Computational Humanities Research

www.cambridge.org/chr

Evaluating large language models with a word-level translation alignment task between Ancient Greek and English

Peter M. Nadel[1] and Gregory Crane[2]

[1]Research Technology, Tufts University, Medford, MA, USA and [2]Classical Studies, Tufts University, Medford, MA, USA

Short Article

Cite this article: Nadel Peter M. and Gregory Crane. 2025. "Evaluating large language models with a word-level translation alignment task between Ancient Greek and English" *Computational Humanities Research*, **1**:e12, https://doi.org/10.1017/chr.2025.10010

Received: 31 January 2025
Revised: 15 August 2025
Accepted: 19 August 2025

Keywords:
artificial intelligence; digital classics; large language models; translation alignment

Corresponding author: Peter M. Nadel; Email: peter.nadel@tufts.edu

Abstract

In this article, we evaluate several large language models (LLMs) on a word-level translation alignment task between Ancient Greek and English. Comparing model performance to a human gold standard, we examine the performance of four different LLMs, two open-weight and two proprietary. We then take the best-performing model and generate examples of word-level alignments for further finetuning of the open-weight models. We observe significant improvement of open-weight models due to finetuning on synthetic data. These findings suggest that open-weight models, though not able to perform a certain task themselves, can be bolstered through finetuning to achieve impressive results. We believe that this work can help inform the development of more such tools in the digital classics and the computational humanities at large.

Plain Language Summary

We propose a new method for aligning translations at the word-level. Using large language models (LLMs), we take several aligned sentences of Ancient Greek and align each word to its modern English counterpart. We evaluate four different LLMs, two proprietary: Claude 3.5 Sonnet and GPT-4o and two open-weight: Llama 3.3 8B and Llama 3.3 70B. Impressed by the results of Claude 3.5 Sonnet, we used it to generate further examples of word-alignment data so that we could then finetune the open-weight models. After doing so, we reevaluated these open-weight models and found a significant increase in their ability to perform this word-level alignment task.

We hope that these findings will give other researchers in the computational humanities more confidence in using synthetic data to finetune open-weight models for specific tasks. Too, we advocate for the release and collation of these synthetic data into a multipurpose dataset for instructional finetuning for common tasks in the computational humanities.

Introduction

Translation alignment is the task of finding semantic or syntactic counterparts between a source text and its translation. It can be performed at different degrees of textual organization, from the document or page levels to the paragraph, sentence and word levels. In this article, we evaluate the ability of autoregressive decoder-only large language models (LLMs) to perform word-level translation alignment between Ancient Greek and modern English.

Word-level translation alignment is an important task in the burgeoning study of computational humanities. Text alignments of this nature can be used by a variety of researchers from across several humanistic disciplines, from those supporting language learners to scholars normally working with texts purely in translation (Aker et al. 2014; Palladino, Foradi, and Yousef 2021; Shi, Zettlemoyer, and Wang 2021). Especially when integrated into multilingual digital editions (MDEs), word-level translation alignments have the opportunity to enrich the experience of any reader (Levchenko 2024).

Previous work

Translation alignment is an important task in natural language processing (NLP) and dates back to Brown et al.'s seminal paper on statistical machine translation. However, their motivation was much different than ours, seeking to develop a method of machine translation which used statistical features gathered from large corpora of pre-aligned sentences to predict translations in a pair of languages (Brown et al. 1993). Building on this work, Och and Ney introduced Giza++, an open-source translation toolkit able to train the models that Brown et al. proposed (Och and Ney 2003). This paradigm of statistical machine translation continued in the next decade,

Dawn [Ἠὼς], out of[ἐκ] her[0] bed[λεχέων] from beside[παρ᾽] noble[ἀγαυοῦ] Tithonus[Τιθωνοῖο],

Figure 1. An example of a single line alignment (Od. 5.1). English is aligned to Ancient Greek with the use of square brackets. A *0* was used when there was no direct alignment between the English and the Ancient Greek.

with Dyer et al.'s "fast_align" (Dyer, Chahuneau, and Smith 2013) and Östling and Tiedemann's EfLoMAI (Östling and Tiedemann 2016).

