Ancient Greek to Modern Greek Machine Translation: A Novel Benchmark and Fine-Tuning Experiments on LLMs and NMT Models

Spyridon Mavromatis[*†], **Sokratis Sofianopoulos**[†], **Prokopis Prokopidis**[†], **Maria Giagkou**[†]
[*]National and Kapodistrian University of Athens, Department of Informatics and Telecommunications
[†]Institute for Language and Speech Processing, Athena RC
spyros.mauromatis@athenarc.gr, s_sofian@athenarc.gr, prokopis@athenarc.gr, mgiagkou@athenarc.gr

Abstract

Machine Translation (MT) for Ancient Greek (AG) to Modern Greek (MG) is a low-resource task, constrained by the lack of large-scale, high-quality parallel data. We address this gap by introducing the AG-MG Parallel Corpus, a new resource containing 132,481 sentence-aligned pairs derived from literary, historical, and biblical texts. We present a novel corpus creation pipeline that combines web-scraped, excerpt-level data with a multi-stage sentence-level alignment, and refinement process. Our method uses VecAlign with LaBSE embeddings, which we first fine-tune on a manually-aligned AG-MG subset, followed by an LLM-based error/misalignment correction phase using Gemini 2.5 Flash to ensure high alignment quality. Furthermore, we provide the first comprehensive benchmark of modern MT models on this task, evaluating three fine-tuning strategies across NMT models (NLLB, M2M100) and a Greek LLM (Llama-Krikri-8B). Our experiments show that fine-tuning yields significant improvements over base models, increasing performance by up to +10.3 BLEU points. Specifically, full-parameter fine-tuning of Llama-Krikri-8B achieves the highest overall performance with a BLEU score of 13.16, while the QLoRA-adapted M2M100-1.2B model demonstrates the largest relative gains and highly competitive results. Our dataset and models represent a significant contribution to Greek NLP.

Keywords: Machine Translation, Ancient Greek, Low-Resource Languages

1. Introduction

Machine Translation (MT) has evolved from rule-based and statistical approaches to neural sequence-to-sequence models and is now being reshaped by Large Language Models (LLMs). Recent studies suggest that LLMs, pre-trained on extensive multilingual data, can rival or complement traditional encoder–decoder MT, particularly in settings with limited resources where parallel data is scarce (Lyu et al., 2023). This trend has led to revisiting under-resourced, stylistically rich language pairs such as Ancient Greek (AG) to Modern Greek (MG), where the source side spans multiple dialects, historical periods, and genres.

A critical bottleneck in machine translation of AG→MG is the lack of large, sentence-aligned parallel corpora. Historical resources often exist at excerpt or document level and contain a lot of noise, such as editorial notes, markup, and orthographic variants. Recent work shows that neural embedding–based aligners can substantially improve sentence alignment accuracy and scalability on classical languages. For example, VecAlign[1] combined with multilingual sentence embeddings (e.g., LaBSE[2] or LASER[3]) has been found effective on Ancient Greek and Latin texts, outperforming traditional approaches (e.g., Hunalign) and handling very long documents (Craig et al., 2023; Thompson and Koehn, 2019; Feng et al., 2020; Yousef et al., 2022b). These developments pave the way for the creation of higher-quality AG→MG datasets suitable for MT.

At the modeling level, two families of systems are promising for Greek: (i) multilingual encoder–decoder NMT models such as M2M100 (Fan et al., 2021) and NLLB (Costa-Jussà et al., 2022), and (ii) open Greek LLMs, notably Llama–Krikri[4] , which offer strong Greek fluency and tokenization while remaining accessible for adaptation (Roussis et al., 2025; Voukoutis et al.). This paper presents a high-quality AG→MG sentence-level parallel corpus, as well as fine-tuning NMT models and a Greek LLM for this task.

In this work, we address the resource gap and we provide a systematic evaluation of modern translation models. Our key contributions include the following:

1. We introduce the **AG-MG Parallel Corpus**, the largest sentence-aligned corpus for this low-resource language pair, containing 132,481 high-quality aligned sentence pairs.

[1] https://github.com/thompsonb/vecalign
[2] https://huggingface.co/sentence-transformers/LaBSE
[3] https://github.com/facebookresearch/LASER
[4] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8685–8698
11–16 May 2026.

The corpus is enriched with extensive metadata, including author, title, segment index, translator, ancient Greek dialect, genre, and era. An excerpt/paragraph-level version of the corpus will also be created soon.

2. We present a **novel hybrid alignment pipeline** for creating the corpus. Our method first uses a domain-adapted LaBSE model for initial alignment and then leverages a LLM (Gemini 2.5 Flash developed by Comanici et al., 2025) for misalignment detection and correction, ensuring superior alignment quality.

3. We conduct the **first large-scale benchmark** of both NMT and LLM-based models for AG-to-MG translation. We compare three different fine-tuning strategies (Full-parameter, LoRA, and QLoRA) on NLLB[5], M2M100[6] and Llama-Krikri-8B, providing a clear picture of the current state-of-the-art on this task.

