, subjective paradigms with systematic and large-scale investigations. Drawing on foundational work by Baker [2], Sun and Li detail how empirically defined phenomena (such as explicitation, simplification, and normalisation) are now observable and quantifiable across parallel and comparable corpora. These developments, aided by advances in text mining, visualisation, and network analysis, have allowed researchers to detect both macro-level trends in translation activity and micro-level stylistic "thumbprints" of individual translators. Their review stresses the value of combining close reading with computational analysis in order to interpret observed variation, establishing a methodological baseline for our comparative work on classical and LLM-generated translations.

### 2.1 Automatic Translation of Ancient Texts

Recent advances in LLMs have redefined the state-of-the-art in ancient language translation. Wannaz and Miyagawa [38] benchmarked leading LLMs, including GPT-4o, Claude Opus, and a domain-specific Coptic Translator, showing that these models can generate English translations for Ancient Greek and Coptic *ostraca* that approach expert-level reliability, especially for Greek due to richer training resources. However, their findings emphasise disparities between languages, performance issues for low-resourced cases, and differences in hallucination rates, translation consistency, and genre sensitivity.

Chamali [6] investigates challenges even for state-of-the-art models by analysing translations of Modern Greek slang and idioms. By creating parallel datasets and evaluating LLMs alongside an NMT baseline, she finds that model performance remains weak for culturally embedded and colloquial language, with diverse error types and informativeness failures, demonstrating the enduring limits of LLMs for morphologically rich and under-resourced languages.

Tekgürler [29] broadens the scope, showing how LLMs approach the translation of 18$^{th}$-century Ottoman Turkish but also highlighting a new set of challenges: contemporary AI content moderation can redact or distort source material, especially with references to violence or sensitive events, raising concerns about the completeness and reliability of automatic historical translation.

Further validation comes from Volk et al. [36], who show that LLMs can produce translations and summaries of 16$^{th}$-century Latin texts close in quality to published human equivalents.

Additionally, the Krikri model [27] has significantly advanced Ancient-to-Modern Greek translation, achieving high benchmark scores through fine-tuning and robust pre-processing. Research increasingly uses LLMs in digital humanities pipelines and comparative translation databases across Greek, Latin, and related languages [37]. Projects in this space combine LLMs, retrieval-augmented generation [22], and hybrid methodologies, supporting digital annotation tools, reading assistance, and deeper stylistic or semantic mapping. The literature consistently highlights the ongoing importance of expert review in validating model outputs and addressing risks such as bias, omissions, and issues with fragmentary or culturally complex historical texts.

### 2.2 Translation Comparisons

Translation comparison studies have gained momentum thanks to digital methods and the increased use of parallel corpora. Palladino et al. [23; 24] present an approach that blends manual word-level alignment with computational metrics to assess translation variation between English and Persian renditions of Ancient Greek sources. Their methodology quantifies overlaps, part-of-speech (POS) correspondence, and alignment categories, enabling precise tracking of where translators choose expansion, omission, or adaptation and showing that indirect (e.g., non-source-language) translations are especially variable.

Parallel corpora and digital concordancers have consequently become central to translation comparison and stylistic analysis. Barlow [4] and others demonstrate how specialised software supports fine-grained comparison of multiple translations, making translator choices, terminology consistency, and stylistic divergence traceable at sentence- and phrase-level. Such studies not only establish rigorous empirical standards for translation research but also contribute to translator education and training. The latest projects expand the scope to direct and indirect translations, systematically examine individual translator styles, and analyse strategies across diverse genres and historical settings, moving the field towards transparent, replicable accounts of translation variation and choice.

## 3 Data

We introduce the edition of the source texts and the translations[2] used in this study, which are analysed with the help of LLMs (see Section 4). Rather than detailing each translation individually,[3] we provide an overview in tabular form, summarising the major editions of Homer's *Odyssey* and Thucydides' *Historiae* across different periods and approaches.

The selected translations span a broad range of publication dates, stylistic approaches, and translational philosophies. Differences arise in the translators' degree of literalness, their handling of narrative voice, and their strategies for rendering cultural concepts: some editions emphasise fidelity to metrical structure or archaic diction (e.g., Wilson's [30], which provides a conversion from the dactylic hexameter to the iambic pentameter, which is more "natural" to English), while others favour accessibility and contemporary language. Such choices lead to variation in the treatment of poetic epithets, authorial interventions, and culturally embedded terms, yielding distinct interpretive outcomes from the same Greek source material.

## 4 Methodology

Our methodology encompasses a comparative analysis of English translations of Ancient Greek texts, produced by both human translators and LLMs. Our workflow ensures consistency, replica-

[2] The year next to the translator's name indicates when the translation was first published.
[3] Key characteristics can be drawn from translators' prefaces and notes.

| **Homer's *Odyssey*** [10] | **Thucydides' *Historiae*** [32] |
|---|---|
| Robert Fitzgerald (1961) [13] | Richard Cawley (1914) [31] |
| Richmond Lattimore (1965) [11] | Rex Warner (1954) [33] |
| Robert Fagles (1996) [12] | Martin Hammond (2009) [34] |
| Anthony S. Kline (2004) [14] | Jeremy Mynott (2013) [35] |
| Emily Wilson (2017) [15] | |
| Peter Green (2018) [16] | |

**Table 3:** Summary of selected English translations analysed for Homer's *Odyssey* and Thucydides' *Historiae*. Publication years correspond to first editions.

bility, and a quantitative foundation for assessing stylistic, lexical, and syntactic features. All the data and the scripts are available on GitHub.[4]

### 4.1 Source Texts and Translations

- **Ancient Greek Sources:** We conducted linguistic analyses of Greek source texts (sentence splitting, tokenisation, POS tagging, lemmatisation) using `odyCy` [18] and saved the output in the CoNLL format.
- **English Translations:** Both human- and LLM-generated translations (see below) underwent the same preprocessing steps as the Ancient Greek sources, albeit using a different spaCy model. We used the English model `en_core_web_md` and saved the result in CoNLL format.