Machine translation, and therefore translation alignment, took on new dimensions with the advent of the transformer architecture, which sought to simplify recurrent or convolutional neural networks by connecting encoder and decoder modules through a novel attention mechanism (Vaswani et al. 2023). Indeed, the most recent studies depart from statistical machine translation alignment and opt to exploit the internal states of encoder modules in transformer models, sometimes known as word embeddings, to align words in source translation sentence pairs. Sabet et al. (Sabet et al. 2021) and Dou and Neubig (Dou and Neubig 2021) employ contextualized and fine-tuned word embeddings to conduct word-level alignment, while Garg et al. (Garg et al. 2019) and Chen et al. (Chen et al. 2020) examine the attention mechanism between the encoder and decoder of the transformer model.

Yousef et al. (Yousef et al. 2022) continued these efforts into the fields of digital classics and the digital humanities with a method that fine-tuned an encoder-only multilingual transformer model on Ancient Greek and Latin sources. In this environment, where languages are generally lacking in resources, like massive publicly available datasets, translation alignment is an especially difficult task, yet all the more valuable to researchers and language learners. Presenting an alignment workflow utilizing contextualized word embeddings, Yousef et al. were able to achieve impressive results on word-level alignment between Ancient Greek and Latin with a pre-trained encoder-only transformer, fine-tuned from XLM-RoBERTa.

In another landmark paper in the field, Yousef et al. (Yousef et al. 2022) demonstrate that the same techniques used to align Latin to Ancient Greek could also be used to align modern languages to Ancient Greek, taking English and Portuguese as examples. Again, Yousef et al. achieve impressive results while comparing the ability of five algorithms for extracting alignment pairs from contextual word embedding similarity scores.

Both of these studies utilized datasets gleaned from the Ugarit Translation Alignment Editor, compiled by Palladino et al. (Palladino et al. 2023). These datasets, with alignments from Latin to Ancient Greek, Portuguese to Ancient Greek and English to Ancient Greek, constitute an invaluable gold standard, critical in developing automated alignment tools.

Since these developments, autoregressive, decoder-only architectures, which power AI systems like ChatGPT, have proven to be skilled at a variety of tasks, especially in machine translation and the parsing of historical texts. However, an evaluation of word-level alignment between a historical language and a modern one has never been conducted with an autoregressive, decoder-only architecture, only with encoder-only architectures. Below, we explore the successes and challenges of doing so and identify relevant lessons which can be applied to other language pairs or other tasks in the computational humanities.

Methodology

Rather than employing encoder-only models to conduct word-level alignment as in Yousef et al., we sought to use autoregressive decoder-only language models to do so. These models are trained to respond to and perform certain user-designed tasks, known as prompts. We selected a set of open-source and proprietary models and evaluated their performance on the word-level alignment task. Based on these results, we then generated new, synthetic alignment data, upon which we finetuned the open-source models and reevaluated them. All prompts used to perform this evaluation and later improvement of open-source models are provided in our additional material.

We first needed a stable set of word-level alignments which we knew were correct so that we could determine the accuracy of the automated systems. To this end, the Ugarit alignments provided by Palladino et al. (Palladino, Shamsian, and Yousef 2022; Palladino et al. 2023) proved invaluable to this study. The quality of these alignments was measured through Inter-Annotator-Agreement (IAA), meaning that after two annotators aligned the same sentence, Palladino et al. calculated the overlap between their alignments. The larger the overlap, the greater agreement between the annotators, the more consistent and more reliable the alignments are. For this dataset, Palladino et al. report an IAA score of 86.08% for Ancient Greek to English alignments (Palladino, Shamsian, and Yousef 2022; Palladino et al. 2023).

There was one drawback, however, to using this dataset as a gold standard. The alignments follow the standard adopted by the North American Chapter of the Association for Computational Linguistics (NAACL), in which indices from source and translation texts are separated by a dash and followed by either an S, to indicate a "sure" alignment or a P, to indicate a "possible" alignment. This format is very useful as it concisely describes the entire alignment, allowing it to be shared and distributed with ease. For autoregressive LLMs, though, this format is very challenging. In early experiments, we found that these models could not consistently render alignments in this format, perhaps to do underlying issues in how their tokenizers render index numbers. To surmount this challenge, we needed to introduce an intermediary format, in which the words themselves were substituted for their index values, with the source word placed in brackets next to the translated words (see Figure 1). In the prompt, we call this a "custom format." We then utilized a simple script which took the gold standard from Palladino et al. and converted to this custom format. All metrics below are taken from a subsequent evaluation.