Our results demonstrate that fine-tuned models significantly outperform their base versions (yielding improvements of up to +10.3 BLEU points). The fully fine-tuned Llama-Krikri-8B LLM achieved the highest overall BLEU score (13.16), while the M2M100-1.2B model showed the largest relative gains. This highlights the potential of both specialized LLMs and adapted NMT architectures for this task. We hope that this study will facilitate future research in Greek NLP and Digital Humanities.

2. Related Work

Our work is positioned at the intersection of classical language resource creation and low-resource machine translation.

2.1. Parallel Corpora and Alignment

Although there are monolingual corpora for Ancient Greek, such as the Diorisis corpus (Vatri and McGillivray, 2018), the creation of large-scale parallel corpora for MT is a more recent challenge. Efforts have focused on establishing gold standards (Yousef et al., 2022a; Palladino et al., 2023) and evaluating alignment methods for classical languages (Yousef et al., 2022b; Craig et al., 2023; Yousef et al., 2023; Sommerschield et al., 2023).

A key finding from this line of work is that neural, embedding-based aligners significantly outperform older, traditional methods (e.g., length-based heuristics or tools like Hunalign). Craig et al. (2023) conducted a detailed evaluation of various tools specifically for **sentence-level alignment** on AG and Latin texts. Their results demonstrated that a combination of VecAlign and LaBSE embeddings provides the most accurate sentence alignments, especially for long and literary documents common in classical studies.

2.2. Machine Translation of Ancient Greek

Machine Translation involving Ancient Greek remains a definitive low-resource task. The majority of published research has focused on translating AG to English (AG→EN). Kolovou (2024), for instance, experimented with OpenNMT on an AG→EN dataset, highlighting the challenges of the task. More recently, Rapacz and Smywiński-Pohl (2025) improved neural models for interlinear translation by incorporating morphological information to better handle the rich inflection of Ancient Greek, focusing on translation into English and Polish (AG→EN/PL).

To our knowledge, the AG→MG language pair is significantly less explored and no large-scale benchmark exists that compares modern, high-capacity NMT models (NLLB, M2M100) and new Greek LLMs (Llama-Krikri) for this task (Sommerschield et al., 2023; Papantoniou and Tzitzikas, 2024). This study seeks to address this gap by providing the first comprehensive evaluation of these models on a new, high-quality dataset.

3. The Ancient-Modern Greek Parallel Corpus

We introduce the AG-MG Parallel Corpus, a new sentence-aligned dataset designed for AG→MG machine translation. This section details the data sources, our alignment methodology, and the final corpus characteristics.

3.1. Data Sources

The corpus aggregates parallel texts from three main types of digital resources, aiming for diversity in genre, dialect, and period of Ancient Greek:

- **Digitized Anthologies:** Collections focusing primarily on classical Attic prose (approx. 5th–4th century BC), offering curated AG texts alongside scholarly MG translations.
- **Comprehensive Digital Libraries:** Larger repositories of Ancient Greek literature containing a wider variety of AG dialects (e.g., Homeric/Epic from the 8th century BC, Ionic, Doric, Attic and later Hellenistic Koine) paired with MG translations. While some translations

[5] https://huggingface.co/docs/transformers/en/model_doc/nllb

[6] https://huggingface.co/docs/transformers/en/model_doc/m2m_100

in these sources date to the early 20th century, we strictly selected only contemporary, monotonic Modern Greek versions, explicitly excluding archaic or Katharevousa translations.

- **Biblical Texts Resources:** Digital versions of the Holy Bible providing AG texts (Septuagint/LXX and New Testament, representing Hellenistic Koine from the 3rd century BC to the 6th century AD) aligned at the verse level with contemporary MG translations.

These sources typically provide texts aligned at the excerpt or paragraph level, necessitating the sentence alignment pipeline described below.

3.2. Preprocessing and Alignment Pipeline

To create a high-quality, sentence-aligned corpus, we implemented a multi-stage pipeline:

1. Scraping and Initial Cleaning: We first scraped the public web sources using standard Python libraries (BeautifulSoup[7] and Scrapy[8]), extracting AG texts, MG translations, and available metadata (author, title, translator, etc.) into JSONL format. Basic cleaning removed residual HTML tags and markup.

2. Deep Cleaning and Segmentation: We applied a more thorough cleaning process to remove noise such as page numbers, editorial brackets, translator comments, and inconsistent punctuation. Both AG and MG texts were then segmented into sentences using the Stanza library[9].