### 4.2 Large Language Models

For translation, we employed three state-of-the-art models: (1) OpenAI's **GPT-4o** (`gpt-4o`), (2) Google's **Gemini 1.5 Pro** (`gemini-1.5-pro`), and (3) Anthropic's **Claude 3.5 Sonnet** (`claude-3-5-sonnet-20240620`). We used corresponding APIs, batch processing and the same prompts overall to ensure comparability. Firstly, we prompted all LLMs[5] to translate the source texts, yielding three additional translations per LLM.

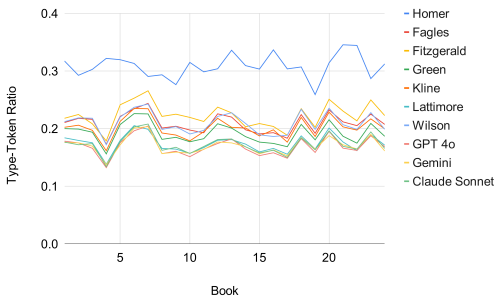
Next, we used Anthropic's Claude 3.5 Sonnet (`claude-3-5-sonnet-20241022`) to align the source and target texts, including the three additional automatic translations.[6] We used the alignment to compute the cosine similarities between source and target and target and target sentences (see below).

### 4.3 Text Processing and Statistical Analysis

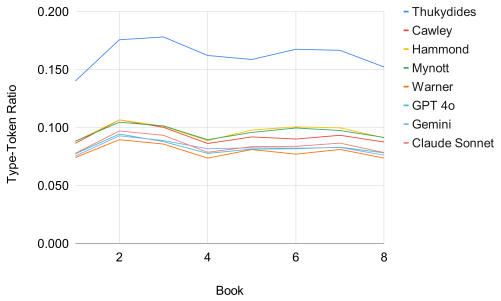
#### *4.3.1 Collocation Analysis*

- **Tooling:** We used the NLTK[7] library to identify and rank frequent bigrams and trigrams.
- **Statistical Association:** We applied pointwise mutual information (PMI), chi-squared, student-t, log-likelihood and frequency counts (with frequency thresholds) to determine salient collocations for each translation.

[4] See `https://github.com/AncientHistory-UZH/echos-of-antiquity`.
[5] See Appendix A.2 for the prompt.
[6] See Appendix A.1 for the prompt.
[7] See `https://www.nltk.org`.




**(a)** Lexical diversity scores of Homer's *Odyssey* and its translations.




**(b)** Lexical diversity scores of Thucydides' *Historiae* and its translations.

**Figure 1:** Lexical diversity (Type–Token Ratio) by book for each source and translation. **x-axis:** Book. **y-axis:** Type–Token Ratio (types/tokens). (Scales differ for Homer vs. Thucydides.)

- **Comparison:** We compared rankings and collocate lists (among the first 1000 collocations each) across LLM and human outputs to assess overlap, divergence, and distinctive phraseological patterns. We used Spearman's $\rho$ and Kendall's $\tau$ to assess differences in the translations.

#### *4.3.2 Translation Comparison Measures*

To quantitatively evaluate the quality and similarity of LLM-generated translations against reference human translations, we employed a set of automatic metrics from the machine translation literature (see Table 4).

| Metric | Description |
|---|---|
| Cosine similarity | Measures the distance between embedded source and target sentences using LaBSE[8] [9]. |
| BLEU [25] | Measures *n*-gram precision between candidate and reference translations, with brevity penalties to control for length differences. |
| ROUGE [20] | Focuses on recall-based overlap of *n*-grams, suitable for evaluating content coverage. |
| METEOR [3] | Incorporates both precision and recall, including stemming, synonymy, and paraphrase matching to capture semantic similarity better. |
| chrF++ [26] | A character *n*-gram based metric combining F-scores for both precision and recall; well-suited for morphologically rich languages and sensitive to minor word-form variation. |

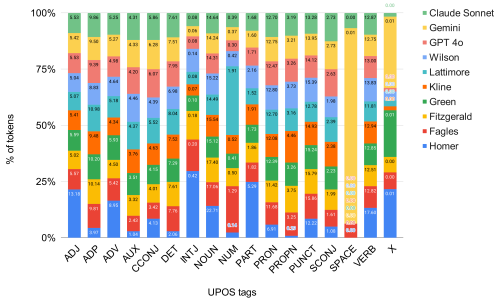
**Table 4:** Automatic translation evaluation metrics used in this study.