Additionally, LLM performance can vary given the amount of context or examples the model is given in the prompt (Wei et al. 2023). In addition to evaluating a particular model's ability to perform the alignment task, we also wanted to test how many examples of correct alignments we could give to the model before we saw a degradation of performance. As is convention, we denote prompts with no examples as "zero-shot," and prompts with examples as "few-shot." We divided the gold standard alignment into two sets, one which we could test on and the other from which we would randomly select a number of examples to populate preexisting

prompts. Below, we evaluate the models at the 1-shot, 2-shot, 3-shot, 4-shot and 5-shot levels. Zero-shot results were not reported as the model output was too variable to ensure that the correct format was followed.

Once an initial evaluation was complete, we chose the model which performed the best to generate synthetic data upon which we could further finetune the open-weight models, hypothesizing that this new data may improve model performance. Synthetic data has been shown to greatly augment smaller models and their ability to perform specific tasks (Kaddour and Liu 2024), though we were unsure whether or not word-level alignment would follow this pattern. We chose an unpublished sentence-aligned edition of Xenophon's *Anabasis* and passed each sentence pair to the best-performing model to generate new word-level alignments. We then employed the *axolotl* framework to perform parameter-efficient finetuning (PEFT) on the open-weight models (Lian 2024; Xu et al. 2023). This software package provides utilities for finetuning across several GPUs, allowing us to leverage the power of the Tufts University High Performance Compute (HPC) Cluster.

Results

To evaluate model performance on word-level alignment tasks, we divided attempts to align words into three categories: correct, wrong and partial. A correct alignment represents a predicted alignment which matched the gold standard exactly and a wrong alignment represents a predicted alignment which completely diverged from the gold standard. A partial alignment was one where the predicted alignment resembled the gold standard to some extent. Take, for example, the following aligned sentences (Od. 5.101):

"And there is not any city of mortals nearby..."
"οὐδέ τις ἄγχι βροτῶν πόλις…".

Many of the models chose to align *οὐδέ* to "And," while the gold standard aligned *δέ* to "And" and *οὐ* to "not." This alignment would be considered partial, as it represents the model's ability to align part of a word or phrase correctly, but could not be classified as correct as it does not match directly with the gold standard. We decided to capture this nuance in our evaluation because, in some cases, these partial alignments would still tell readers something about the relationship between the source and translation texts. Indeed, regarding these partial alignments as wrong, we contend, would have skewed our results just as much as counting them as correct. As in the example given above, there is some ambiguity in the way in which we have asked the LLMs to align the text. *οὐδέ* is a single word and therefore ought to be aligned to another single word, like "And." The model may not be aware that breaking up *οὐδέ* is allowed in the task, despite seeing examples of it in the prompt. A further exploration of this point can be found in the Discussion section.

Thankfully, the NAACL standard supports this kind of ambiguity with their "possible" tag, and as a result, we were able to use the same metrics of precision, recall, F1 score and alignment error rate (AER) as Yousef et al. (Yousef et al. 2022), with precision defined as

$$\frac{|A \cap P|}{|A|}.$$

Recall is defined as

$$\frac{|A \cap S|}{|S|}.$$

F1 score is defined as

$$\frac{2 * precision * recall}{precision + recall}.$$

AER is defined as

$$1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|},$$

where

$$A$$

represents the set of alignments predicted by the model,

$$P$$

represents the "possible" alignments and

$$S$$

represents the "sure" alignments.

The table below describes the results obtained for word-level alignment across four different LLMs, two open-weight and two proprietary. For ease of comparison, we have chosen to report F1 and AER (Table 1).

As described above, we used the best-performing model, Claude 3.5 Sonnet, to generate a thousand examples of correct alignments from a sentence-aligned edition of Xenophon's *Anabasis*, costing approximately $15 from Claude's API service. This dataset could then be used to finetune the open-weight models on this specific task. Find in Table 2 the results of these finetuned models for word-level alignment. Too, the F1 scores

Table 1. F1 scores before finetuning compared to Yousef et al. (Yousef et al. 2022). Bold values indicate the highest score.