3. Fine-Tuned Embedding Alignment: For the non-Bible sources where sentence alignment was needed, we employed VecAlign, following prior work (Craig et al., 2023). Crucially, instead of using off-the-shelf embeddings, we fine-tuned LaBSE (Feng et al., 2020) on **1,000 manually aligned AG-MG sentence pairs**, varying in genre and ancient Greek dialect and extracted from our sources. This domain adaptation step significantly improved the embedding quality for the specific nuances of AG-MG alignment. VecAlign was then run using these fine-tuned embeddings to generate initial sentence alignments. For the Bible texts, alignment was straightforward using the existing verse indices.

4. LLM-Based Refinement: While the fine-tuned VecAlign+LaBSE approach yielded good results, manual inspection revealed residual misalignments, particularly with non-literal translations or sentence splitting/merging. To ensure the highest possible quality of the corpus, we implemented a refinement step using the Gemini 2.5 Flash API (Comanici et al., 2025). We designed prompts[10] that instructed the model to evaluate the semantic equivalence of proposed AG-MG pairs from VecAlign and flag or correct likely misalignments (e.g., 1-to-many, many-to-1, many-to-many, or incorrect 1-to-1 mappings). This automated LLM-based verification proved highly effective in identifying and fixing errors. After applying the LLM refinement, we conducted a careful human evaluation on a subset of 150 aligned pairs and estimated that our dataset contains approximately 95% correct alignments. The 5% of alignment errors are mainly caused by inconsistent MG translations and translator comments.

5. Deduplication and Multi-Reference Handling: Following Lee et al. (2021), we performed deduplication on all splits based on the MG sentences to remove near-identical translation variants that might skew model training. However, drawing inspiration from multi-reference MT training (Zheng et al., 2018; Khayrallah et al., 2020), when sources provided distinct translations from different translators for the same AG sentence, we retained these variants specifically within the training set, enriching the corpus with translation and stylistic variation.

3.3. Data Card and Statistics

The final corpus consists of **132,481** sentences per language (AG→MG), with **2,311,448** AG tokens/words (avg. 17.45 per sentence) and **3,060,949** MG tokens/words (avg. 23.10 per sentence) as presented in Table 1. Each dataset entry preserves metadata for *author, title, segment index, url, translator, ancient Greek dialect, genre, and era* and enables comparisons to excerpt-level variants (kept separately for future excerpt/paragraph-level work).

[7] https://beautiful-soup-4.readthedocs.io/en/latest/

[8] https://scrapy.org/

[9] https://stanfordnlp.github.io/stanza/

[10] *You are a cautious data editor. Your job is to fix alignment errors between Ancient Greek ("grc") and Modern Greek ("ell") sentences inside a JSONL file. Do not translate or paraphrase any text. Only merge/split and reorder within a row while keeping the original order of sentences.*

[11] Tokens (or words) are counted as whitespace-separated items. The exact number of Ancient and Modern Greek sentences may be higher, since aligned sentence pairs can contain multiple sentences per side.

	Ancient Greek	Modern Greek
Sentences	132,481	132,481
Tokens / Words	2,311,448	3,060,949
Avg. tokens per sent.	17.45	23.10

Table 1: Sentence and token statistics for the released sentence-level corpus.[11]

Split	Sentence Pairs
Train	128,231
Dev	2,000
Test (main reporting set)	2,000
Stress (rarer dialects: Ionic, Doric, Homeric)	250
Total	**132,481**

Table 2: Corpus splits for AG→MG experiments.

4. Experimental Setup

We evaluate the effectiveness of our AG-MG Parallel Corpus by fine-tuning several state-of-the-art NMT and LLM models. This section details the dataset splits, models, fine-tuning procedures, and evaluation metrics used.

4.1. Dataset Splits

We split the 132,481 sentence pairs as described in Table 2. The training set comprises 128,231 pairs. We use separate development (Dev) and test sets, each containing 2,000 pairs, randomly sampled while ensuring no overlap with the training set or each other. Additionally, we created a small stress set of 250 pairs specifically containing rarer AG dialects (Ionic, Doric, Homeric) to evaluate model generalization beyond the dominant Attic and Koine dialects. A detailed breakdown of the exact dialectal composition for each of these splits is provided in Appendix B.

4.2. Models

We benchmark two types of models:

- **Multilingual NMT Models:** We selected strong open-source encoder-decoder models known for their ability to handle multiple languages:
  - facebook/nllb-200-distilled-600M (Costa-Jussà et al., 2022)
  - facebook/nllb-200-distilled-1.3B (Costa-Jussà et al., 2022)
  - facebook/m2m100_1.2B (Fan et al., 2021)

Ancient Greek Dialect	Sentences
Attic (Αττική)	75,887
Ionic (Ιωνική)	8,725
Doric (Δωρική)	2,614
Homeric / Epic (Ομηρική, Επική)	6,729
Hellenistic Koine (Ελληνιστική Κοινή)	38,526
Total	**132,481**

Table 3: Distribution of aligned sentence pairs across major Ancient Greek dialects in the corpus.

- **Greek LLM:** We used ilsp/Llama-Krikri-8B-Instruct (Roussis et al., 2025), an open Greek LLM based on the Llama architecture, selected for its strong native Greek fluency.