## 5 Preliminary Results and Analyses

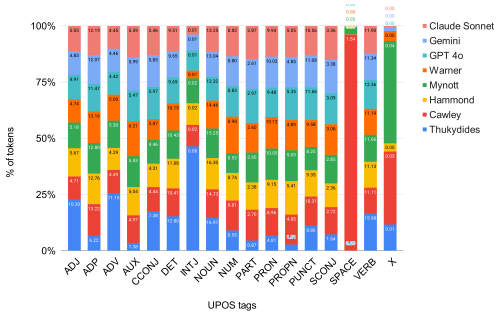
### 5.1 Statistical Analysis of Translations

#### *5.1.1 Lexical Analysis*

Figure 1 shows the Type-Token Ratio (TTR) of the two source texts and their translations. We notice a strong difference between the Ancient Greek sources: Homer's average TTR across all books is 0.31, while Thucydides' is 0.16, indicating Homer's greater vocabulary variation. For




**(a)** POS profile of Homer's *Odyssey* and its translations.




**(b)** POS profile of Thucydides' *Historiae* and its translations.

**Figure 2:** UPOS distribution profiles for source and translations (grouped bar charts). **x-axis**: UPOS tag. **y-axis**: Percentage of tokens.

the translations, we note that the automatically generated translations are always at the bottom. This suggests either a more consistent translation of specific phrases or a limited command of the vocabulary. Pairwise t-tests between human and automatic translation show that the differences of the TTR of the *Odyssey* between GPT-4o, Gemini, Claude Sonnet and Lattimore are not statistically significant. The t-tests for *Historiae* show no significant differences whatsoever. Notably, the TTRs of different translations follow similar patterns (e.g., low in Book 3, high in Book 7), with some peaks and dips also aligning with the source texts (e.g., Book 19).

#### *5.1.2 Part-of-Speech Tag Patterns*

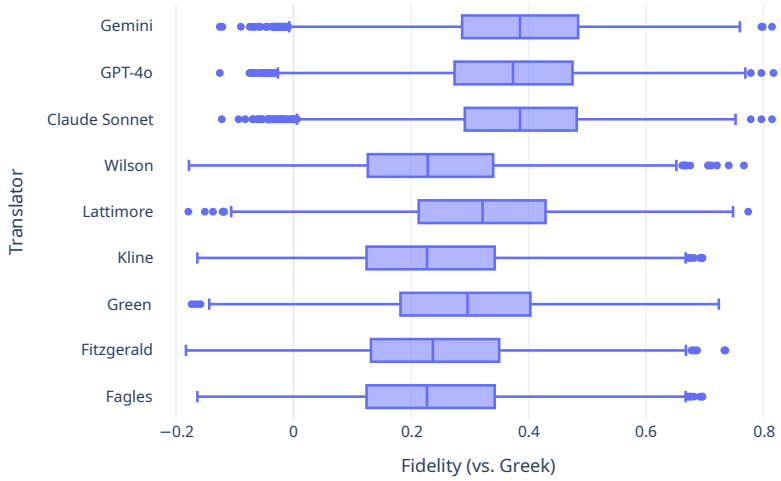
The POS tag profiles in Figure 2 again show a difference between the two genres, especially in the usage of nouns and verbs. Homer's language is distinctively more "noun-heavy" than Thucydides', while the latter's usage of verbs is much higher.

### 5.2 Collocations

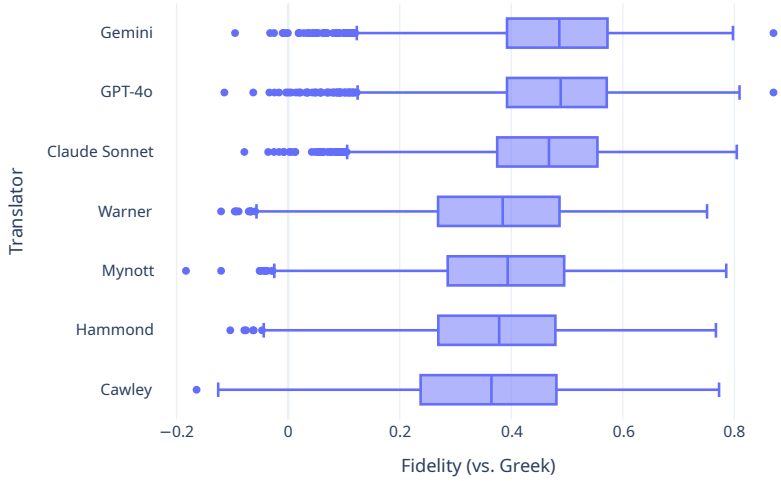
#### *5.2.1 Automatic alignment*

For the preliminary collocation analysis, we refer to Appendix B. Collocation profiles suggest that human translations, despite interpretive variation, often show higher intercorrelation in their phraseological patterns, particularly for established translators of the same era and literary orientation. The p-values of Kendall's $\tau$ analysis can indicate which LLM translations are closely related to specific reference translations. E.g., the trigrams identified in Thucydides with the help of the chi-square measure between GPT-4o and Cawley show a p-value of 0.007, suggesting that the commonality among the first 1000 trigrams in both translations does not occur by chance. As such, some LLMs align more closely with their likely training exemplars, while others exhibit patterns distinct from those of any individual human translator.

Concretely, in Table 2 (Thuc. I.1.3), Claude's rendering tracks Cawley's clause structure and connective choices unusually closely; GPT-4o and Gemini combine elements characteristic of Hammond and Mynott (e.g., treatment of "remote antiquity" vs. "preceding period"). This pattern is consistent with our broader collocation–rank correlations, suggesting that LLMs preferentially reproduce entrenched translation routines. While this does not prove direct memorisation, it supports the hypothesis of stylistic imprint from widely circulated public-domain or widely quoted translations.