Model	1-shot	2-shot	3-shot	4-shot	5-shot
	F1	F1	F1	F1	F1
meta-llama/Meta-Llama-3.3-8B-Instruct	0.0000	**0.0128**	0.0000	0.0000	0.0000
meta-llama/Meta-Llama-3.3-70B-Instruct	0.0056	0.0051	0.0061	0.0000	**0.0089**
GPT-4o	0.7780	0.7758	0.8010	0.7846	**0.8212**
Claude 3.5 Sonnet	0.8451	0.8438	**0.8502**	0.8432	0.8452
Alignment method	Softmax	Entmax15	Match	Argmax	Itermax
Yousef et al. Experiment 6 with XLM-R	0.7768	0.7582	**0.8150**	0.7495	0.7990

Table 2. AER scores before finetuning compared to Yousef et al. (Yousef et al. 2022). Bold values indicate the highest score.

Model	1-shot AER	2-shot AER	3-shot AER	4-shot AER	5-shot AER
meta-llama/Meta-Llama-3.3-8B-Instruct	1.000	**0.9872**	1.000	1.000	1.000
meta-llama/Meta-Llama-3.3-70B-Instruct	0.9944	0.9949	0.9939	1.000	**0.9911**
GPT-4o	0.2220	0.2242	0.1990	0.2125	**0.1788**
Claude 3.5 Sonnet	0.1549	0.1562	**0.1498**	0.1568	0.1548
Alignment method	Softmax	Entmax15	Match	Argmax	Itermax
Yousef et al. Experiment 6 with XLM-R	0.2189	0.2369	**0.1872**	0.2454	0.1973

Table 3. F1 scores of open-weight models after finetuning. Bold values indicate the highest score.

Model	1-shot F1	2-shot F1	3-shot F1	4-shot F1	5-shot F1
meta-llama/Meta-Llama-3.3-8B-Instruct	0.7722	0.7721	0.1563	0.7797	**0.7903**
meta-llama/Meta-Llama-3.3-70B-Instruct	0.7951	0.8171	0.8233	0.8288	**0.8288**

Table 4. AER scores of open-weight models after finetuning. Bold values indicate the highest score.

Model	1-shot AER	2-shot AER	3-shot AER	4-shot AER	5-shot AER
meta-llama/Meta-Llama-3.3-8B-Instruct	0.2278	0.2279	0.8437	0.2203	**0.2097**
meta-llama/Meta-Llama-3.3-70B-Instruct	0.2049	0.1829	0.1767	0.1712	**0.1712**

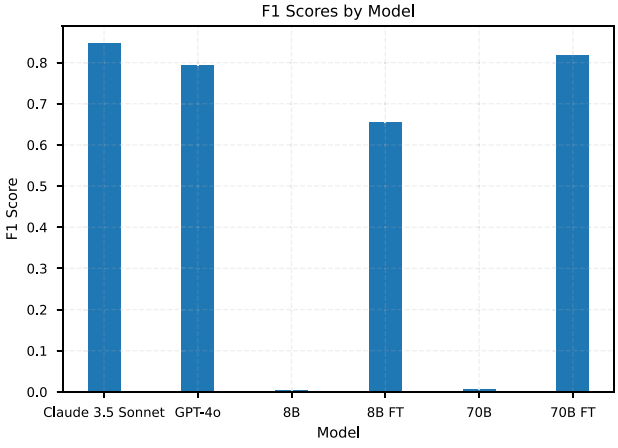

Figure 2. Comparison of model performance across different shot counts, showing the relationship between precision and recall scores. The diagonal dashed line represents equal precision and recall scores. Finetuned models (lighter colors) generally show improved performance over their base versions (darker colors).

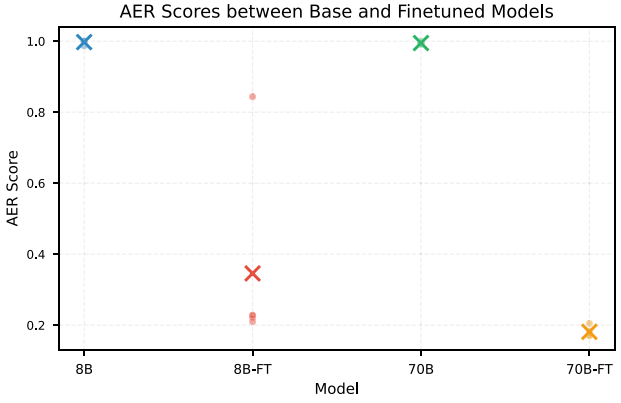

Figure 3. Distribution of performance between base and finetuned Llama models, across different shots. Finetuned models show significant improvement from their model counterparts.

after finetuning show dramatic improvement (Table 3), with corresponding improvements in AER scores (Table 4).