For each model, we evaluate both its zero-shot (base) performance and its performance after fine-tuning.

4.3. Fine-Tuning Strategies

We employed three fine-tuning strategies to adapt the models to the AG→MG task:

- **Full Fine-Tuning (Full FT):** All model parameters were updated during training. Applied to M2M100-1.2B and Llama-Krikri-8B.
- **LoRA (Low-Rank Adaptation):** (Hu et al., 2022) A parameter-efficient fine-tuning (PEFT) method that injects trainable low-rank matrices into the model layers, freezing the original weights. Applied to NLLB-600M and NLLB-1.3B.
- **QLoRA (Quantized LoRA):** (Dettmers et al., 2023) A more memory-efficient PEFT method combining 4-bit quantization with LoRA. Applied to M2M100-1.2B and Llama-Krikri-8B.

4.4. Training Details

Fine-tuning was performed using either Google Colab Pro (L4 GPU) or the CINECA[12] supercomputing infrastructure (using 1 to 4 NVIDIA A100 64GB GPUs[13]) for the larger models and full fine-tuning runs. Key hyperparameters are summarized in Table 4.

All experiments used the Hugging Face `transformers`[14] library, with `peft`[15] for

[12] https://docs.hpc.cineca.it/index.html

[13] The LEONARDO supercomputer features custom NVIDIA Ampere A100 accelerators equipped with 64GB of HBM2e memory, a specific hardware configuration designed for this EuroHPC system.

[14] https://huggingface.co/docs/transformers

[15] https://huggingface.co/docs/peft

LoRA/QLoRA and `bitsandbytes`[16] for quantization. Training employed AdamW optimizer variants (`paged_adamw_8bit` or `adamw_torch_fused`) with a cosine learning rate scheduler and warmup. We used mixed precision (BF16 on A100s, FP16 on L4). For NLLB and M2M100 models, we expanded the tokenizer vocabulary with 148 (for the NLLB models) and 122 (for the M2M100 models) Ancient Greek characters/tokens missing from the base models and resized the model embeddings accordingly, unfreezing them during PEFT to allow adaptation. The Full FT of Krikri-8B utilized `DeepSpeed ZeRO Stage 3` for distributed training across 4 A100s. Models were trained for 5 epochs, selecting the best checkpoint based on validation loss.

4.4.1. Vocabulary Adaptation and Smart Initialization

A significant challenge in adapting multilingual NMT models like M2M100 and NLLB to Ancient Greek is related to how these models handle Ancient Greek tokens. Ancient Greek utilizes the Polytonic system, featuring diacritics such as *oxia* (acute accent), *varia* (grave accent), *perispomeni* (circumflex accent), *psili* (smooth breathing), *dasia* (rough breathing) and more, which is largely absent from the pre-training data of modern multilingual models. Consequently, standard tokenizers treat many Polytonic characters as unknown tokens (`<unk>`), leading to input fragmentation or complete "input blindness", as observed in the base M2M100 model.

To overcome this, we implemented a robust four-stage vocabulary adaptation pipeline:

1. **Token Discovery:** We scanned the entire Ancient Greek training corpus to identify characters that the base tokenizer could not resolve. This process revealed 122 missing Ancient Greek characters/tokens for the M2M100 and 148 for the NLLB models.

2. **Dictionary Update:** These identified characters were explicitly added to the model's tokenizer, assigning them unique IDs and preventing them from being mapped to `<unk>`.

3. **Embedding Resizing:** We structurally resized the model's input and output embedding matrices to accommodate the newly added token IDs.

4. **Smart Initialization (Weight Transplant):** Initializing new embeddings with random noise can significantly slow down convergence, as the model must learn the semantic value of these characters from scratch. Instead, we employed a "smart initialization" strategy. For each new Polytonic character (e.g., *ᾷ*), we programmatically identified its closest base character in the existing vocabulary (e.g., *α*) by stripping diacritics. We then copied the pre-trained embedding weights of the base character to the new Polytonic token. This effectively provides the model with a head start, allowing it to treat the new character as a variation of a known vowel or consonant rather than a random symbol.

This methodology ensured that fine-tuning focused on learning the syntactic and dialectal nuances of Ancient Greek rather than learning basic character representations, leading to the substantial performance gains observed in our experiments.

It is important to note that this vocabulary adaptation process was applied exclusively to the NMT models (M2M100 and NLLB). The LLM evaluated in this study (Llama-Krikri) utilizes tokenizer architectures with significantly larger vocabularies and byte-level fallback mechanisms (e.g., BPE or TikToken), which inherently handle rare Polytonic characters without requiring explicit vocabulary expansion.