**(a)** Distribution of translation fidelities of Homer's *Odyssey*.




**(b)** Distribution of translation fidelities of Thucydides' *Historiae*.

**Figure 3:** Distribution of translation fidelities of the translations. On the x-axis, we see the range of the cosine similarities, while the y-axis shows the respective translations (not the difference in scale (x-axis)).

| Machine Translation | Human Reference | Semantic Similarity | BLEU | ROUGE | METEOR | chrF++ |
|---|---|---|---|---|---|---|
| Claude Sonnet | Fagles | 0.576 | 0.060 | 0.300 | 0.279 | 24.981 |
| | Fitzgerald | 0.511 | 0.055 | 0.250 | 0.250 | 21.824 |
| | Green | 0.685 | 0.099 | 0.392 | 0.383 | 29.435 |
| | Kline | 0.576 | 0.060 | 0.300 | 0.279 | 24.981 |
| | Lattimore | 0.744 | 0.140 | 0.445 | 0.424 | 35.963 |
| | Wilson | 0.554 | 0.066 | 0.296 | 0.304 | 24.884 |
| GPT-4o | Fagles | 0.562 | 0.052 | 0.286 | 0.265 | 23.759 |
| | Fitzgerald | 0.500 | 0.045 | 0.240 | 0.239 | 20.761 |
| | Green | 0.666 | 0.083 | 0.367 | 0.355 | 27.398 |
| | Kline | 0.562 | 0.052 | 0.286 | 0.265 | 23.759 |
| | Lattimore | 0.729 | 0.123 | 0.426 | 0.401 | 34.000 |
| | Wilson | 0.546 | 0.060 | 0.289 | 0.297 | 24.127 |
| Gemini | Fagles | 0.567 | 0.054 | 0.287 | 0.268 | 24.180 |
| | Fitzgerald | 0.504 | 0.049 | 0.244 | 0.243 | 21.261 |
| | Green | 0.674 | 0.089 | 0.375 | 0.366 | 28.270 |
| | Kline | 0.567 | 0.054 | 0.287 | 0.268 | 24.180 |
| | Lattimore | 0.736 | 0.131 | 0.430 | 0.411 | 34.911 |
| | Wilson | 0.546 | 0.058 | 0.285 | 0.293 | 24.091 |

**Table 5:** Translation measures of Homer's *Odyssey*.

| Machine Translation | Human Reference | Semantic Similarity | BLEU | ROUGE | METEOR | chrF++ |
|---|---|---|---|---|---|---|
| Claude Sonnet | Cawley | 0.668 | 0.144 | 0.374 | 0.379 | 35.73 |
| | Hammond | 0.752 | 0.155 | 0.416 | 0.426 | 40.14 |
| | Mynott | 0.777 | 0.176 | 0.456 | 0.457 | 42.198 |
| | Warner | 0.755 | 0.152 | 0.412 | 0.408 | 40.152 |
| GPT-4o | Cawley | 0.646 | 0.092 | 0.323 | 0.322 | 31.111 |
| | Hammond | 0.729 | 0.118 | 0.374 | 0.373 | 36.138 |
| | Mynott | 0.756 | 0.129 | 0.411 | 0.398 | 37.598 |
| | Warner | 0.729 | 0.108 | 0.363 | 0.351 | 35.34 |
| Gemini | Cawley | 0.661 | 0.125 | 0.349 | 0.363 | 34.598 |
| | Hammond | 0.737 | 0.138 | 0.387 | 0.401 | 38.653 |
| | Mynott | 0.764 | 0.154 | 0.423 | 0.429 | 40.297 |
| | Warner | 0.74 | 0.136 | 0.382 | 0.385 | 38.592 |

**Table 6:** Translation measures for Thucydides' *Historiae*.

### *5.2.2 Evaluation of Translation*

Figure 3 shows the box plots of cosine similarity scores between each sentence of the source and target texts in the embedding space created by LaBSE. As with the TTR and POS tag profiles, we observe differences between Homer and Thucydides. Plots (a) and (b) also show that the similarities between the source and automatically generated translations are greater. Since LaBSE was trained on the Common Crawl corpus and Wikipedia, it is possible that the Ancient Greek source texts and translations in the public domain were in the training data.

Tables 5 and 6 show the general evaluation according to the measures introduced in Section 4.3.2. The semantic similarity score indicates which human translation is closest to the automatic translation. For Homer, Lattimore's translation is, on average, the most similar to all the automatic translations. We must thus hypothesise that Lattimore's text has been included in the training data of all three LLMs. For Thucydides, this is the case for Hammond, Mynott, and Warner. The overall lower BLEU scores for the *Odyssey* when compared to the *Historiae* again emphasise the genre difference. It appears that translating poetry yields more diverse results. The remaining scores correlate well with the semantic similarity and BLEU scores.

As concerns further interpretation, we note that BLEU/ROUGE (*n*-gram overlap) emphasise surface correspondence. In contrast, METEOR and chrF++ better tolerate paraphrase and morphological variation, and the sentence-embedding similarity foregrounds semantic proximity. The consistent proximity of LLM outputs to Lattimore (Homer) and to Hammond/Mynott/Warner (Thucydides) indicates convergence on dominant rendering patterns rather than uniformly higher "quality." In other words, alignment seems stylistic-semantic: models mirror phrasing conventions that likely occur in pre-existing translations, while maintaining adequate semantic coverage. This explains why poetry (Odyssey) shows relatively lower overlap scores than prose (Historiae): meter and figurative density invite broader paraphrastic space, where models track conventional solutions but rarely expand the interpretive palette beyond them.