Plotting the F1 scores in Figure 2 helps us to understand how these models perform relative to each other. We see that while Llama 3.3 8B-Instruct and Llama 3.3 70B-Instruct lag significantly behind both GPT-4o and Claude 3.5 Sonnet, their finetuned counterparts excel and reach performance comparable to these proprietary models.

These findings are confirmed by Figure 3, which tells us that the finetuned models outstrip their base models, even as we control for the number of examples contained in the prompts. These results confirm the hypothesis mentioned above, that synthetic

data produced by Claude 3.5 Sonnet can greatly improve smaller model performance even on difficult tasks like word-level translation alignment in a historical language. Too, note the outlier we see in the AER Score of the finetuned Llama 3.3 8B-Instruct. We explore the reasons for this anomaly in our Discussion section.

Discussion

The ramifications of this study are twofold. First, we show that LLMs can produce useful, though not perfect, word-level alignments between Ancient Greek source texts and their modern English translations, especially when prompts contain multiple examples. Second, we illustrate that the production of synthetic data, and finetuning thereon, can significantly improve even small models and make open-weight models competitive with even the largest proprietary ones.

Alignment error analysis

As for many other tasks, Claude 3.5 Sonnet excels at the word-level alignment task. While this may be expected from more general evaluations (OpenRouter 2025), we were surprised with how poorly, in comparison, GPT-4o performed. Though this model is usually quite strong, especially in multilingual contexts, we find that it is distinctly worse at the word-level alignment task than Claude 3.5 Sonnet. Much more work should be pursued in evaluating and comparing these models for historical or under-resourced language tasks.

Though Claude 3.5 Sonnet achieved impressive results, it behooves us to examine what kinds of mistakes it makes as compared to the human gold standard. Taking the 5-shot results, of the 138 possible gold standard alignments, 64 were evaluated as wrong or partial. Of these, 27 alignments were to Ancient Greek adverbs and conjunctions like ἀλλὰ, μάλα and ἔνθα, not to mention notoriously difficult to translate words like δέ and τε. Another 11 mistakes were aligned to Ancient Greek determiners or articles, all variants of ὅς and ὁ. These categories of words can pose challenges to even human annotators, let alone an automated system. Often these mistakes mixed up words with similar meanings or senses, as in the example below (Od. 5.103):

"But after all, it is not at all possible that the design of aegis-holding Zeus&"

"ἀλλὰ μάλ᾽ οὔ πως ἔστι Διὸς νόον αἰγιόχοιο…"

Claude 3.5 Sonnet incorrectly aligns the whole phrase "But after all" to ἀλλὰ alone rather than in the gold standard ἀλλὰ μάλ᾽. As a result, it then aligns μάλ᾽ to "at all" and πως to "possible." This alignment, though deviating from the meaning of the original passage and ultimately incorrect, follows word senses common to Ancient Greek as described in Liddell Scott Jones, with ἀλλὰ often being translated with an entire English phrase, μάλ᾽ typically strengthening assertions as the alignment of this line suggests, and πως used "sometimes merely to qualify [a phrase's] force, when it cannot be always rendered by any one English equivalent" (Liddell and Scott 1940). We hypothesize that the mistakes that we see in the example above may be related to the autoregressive nature of the LLM. As the model is only able to predict the next token linearly, it lacks any ability to look backwards and fix anything in an earlier part of its prediction. When ἀλλὰ is linked to "But after all," for example, the model cannot go back and add μάλ᾽ once it realizes that this word is part of a larger, already aligned, phrase, as a human might. Instead, it must align it to whatever remains of the English sentence.