4.5. Evaluation Metrics

We evaluate translation quality using a comprehensive suite of standard metrics:

- **BLEU:** (Papineni et al., 2002) SacreBLEU implementation with `13a` and `intl` tokenization.
- **chrF/chrF++:** (Popović, 2015, 2017) Character n-gram metrics, with chrF++ using word n-grams (`word_order=2`).
- **TER (Translation Edit Rate):** (Snover et al., 2006) Measures the number of edits required to match the reference. Lower is better.
- **BERTScore:** (Zhang et al., 2019) Computes semantic similarity using contextual embeddings (using `xlm-roberta-large`). We report F1.
- **COMET:** (Rei et al., 2020) A neural metric trained to predict human judgments of translation quality (using `Unbabel/wmt22-comet-da`). Higher is better.

[16] https://github.com/bitsandbytes-foundation/bitsandbytes

[18] Calculated using DeepSpeed config: 2 (micro-batch) * 8 (grad accum) * 4 (GPUs).

Model	Adaptation	Quant.	Rank (r)	Eff. Batch	LR	Platform
NLLB-600M	LoRA	8-bit	16	16	2e-4	L4 (Colab)
NLLB-1.3B	LoRA	8-bit	16	12	1e-4	L4 (Colab)
M2M100-1.2B	QLoRA	4-bit NF4	16	12	1e-4	L4 (Colab)
M2M100-1.2B	Full FT	None	N/A	12	1e-4	1xA100 (CINECA)
Krikri-8B-IT	QLoRA	4-bit NF4	32	16	8e-5	1xA100 (CINECA)
Krikri-8B-IT	Full FT	None	N/A	64[18]	2e-5	4xA100 (CINECA, DS-Z3)

Table 4: Key hyperparameters for fine-tuning experiments. Eff. Batch = per-device batch size $\times$ gradient accumulation steps $\times$ num GPUs. LR = Learning Rate. All experiments ran for 5 epochs.

All metrics were calculated using standard implementations, comparing model outputs against the reference translations in the Test and Stress sets.

5. Results and Analysis

We present the evaluation results on the Test and Stress sets, comparing the zero-shot (Base) performance of the pre-trained models against their fine-tuned versions across all metrics.

5.1. Results on Test Set

Table 5 summarizes the performance of all models on the main Test set (2,000 pairs).

The results clearly demonstrate the effectiveness of fine-tuning. All fine-tuned models achieve significant improvements over their base zero-shot counterparts across all metrics, with BLEU scores increasing by up to **+10.3** points for the QLoRA-adapted M2M100 model (using its corresponding base score as reference).

The Greek LLM, Llama-Krikri-8B-Instruct, emerges as the top overall performer. Its base performance (BLEU 8.29) is notably higher than the NMT base models, indicating strong inherent capabilities for Greek-related tasks. After full fine-tuning, it reaches the highest BLEU (13.16) and chrF++ (34.71) across all models. Furthermore, its QLoRA variant achieves the highest COMET score (0.713), suggesting high perceived translation quality.

A striking observation concerns the performance of the base M2M100-1.2B, NLLB-600M and NLLB-1.3B models, which performed very poorly. The base M2M100-1.2B model recorded the lowest BLEU score (0.62) among all experiments. These near-zero performances are not due to poor grammar but rather a complete failure to process the Polytonic script. Since the base tokenizers treat most Ancient Greek characters (e.g., vowels with breathing marks and varia/oxia) as unknown tokens or noise, the models suffer from severe hallucinations. This validates the necessity of our vocabulary expansion and embedding resizing strategy, which unlocked the true potential of the models. While the base zero-shot performance of the M2M100-1.2B model is very low due to its limited exposure to Polytonic Greek, fine-tuning with QLoRA achieves a BLEU of 10.96, the lowest TER (82.99), and the highest BERTScore (0.911), showing a massive relative improvement (+10.34 ΔBLEU). This observation highlights the paramount role of data quality in resource-constrained tasks. It suggests that a high-quality, curated dataset, such as our AGMG corpus, contributes significantly to a fine-tuned model's translation efficacy, effectively mitigating the effects of a less capable base model.

5.2. Results on Stress Set (Rare Dialects)

Table 6 shows performance on the Stress set (250 pairs), designed to test generalization to rarer AG dialects (Ionic, Doric, Homeric).

As expected, performance drops on the Stress set compared to the main Test set for all models, indicating the challenge posed by out-of-distribution dialects.

The Krikri-8B model maintains its leading position. The fully fine-tuned variant achieves the highest BLEU (**12.80**), chrF++ (**35.90**), and lowest TER (**81.40**) on this set. Meanwhile, the Krikri-QLoRA model matches the highest BERTScore F1 (**0.911**) and achieves the highest COMET score (**0.717**). The base Krikri model again starts from a higher baseline (BLEU 6.55) compared to the NMT models on this harder set. This suggests that the pre-training of Krikri might provide better generalization or robustness to dialectal variation within Greek.