## 6 Conclusion

Our findings demonstrate that both human experts and state-of-the-art LLMs produce translations of Ancient Greek that are systematically distinct in measurable ways, particularly in style, lexicon, and fidelity to the source. Lexical diversity analyses confirm significant differences between poetry and historiography, with Homer's original demonstrating a vastly richer vocabulary than Thucydides, and these patterns are reliably reflected in both human and machine translations. POS distributions further reveal genre-specific syntactic tendencies and highlight the degree to which translators (and LLMs) reproduce or diverge from the source structure.

Automatic evaluation metrics indicate that LLM outputs are often closest to specific human translations (notably Lattimore for Homer), suggesting that the training data overlap or algorithmic convergence occurs on a dominant translation style. Yet, the lower scores for poetry versus prose also confirm that genre affects translation diversity and model performance. Our results suggest that while LLMs can approximate established translation choices, they may reinforce conventional renderings rather than innovate or meaningfully expand interpretive space. Finally, the combined analyses point to both the promise and clear constraints of LLM-based translation for complex, literary source material: machine outputs are consistent and sometimes convergent, but accurate interpretative diversity still emerges more strongly among human translators.

We therefore understand fidelity not merely as textual correctness but as the preservation and transformation of ancient conceptual categories. By quantifying where renderings converge (stylistic imprint, genre constraints) and where they diverge (interpretive breadth, cultural framing), we approximate an "echo" of the ancient mind, i.e., those durable constraints that persist across modern horizons. The next step is integrative: combine these quantitative maps with targeted close readings and controlled concept inventories to model interpretive distance explicitly. Rather than seeking a single correct translation, we reconstruct how antiquity is being understood: which parts resonate stably and which are reshaped in reception.

## Limitations

This research is limited by the selection of a small number of source texts and translations, which focus solely on Ancient Greek historiography and epic. Evaluation is restricted to English outputs, and the LLMs tested are limited to the leading available models at the time of writing. The alignment assessment was only partially automated and used a subset of translation pairs; broader validation will be required for generalisable conclusions. Finally, automatic evaluation metrics may not fully capture literary subtlety or deep interpretive fidelity.

## Ethical Considerations

We adhere to the ethical guidelines of both our institution and the conference. All primary texts are in the public domain. Translations used are either public domain or used in limited scope under academic fair use/fair dealing exceptions. No copyrighted material was shared beyond permissible bounds with external systems. No sensitive personal data is present. Automated analyses were transparently documented for reproducibility, and all generated LLM outputs were critically evaluated to avoid propagation of misleading or culturally insensitive material. We further acknowledge the potential for LLM tools to introduce historical or cultural biases and flag this as an ongoing area of concern and scrutiny.

## References


[1] Armstrong, Richard Hamilton. "Translating Ancient Epic". In: *A Companion to Ancient Epic*. Wiley Online Library, 2005, pp. 174–195.

[2] Baker, Mona. "Corpus Linguistics and Translation Studies: Implications and Applications". In: *Text and Technology*. John Benjamins, 1993, pp. 233–250. URL: https://www.jbe-platform.com/content/books/9789027285874-z.64.15bak.

[3] Banerjee, Satanjeev and Lavie, Alon. "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*, ed. by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss. 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[4] Barlow, Michael. "Parallel Concordancing and Translation". In: *Proceedings of Translating and the Computer 26*. 2004. URL: https://aclanthology.org/2004.tc-1.9/.

[5] Benjamin, Walter. "Die Aufgabe des Übersetzers". In: *Gesammelte Schrifte*, ed. by Tillmann Rexroth. XYZ: Suhrkamp, 1972, pp. 9–21.

[6] Chamali, Irene. "It's All Greek to Them: Challenges in Translating Greek Slang and Idioms via LLMs and NMT". 2025.

[7] Derrida, Jacques. "Des Tours de Babel". In: *Differences in Translation*, ed. by Joseph F. Graham. Ithaca and London: Cornell University Press, 1987.

[8] Durantine, Peter. "The Challenge of Translating 'The Odyssey'". Accessed: July 18, 2025. 2018. URL: https://www.fandm.edu/stories/the-challenge-of-translating-the-odyssey.html.

[9] Feng, Fangxiaoyu, Yang, Yinfei, Cer, Daniel, Arivazhagan, Naveen, and Wang, Wei. "Language-agnostic BERT Sentence Embedding". 2022. arXiv: 2007.01852 [cs.CL]. URL: https://arxiv.org/abs/2007.01852.

[10] Homer. *The Odyssey*. Vol. 1-2. Cambridge, MA and London: Harvard University Press and William Heinemann, Ltd., 1919.

[11] Homer. *The Odyssey*. 1967. Trans. by Richmond Lattimore as *The Odyssey of Homer* (Pymble: Harper Collins e-books).

[12] Homer. *The Odyssey*. 1996. Trans. by Robert Fagles as *The Odyssey* (London: Penguin Books).

[13] Homer. *The Odyssey*. 1998. Trans. by Robert Fitzgerald as *The Odyssey* (New York: Farrar, Strauss and Giroux).

[14] Homer. *The Odyssey*. 2004. Trans. by Anthony. S. Kline as *The Odyssey* (Poetry in Translation).