This limitation could be alleviated in several different ways. One method in particular has become popular for dealing with similar obstacles in other disciplines. Using an agent-based or agentic approach, we could, rather than using just one instance of the model, exploit a system of model calls, which would query the model several times per each sentence pair at each stage moving closer to a final alignment (Wu et al. 2023). For example, one agent could first provide a morphosyntactic parsing of the Greek and English and then pass this information to another agent that would compose a preliminary alignment. All of this information could then be handed off to a third agent, whose job it would be to edit the alignment, looking out for overlapping word senses as described above. A last agent could focus on reformatting all of this work into a single, valid alignment. Though the original problem, that a model is unable to revise its work as a human might, remains, we can use successive model runs to iterate and improve on the result. That said, this approach is much more costly when it comes to computational resources, as the model could be called up to three extra times compared to the original. Further work in this direction should be pursued, but with a close eye on how accuracy changes with respect to cost-per-token.

Finetuning open-weight models

Despite the error analysis above, the results from Claude 3.5 Sonnet were sufficient for us to experiment with finetuning open-weight models on synthetic data it generated. This practice is common in other fields, but the computation humanities lag in the adoption of this technique. We hope that our results will inspire confidence in this approach. To that end, all of the material used to finetune our models, including data and configuration files, will be released alongside this publication.

Despite our goal to encourage the use of synthetic data, at the same time, it would be a mistake not to mention that Claude 3.5 Sonnet's results were far from perfect. We explored common errors above, even suggesting that they may be the same errors that humans would make on the same data. That said, finetuning on this "silver" quality data will perpetuate these same errors that we already see in the error analysis above. Indeed, our finetuned models merely mimic the same errors as Claude 3.5 Sonnet, errors which for the time being are tolerable, but errors all the same. One potential augmentation to this approach would be to add human intervention and oversight to the Claude-generated alignments. Though creating alignment is often time-consuming for human experts, revising and validating existing alignments is much less so. While not trivial for human annotators either, amending alignments which are mostly (around 85%) correct is a much less taxing task than composing them from scratch. Re-entering human expertise into the workflow then would take advantage of Claude 3.5 Sonnet's ability in parsing ancient languages, while also improving its output before finetuning. We hypothesize that the resulting finetuned models would surpass Claude 3.5 Sonnet, while still providing the additional benefits of a more parsimonious and secure model.

Advantages

We see that finetuning, in both cases, improved AER score, for Llama 3.3 8B Instruct by about .75 and for Llama 3.3 70B Instruct

Table 5. Standard deviation on the best scores of proprietary models and finetuned open-weight models compared to Yousef et al.

Model	Best precision STD	Best recall STD	Best F1 STD	Best AER STD
Claude 3.5 Sonnet	0.1242	0.1187	0.1065	0.1065
GPT-4o	0.1942	0.1861	0.1787	0.1787
meta-llama/Meta-Llama-3-8B-Instruct	0.1544	0.1488	0.1388	0.1388
meta-llama/Meta-Llama-3-70B-Instruct	0.1187	0.1168	0.1017	0.1017
Yousef et al.	0.1149	0.1176	0.1098	0.1030

Note: Yousef et al. did not report these numbers in their study, but they were calculated by us for this study.

Table 6. Average time to align one sample, with Yousef et al. far surpassing any decoder-only architecture. Bold values indicate the highest score.

Model	Average time to align (seconds) averaged over 100 samples
Claude 3.5 Sonnet	6.780
GPT-4o	4.690
meta-llama/Meta-Llama-3-8B-Instruct	9.210
meta-llama/Meta-Llama-3-70B-Instruct	29.99
Yousef et al. with Match	**0.1075**

Note: As in Table 5, this figure was not reported in Yousef et al., but was taken from our own testing of the model in Yousef et al.

by about 0.8. Most impressive to us is the stark improvement observed in the smaller 8B variant. Though 70B achieves a higher score, 8B not only approaches GPT-4o, but does so in a size small enough to fit on one Nvidia T4 GPU, which can be accessed for free through a variety of cloud-based compute resources, including Google Colab. Most humanists, computational or otherwise, do not have access to an HPC or the most cutting-edge hardware. Thus, offline LLMs, finetuned for a specific purpose, are often out of the toolkit for most researchers. In this case, however, we are able to provide to researchers a state-of-the-art model which can be easily and efficiently used, along with instructions to use it in the Python notebook provided alongside this publication.