The M2M100 models continue to show strong relative gains, with the QLoRA variant achieving a +9.45 BLEU improvement over its base. The NLLB models show more modest results, though the LoRA adaptations still vastly improve upon their near-zero base scores.

Model	Method	BLEU↑	chrF++↑	TER↓	BERTScore F1↑	COMET↑	△BLEU
			NMT Models				
NLLB-600M	Base	1.55	16.86	106.80	0.880	0.539	-
	+ LoRA	7.43	29.31	88.32	0.903	0.667	+5.88
NLLB-1.3B	Base	2.15	17.78	106.41	0.885	0.573	-
	+ LoRA	8.01	30.02	87.74	0.905	0.687	+5.86
M2M100-1.2B	Base	0.62	10.70	100.50	0.858	0.475	-
	+ QLoRA	10.96	33.09	**82.99**	**0.911**	0.710	**+10.34**
	+ Full FT	9.60	31.16	83.43	0.908	0.692	+8.98
			LLM				
Krikri-8B-IT	Base	8.29	29.87	88.13	0.895	0.695	-
	+ QLoRA	11.90	34.07	84.16	0.906	**0.713**	+3.61
	+ Full FT	**13.16**	**34.71**	83.68	0.848	0.702	+4.87

Table 5: Main results on the Test set (2,000 sentence pairs). BLEU scores are BLEU-13a. △BLEU shows improvement over the base model. Best score per metric in bold.

Model	Method	BLEU↑	chrF++↑	TER↓	BERTScore F1↑	COMET↑	△BLEU
			NMT Models				
NLLB-600M	Base	0.77	14.40	118.13	0.866	0.484	-
	+ LoRA	5.65	28.74	88.01	0.900	0.638	+4.89
NLLB-1.3B	Base	1.25	16.15	107.03	0.873	0.525	-
	+ LoRA	5.68	28.94	88.24	0.900	0.656	+4.43
M2M100-1.2B	Base	0.07	9.37	100.34	0.840	0.428	-
	+ QLoRA	9.52	33.30	81.95	**0.911**	0.691	**+9.45**
	+ Full FT	8.16	31.12	83.11	0.907	0.664	+8.09
			LLM				
Krikri-8B-IT	Base	6.55	28.98	87.38	0.900	0.695	-
	+ QLoRA	10.37	34.09	82.28	**0.911**	**0.717**	+3.82
	+ Full FT	**12.80**	**35.90**	**81.40**	0.884	0.716	+6.25

Table 6: Results on the Stress set (250 sentence pairs with rarer dialects). BLEU scores are BLEU-13a. △BLEU shows improvement over the base model. Best score per metric in bold.

6. Conclusion

In this paper, we addressed the critical scarcity of resources for Ancient Greek (AG) to Modern Greek (MG) machine translation, a low-resource task compounded by the significant dialectal, historical, and genre-based diversity of the Ancient Greek source texts. Our primary contribution is the introduction of the AG-MG Parallel Corpus, the largest sentence-aligned dataset for this pair, containing 132,481 high-quality pairs. This new resource is derived from diverse literary, historical, and biblical texts and is comprehensively annotated with metadata for author, dialect, genre, and era, enabling more granular analysis in future work.

We detailed the novel hybrid alignment pipeline developed to create this corpus. This method improves upon existing approaches by first fine-tuning LaBSE embeddings on a 1,000-pair manually-aligned AG-MG subset before using them with VecAlign for initial alignment. A crucial second stage involved an LLM-based refinement step using Gemini 2.5 Flash, which was prompted to act as a "*cautious data editor*" to perform misalignment detection and correction, ensuring the high quality of the final dataset.

Furthermore, we presented a comprehensive benchmark comparing modern NMT models (NLLB-600M, NLLB-1.3B, M2M100-1.2B) and a specialized Greek LLM (Llama-Krikri-8B-Instruct) on this new task. Our experiments evaluated three different adaptation strategies, including full fine-tuning, LoRA, and QLoRA. The results unequivocally confirm that fine-tuning on our corpus yields significant improvements over zero-shot per-

formance, with BLEU scores increasing by up to +10.3 points. The fully fine-tuned Llama-Krikri-8B model achieved the highest absolute performance on the main test set, reaching a BLEU score of 13.16 and the highest chrF++ score of 34.71. Meanwhile, the QLoRA-adapted M2M100-1.2B model demonstrated the most substantial relative gains, improving from a near-zero base score to 10.96 BLEU.

The Llama-Krikri-8B model's strength was particularly evident on our Stress set, which focused on rarer AG dialects (Ionic, Doric, Homeric). On this challenging set, the fully fine-tuned Krikri model achieved the highest BLEU (12.80) and chrF++ scores, while its QLoRA-adapted variant achieved the highest COMET (0.717). This indicates superior generalization and robustness to dialectal variation compared to standard NMT models. This, combined with the fact that its full fine-tuning version performed worse on some metrics, suggests QLoRA may be a more stable and effective adaptation strategy for this LLM in this low-resource scenario.