[15] Homer. *The Odyssey*. 2018. Trans. by Emily Wilson as *The Odyssey* (New York: Norton & Company).

[16] Homer. *The Odyssey*. 2018. Trans. by Peter Green as *The Odyssey* (Oakland: University of California Press).

[17] Jakobson, Roman. “On Linguistic Aspects of Translation”. In: *Harvard University Press* (1959).

[18] Kostkan, Jan, Kardos, Márton, Mortensen, Jacob Palle Bliddal, and Nielbo, Kristoffer Laigaard. “OdyCy – A general-purpose NLP pipeline for Ancient Greek”. In: *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, ed. by Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 128–134. DOI: 10.18653/v1/2023.latechclfl-1.14. URL: https://aclanthology.org/2023.latechclfl-1.14/ (visited on 10/04/2025).

[19] Lengauer, Włodzimierz. “Traduttore traditore: Is It Possible to Translate Ancient Greek Texts?” In: *Przekładaniec Issues in English* (2013), pp. 15–27. DOI: 10.4467/16891864ePC.13.034.1451.

[20] Lin, Chin-Yew. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[21] Martindale, Charles. *Redeeming the Text: Latin Poetry and the Hermeneutics of Reception*. Cambridge University Press, 1993.

[22] Miyagawa, So. “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*, ed. by Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar. 2025, pp. 33–40. DOI: 10.18653/v1/2025.nlp4dh-1.4.

[23] Palladino, Chiara, Shamsian, Farnoosh, and Yousef, Tariq. “Using Parallel Corpora to Evaluate Translations of Ancient Greek Literary Texts. An Application of Text Alignment for Digital Philology Research”. In: *Journal of Computational Literary Studies 1*, no. 1 (2022). DOI: https://doi.org/10.48694/jcls.100.

[24] Palladino, Chiara, Shamsian, Farnoosh, Yousef, Tariq, Wright, David J., Ferreira, Anise d’Orange, and Reis, Michel Ferreira dos. “Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards”. In: *Journal of Open Humanities Data* (2023). DOI: 10.5334/johd.131.

[25] Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[26] Popović, Maja. “chrF++: words helping character n-grams”. In: *Proceedings of the Second Conference on Machine Translation*. 2017, pp. 612–618. DOI: 10.18653/v1/W17-4770.

[27] Roussis, Dimitris, Voukoutis, Leon, Paraskevopoulos, Georgios, Sofianopoulos, Sokratis, Prokopidis, Prokopis, Papavasileiou, Vassilis, Katsamanis, Athanasios, Piperidis, Stelios, and Katsouros, Vassilis. “Krikri: Advancing Open Large Language Models for Greek”. 2025. arXiv: 2505.13772 [cs.CL]. URL: https://arxiv.org/abs/2505.13772.

[28] Sun, Yifeng and Li, Dechao. “Digital Humanities Approaches to Literary Translation”. In: *Comparative Literature Studies 57*, no. 4 (2020), pp. 640–654. URL: https://www.jstor.org/stable/10.5325/complitstudies.57.4.0640 (visited on 07/18/2025).

[29] Tekgurler, Merve. “LLMs for Translation: Historical, Low-Resourced Languages and Contemporary AI Models”. 2025. arXiv: 2503.11898 [cs.CL]. URL: https://arxiv.org/abs/2503.11898.

[30] *The Odyssey*. Trans. by Emily Wilson. New York: Norton & Company.

[31] Thucydides. *Historiae*. 1914. Trans. by Richard Cawley as *History of the Peloponnesian War* (London: J. M. Dent & Sons Ltd.).

[32] Thucydides. *Historiae*. Vol. 1-2. Oxford: Oxford University Press, 1942.

[33] Thucydides. *Historiae*. 1954. Trans. by Rex Warner as *History of the Peloponnesian War* (London: Penguin Books).

[34] Thucydides. *Historiae*. 2009. Trans. by Martin Hammond as *The Peloponnesian War* (Oxford: Oxford University Press).

[35] Thucydides. *Historiae*. 2013. Trans. by Jeremy Mynott as *The War of the Peloponnesians and the Athenians* (Cambridge: Cambridge University Press).

[36] Volk, Martin, Fischer, Dominic Philipp, Fischer, Lukas, Scheurer, Patricia, and Ströbel, Phillip Benjamin. “LLM-based Machine Translation and Summarization for Latin”. In: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, ed. by Rachele Sprugnoli and Marco Passarotti. May 2024, pp. 122–128. URL: https://aclanthology.org/2024.lt4hala-1.15/.

[37] Voukoutis, Leon, Roussis, Dimitris, Paraskevopoulos, Georgios, Sofianopoulos, Sokratis, Prokopidis, Prokopis, Papavasileiou, Vassilis, Katsamanis, Athanasios, Piperidis, Stelios, and Katsouros, Vassilis. “Meltemi: The first open Large Language Model for Greek”. 2024. arXiv: 2407.20743 [cs.CL]. URL: https://arxiv.org/abs/2407.20743.

[38] Wannaz, Audric-Charles and Miyagawa, So. “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*, ed. by Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni. 2024, pp. 463–471. DOI: 10.18653/v1/2024.nlp4dh-1.44.