Disadvantages

Compared to Yousef et al., both the proprietary and finetuned models seem to excel. Based solely on the metrics presented above, it would be easy to draw that very simple conclusion. We would be remiss not to mention certain ways, in which our method lags behind that of Yousef et al. First, we note much more variance in F1 and AER in our models compared to Yousef et al. Take, for example, the 3-Shot result in the finetuned Llama 3.3 8B Instruct. The F1 drops to a meager 0.1563, signifying that though this model can for the most part match the results in Yousef et al., it is perhaps less consistent than Yousef et al. In fact, we see this trend in most of the models studied, as shown in Table 5.

Additionally, the time it took for the models in this study to align the text far exceeded the time it took for Yousef et al. This is a serious consideration, especially when put in the context of all translated Greek literature. Again, though our finetuned models produce potentially better results, the amount of time they take to complete this task may be too long for some scholars. See Table 6 for a detailed comparison.

Conclusions

We argue that, while LLMs have the potential to revolutionize humanities research, they should be used in thoughtful and parsimonious ways. To that end, we demonstrate how the creation of a synthetic dataset from a much larger and more expensive model can provide researchers with a free-to-use open-weight alternative that is both more efficient and cheaper.

In this study, we only created a single type of instruction, focusing on translation alignment for finetuning, but synthetic data for other tasks could be included as well. In the future, we hope to build an instructional finetuning datasets for LLM training which revolves around the parsing, translating and aligning ancient languages. This dataset could then be used to enhance the performance of any open-weight LLM on a variety of tasks relevant to digital classics. We encourage similar efforts from other fields and disciplines as the tasks commonly found in publicly accessible instructional finetuning datasets are not sufficient for the tasks in the computational humanities, thus limiting their utility.

If the computational humanities seek to expand their toolkit to include LLMs, researchers must adopt creative and inventive methods for utilizing them. In this article, we have shown one approach for doing so. However, more than just this finding, we hope to inspire other scholars to employ LLMs in new and unexpected ways.

Acknowledgements. We are grateful for the assistance of the Perseus team: Alison Babeu, James Tauber, Charles Pletcher, Cliff Wulfman, Farnoosh Shamsian, Sarah Abowitz and Serguisz Kazmierski.
We also acknowledge the Tufts University High Performance Compute Cluster (https://it.tufts.edu/high-performance-computing) which was utilized for the research reported in this article.

Data availability statement. Below, please find the prompt used to generate the word-level alignments:

```
<title>Word Alignment Task</title>
<description>
    You will be given a sentence in two languages. Your task is to align the
        words in the two sentences in the custom format indicated in the
        examples.
</description>
<notes>
    <note>Align the sentences in the custom format with brackets.</task>
    <note>Be sure to consider the complex relationship that small functional
        words play in the sentence alignment.</task>
    <task>If you are alignment a highly inflected language to a language which
         is not as inflected, be sure to align all of a prepositional phrase
        in the lower inflected language to what may be single word in the
        highly inflected language.</task>
</notes>
<examples>
\t{examples}
</examples>
<output\_format>
    ‘‘‘xml
    <alignment>
        <lang1>[lang1]</lang1>
        <lang2>[lang2]</lang2>
        <custom>[custom]</custom>
    </alignment>
    ‘‘‘
</output\_format>
<to\_align>
    <lang1>{lang1}</lang1>
    <lang2>{lang2}</lang2>
</to\_align>
```

listing 1. Training prompt template.

All code, data and examples can be found on our GitHub repository here: https://github.com/pnadelofficial/word-level-alignment.

Author contributions. Ideas and theoretical details were developed jointly by the authors. Mr. Nadel provided implementation in code as well as managed storage of models. Dr. Crane provided all data used.

Funding Statement. No external funds were used to complete this work. Any costs were payed by the authors.

Competing interests. The authors declare none.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the United States.

Disclosure of use of AI tools. Outside of the use of AI tools in the content of this work, they were also used to help design figures and tables.

References

Aker, Ahmet, Monica Paramita, Marcis Pinnis, and Robert Gaizauskas. 2014. "Bilingual Dictionaries for all EU Languages." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by **Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis**, 2839–45. Reykjavik: European Language Resources Association (ELRA). https://aclanthology.org/L14-1623/

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics* 19, no. 2: 263–311. https://aclanthology.org/J93-2003/.