While acknowledging limitations such as domain bias toward literary/biblical texts and the need for human evaluation, this work provides a new, high-quality dataset and a clear benchmark for a challenging task. To foster further research in Greek NLP and Digital Humanities, we will make our fine-tuned models available. However, as a precautionary measure due to the complex and often uncertain copyright status of the source materials, we are unable to release the compiled corpus directly.

7. Limitations

While our work provides a significant new resource and benchmark, several limitations should be acknowledged. First, the **corpus composition** reflects the available digital sources, primarily literary, philosophical, and biblical texts. This may introduce domain bias, and models trained on it might perform less optimally on other genres. Second, despite our LLM-based refinement, the **alignment quality** is estimated at 95% based on a manual review of 150 examples. While this sample indicates high quality, a larger human evaluation could provide a more robust assessment of alignment errors, particularly for highly non-literal translations or complex sentence rearrangements inherent in literary texts. Third, our **evaluation** primarily relies on automatic metrics. While comprehensive, these may not fully capture translation adequacy and fluency, especially for a language pair with rich morphology and stylistic variation like AG-MG. Large-scale human evaluation and error analysis would permit a deeper quality assessment. Furthermore, due to the small size of the Stress set (250 pairs), differences in metrics like BLEU between models may lack statistical reliability, and confidence intervals should be considered in future evaluations. Fourth, our **modeling experiments** explored specific architectures and PEFT methods. Other models or fine-tuning techniques might yield different results.

Notably, this work focused exclusively on **sentence-level translation**. Our original data sources were often excerpt-based, and we retained this paragraph-level alignment separately. Future work should explore leveraging this longer-context data, potentially by fine-tuning LLMs like Krikri-8B for document-level AG-MG translation, which might better handle discourse phenomena and context-dependent translations, especially for literary texts (Karpinska and Iyyer, 2023; Wang et al., 2023). Analyzing model performance across different dialects and genres using the corpus metadata is another area for future research. Finally, our experiments focused solely on the AG→MG translation direction. While our parallel corpus could potentially be reversed to train MG→AG models, generating morphologically rich and syntactically complex Ancient Greek poses significantly greater challenges, representing a distinct direction for future investigation.

8. Ethical Considerations

The data used in this work was compiled from publicly accessible online resources, primarily intended for educational and research purposes. We believe our use aligns with the intended purpose of these digital libraries. The created dataset, AG-MG Parallel Corpus, consists of historical texts and their modern translations. While the source texts themselves may reflect historical biases, our processing did not introduce new biases beyond potential selection effects from the source availability.

The models fine-tuned are based on open-source pre-trained models (NLLB, M2M100, Llama-Krikri). Our intended use for the fine-tuned models and the corpus is for research, education, and cultural heritage preservation within the NLP and Digital Humanities communities. Potential misuse seems minimal given the specific language pair and domain, but users should be aware that, although MT quality is high according to metrics, it is not perfect and should not be relied upon for critical applications without human oversight.

9. Acknowledgments

This work was supported in part by a thesis scholarship granted to the first author by the Institute for Language and Speech Processing (ILSP), Athena Research Center. This work was also supported in part by the PHAROS project (Grant Agreement No. 101234269). We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the LEONARDO supercomputer, hosted by CINECA (Italy) and the LEONARDO consortium through the EuroHPC Development Access call.

10. Bibliographical References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Caroline Craig, Kartik Goyal, Gregory Crane, Farnoosh Shamsian, and David A Smith. 2023. Testing the limits of neural sentence alignment models on classical greek and latin texts and translations. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. *arXiv preprint arXiv:2004.14524*.

Ourania Kolovou. 2024. Machine translation from ancient greek to english: Experiments with opennmt.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F Wong, Siyou Liu, and Longyue Wang. 2023. A paradigm shift: The future of machine translation lies with large language models. *arXiv preprint arXiv:2305.01181*.

Chiara Palladino, 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*, 9:22.

Katerina Papantoniou and Yannis Tzitzikas. 2024. Nlp for the greek language: A longer survey. *arXiv preprint arXiv:2408.10962*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Maciej Rapacz and Aleksander Smywiński-Pohl. 2025. Low-resource interlinear translation: Morphology-enhanced neural models for ancient greek. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 145–165.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Dimitris Roussis, Leon Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2025. Krikri: Advancing open large language models for greek. *arXiv preprint arXiv:2505.13772*.

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.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando De Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1342–1348.

L Voukoutis, D Roussis, G Paraskevopoulos, S Sofianopoulos, P Prokopidis, V Papavasileiou, A Katsamanis, S Piperidis, and V Katsouros. Meltemi: The first open large language model for greek. arxiv 2024. *arXiv preprint arXiv:2407.20743*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Tariq Yousef, Chiara Palladino, and Farnoosh Shamsian. 2023. Classical philology in the time of ai: Exploring the potential of parallel corpora in ancient language. In *Proceedings of the Ancient Language Processing Workshop*, pages 179–192.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022a. An automatic model and gold standard for translation alignment of Ancient Greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France. European Language Resources Association.