## A Prompts

Listing 1: Prompt generated by Claude to align the Ancient Greek text segments to the translations. The placeholders {xml_content} and {chapter} have been replaced with the respective XML parts during batch processing.

```
You will be given two inputs: an XML file containing an ancient
   Greek text split into sentences, and a text file containing
   an English translation of the same text. Your task is to
   align the sentences from the ancient Greek text to the
   corresponding parts of the English translation.
Here is the Greek XML input:

<greek_xml>{xml_content}</greek_xml>
```

```
Here is the English translation:

<english_translation>{chapter}</english_translation>.

The alignment can be 1-to-many, i.e., one Greek sentence can
   correspond to several English sentences. Follow these steps
   to complete the task:
1. Preprocess the inputs:
a. Parse the Greek XML to extract individual sentences and their
    IDs.
b. Split the English translation into sentences.
2. For each Greek sentence:
a. Identify the corresponding sentences in the English
   translation.
b. Use a combination of the following techniques to find the
   best matching English sentence(s):
- Length-based matching (compare the number of words)
- Key term matching (look for proper nouns, numbers, or other
   distinctive words)
- Semantic similarity (use context to determine if the content
   matches)
c. If a single Greek sentence aligns with multiple English
   sentences, group them together.
d. If multiple Greek sentences align with a single English
   sentence, keep them grouped with that English sentence.
3. Record the alignments in the following format
<alignment confidence="1.0">
<greek_sentence id="[ID from XML]">[Greek sentence]</
   greek_sentence>
<english_translation>[Corresponding English sentence(s)]</
   english_translation>
</alignment>
4. Here are examples of good and bad alignments:
Good alignment:
<alignment confidence="1.0">  <greek_sentence id="1.1">
Bad alignment:
<alignment confidence="0.1">  <greek_sentence id="1.2">

Handle these edge cases as follows:
1. If a Greek sentence is split across multiple XML elements,
   combine them before aligning.
2. If the English translation contains footnotes or explanatory
   notes, align them separately and mark them as notes:
<note>[Note content]</note>
3. If there are quotes or dialogue in either text, ensure they
   are aligned properly, maintaining the structure of the
   conversation.

Provide your alignments in the order they appear in the original
```

```
 Greek text. Be thorough and precise in your alignments, and
use the tentative and unaligned tags when necessary. Add a
confidence attribute to each alignment, showing how confident
 you are about the alignment on a scale from 0 to 1. Just
output the resulting XML nicely formatted without any
additional comments. Process the whole file without asking to
 proceed with more.
```

Listing 2: Prompt generated by Claude to align the Ancient Greek text segments to the translations. The placeholder `{xml_content}` has been replaced with the respective XML parts during batch processing.

```
You are tasked with translating an XML document containing
   Ancient Greek text into English while preserving the original
    XML structure. Here is the input XML:

<greek_xml>{xml_content}</greek_xml>

Follow these steps to complete the translation:
1. Process each <sentence> element within the XML structure.
2. For each <sentence> element:
a. Extract the Ancient Greek text.
b. Translate the text into English to the best of your ability.
c. Replace the original Greek text with your English translation
   .
3. Maintain the exact XML structure, including all div elements,
    their IDs, and nested sentence elements.
4. Preserve any whitespace or formatting present in the original
    XML.
5. Do not alter, add, or remove any XML tags or attributes.
6. After processing all sentences, output the entire translated
   XML document, maintaining its original structure.

Begin your response with <translated_xml> and end it with </
   translated_xml>. Ensure that the content between these tags
   is a valid XML document with the same structure as the input,
    but with English translations replacing the Greek text in
   each <sentence> element.
```

## B Collocation analysis

These heatmaps show the collocation comparison between translations ( `PMI` = pointwise mutual information, `chi_sq` = chi square).