Chen, Yun, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. "Accurate Word Alignment Induction from Neural Machine Translation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by **Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu**, 566–76, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.42

Dou, Zi-Yi, and Graham Neubig. 2021. "Word Alignment by Fine-tuning Embeddings on Parallel Corpora." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, edited by **Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty**, 2112–28, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.181

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. "A Simple, Fast, and Effective Reparameterization of IBM Model 2." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by **Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff**, 644–48. Atlanta: Association for Computational Linguistics. https://aclanthology.org/N13-1073/

Garg, Sarthak, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. "Jointly Learning to Align and Translate with Transformer Models." Preprint, arXiv:1909.02074. https://arxiv.org/abs/1909.02074

Kaddour, Jean, and Qi Liu (2024). "Synthetic Data Generation in Low-Resource Settings via Fine-Tuning of Large Language Models." Preprint, arXiv:2310.01119. https://arxiv.org/abs/2310.01119

Levchenko, Maria (2024). "Automatic Translation Alignment Pipeline for Multilingual Digital Editions of Literary Works." Preprint, arXiv: 2410.13255. https://arxiv.org/abs/2410.13255

Lian, Wing. 2024. "Axolotl." Version 0.6.0. https://github.com/axolotl-ai-cloud/axolotl

Liddell, Henry George, and Robert Scott (1940). *A Lexicon Abridged from Liddell and Scott's Greek-English Lexicon*. New York: Clarendon Press.

Och, Franz Josef, and Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics* 29, no. 1: 19–51. https://doi.org/10.1162/089120103321337421.

OpenRouter . 2025. "LLM Rankings." https://openrouter.ai/rankings. (visited on 01/21/2025).

Östling, Robert, and Jörg Tiedemann. 2016. "Efficient Word Alignment with Markov Chain Monte Carlo." *The Prague Bulletin of Mathematical Linguistics* 106, 125–146. https://doi.org/10.1515/pralin-2016-0013.

Palladino, Chiara, Maryam Foradi, and Tariq Yousef. 2021. "Translation Alignment for Historical Language Learning: a Case Study." *Digital Humanities Quarterly* 15, no. 3. http://www.digitalhumanities.org/dhq/vol/15/3/000563/000563.html

Palladino, Chiara, Farnoosh Shamsian, and Tariq Yousef. 2022. "Translation Alignment: Ancient Greek to English. Annotation Style Guide and Gold Standard." Version 1.0. Zenodo. https://doi.org/10.5281/zenodo.7362097.

Palladino, Chiara, Farnoosh Shamsian, Tariq Yousef, David J. Wright, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2023. "Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards." *Journal of Open Humanities Data* 9, no. 1: 22. https://doi.org/10.5334/johd.131.

Sabet, Masoud Jalili, Philipp Dufter, François Yvon, and Hinrich Schütze. 2021. "SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings." Preprint, arXiv: 2004.08728. https://arxiv.org/abs/2004.08728.

Shi, Haoyue, Luke Zettlemoyer, and Sida I. Wang. 2021. "Bilingual Lexicon Induction via Unsupervised Bitext Construction and Word Alignment." Preprint, arXiv: 2101.00148. https://arxiv.org/abs/2101.00148.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. "Attention Is All You Need." Preprint, arXiv: 1706.03762. https://arxiv.org/abs/1706.03762.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." Preprint, arXiv: 2201.11903 [cs.CL]. https://arxiv.org/abs/2201.11903

Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang (2023). "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation." Preprint, arXiv: 2308.08155. https://arxiv.org/abs/2308.08155

Xu, Lingling, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang (2023). "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment." Preprint, arXiv: 2312.12148. https://arxiv.org/abs/2312.12148

Yousef, Tariq, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022. "An Automatic Model and Gold Standard for Translation Alignment of Ancient Greek." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, edited by **Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis**, 5894–905. Marseille: European Language Resources Association.

Yousef, Tariq, Chiara Palladino, David J. Wright, and Monica Berti. 2022. "Automatic Translation Alignment for Ancient Greek and Latin." In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, edited by **Rachele Sprugnoli and Marco Passarotti**, 101–7. Marseille: European Language Resources Association. https://aclanthology.org/2022.lt4hala-1.14/.