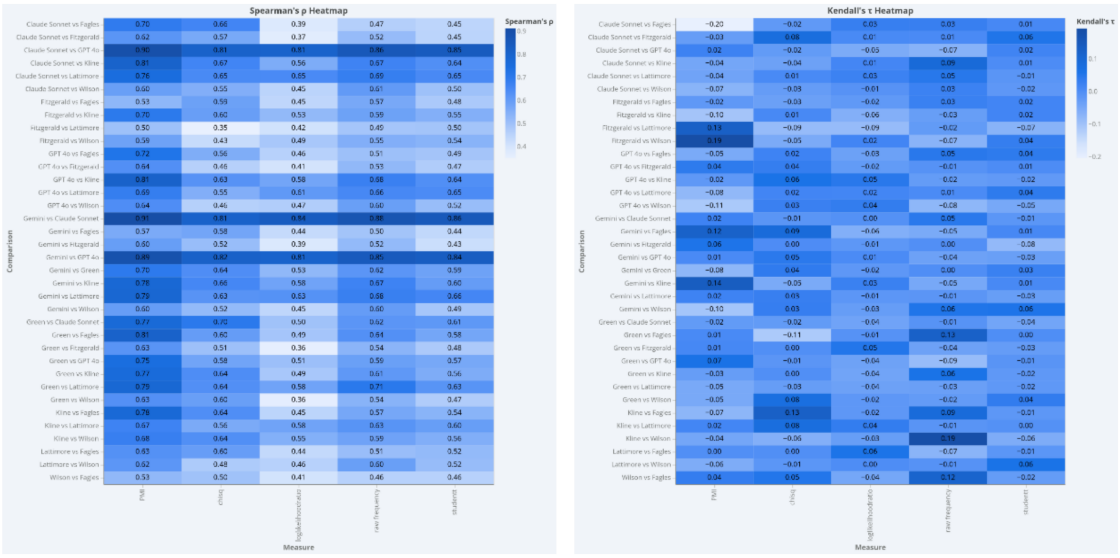



**Figure 4:** Statistics of bigrams of the Homer translations (Spearman's $\rho$ left, Kendall's $\tau$ right).

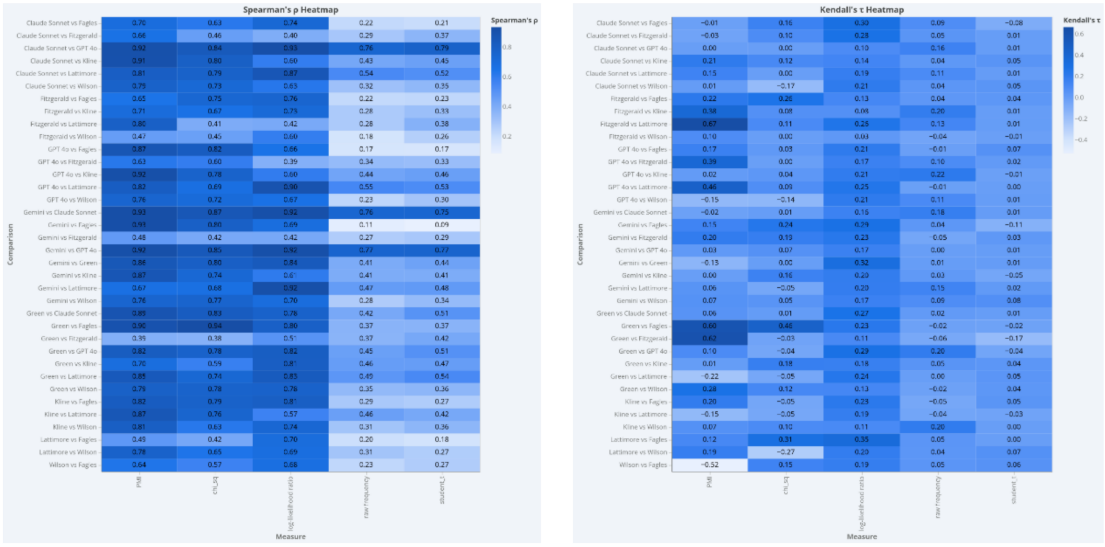



**Figure 5:** Statistics of trigrams of the Homer translations (Spearman's $\rho$ left, Kendall's $\tau$ right).

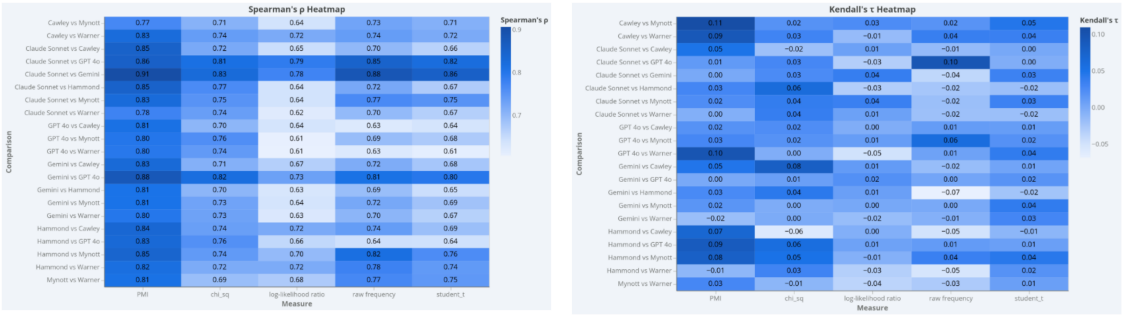



**Figure 6:** Statistics of bigrams of the Thucydides translations (Spearman's $\rho$ left, Kendall's $\tau$ right).

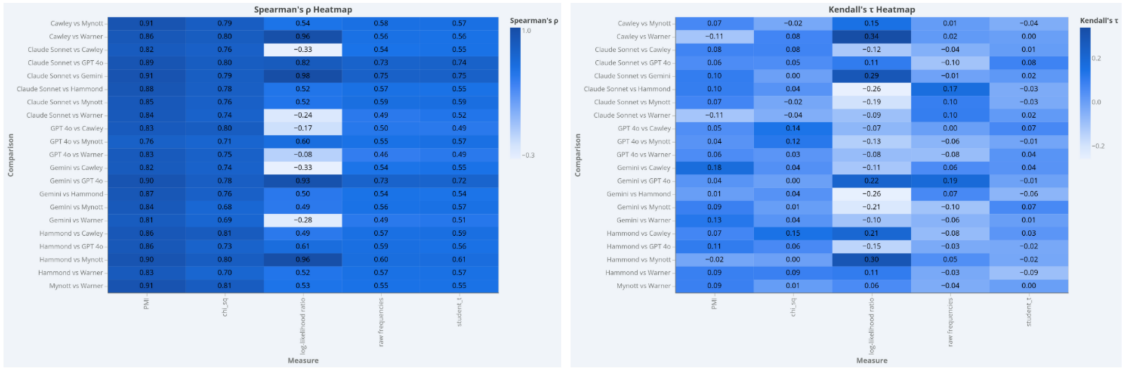



**Figure 7:** Statistics of trigrams of the Thucydides translations (Spearman's $\rho$ left, Kendall's $\tau$ right).