Tariq Yousef, Chiara Palladino, David J Wright, and Monica Berti. 2022b. Automatic translation alignment for ancient greek and latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. *arXiv preprint arXiv:1808.09564*.

11. Language Resource References

Vatri, Alessandro and McGillivray, Barbara. 2018. *The Diorisis Ancient Greek Corpus: Linguistics and Literature*. Brill.

Appendix A. Corpus Creation Pipeline

Figure 1 illustrates the multi-stage hybrid alignment pipeline used to create the AG-MG Parallel Corpus, combining neural embeddings with LLM-based refinement.

Appendix B. Detailed Corpus Statistics

This section provides a granular breakdown of the AG-MG Parallel Corpus. Tables 7 through 10 detail the dialect distributions across the specific data splits.

Figure 2 visualizes the overall dominance of the Attic and Koine dialects. Finally, as shown in Table 1, the corpus exhibits a significant token count disparity between the source and target languages. The Modern Greek side contains approximately 32% more tokens than the Ancient Greek side (3.06M vs. 2.31M). This expansion is linguistically expected and reflects the typological shift from Ancient Greek, a highly synthetic language rich in inflection, to Modern Greek, which is more analytic. Ancient Greek often encodes grammatical information (such as subject, tense, and mood) into single word endings, whereas Modern

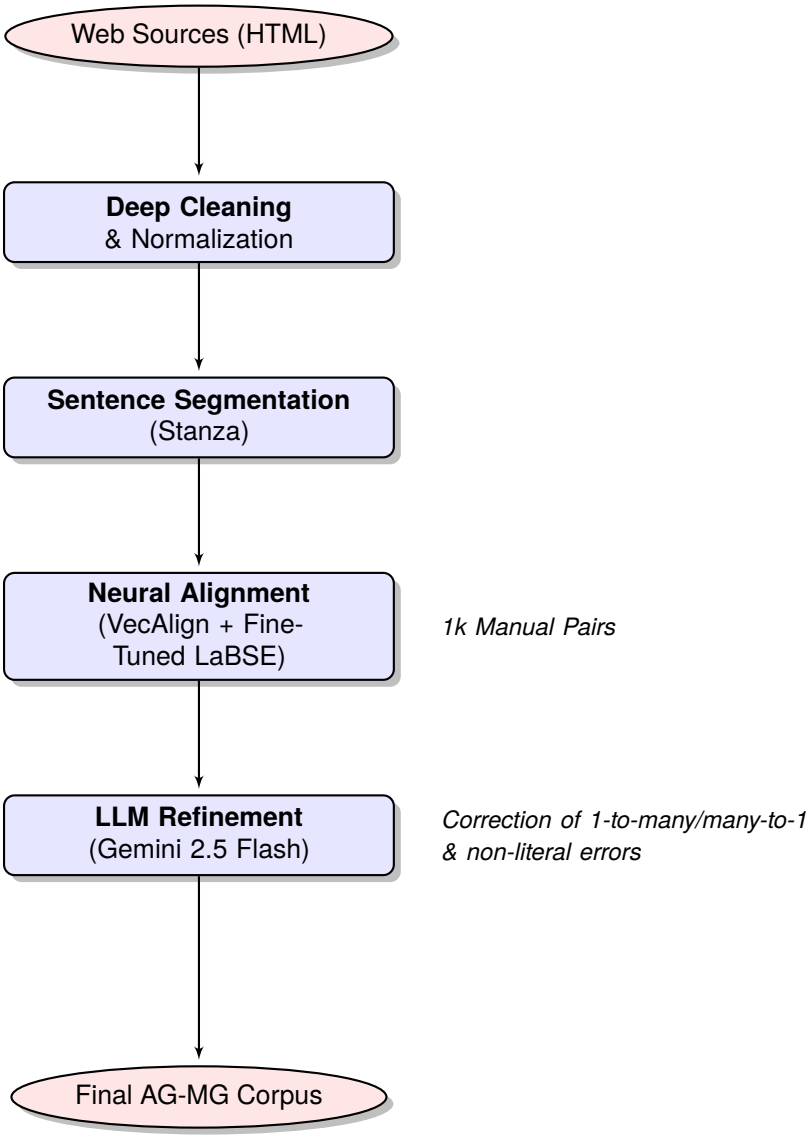

Figure 1: The hybrid corpus creation pipeline, combining neural embedding-based alignment with LLM-based refinement.

Greek frequently employs periphrastic constructions, auxiliary verbs, and particles to convey the same meaning. This is also visually evident in the token distribution histogram (Figure 3), where the Modern Greek distribution is consistently shifted towards higher token counts compared to the Ancient Greek counterpart.

Dialect	Rows	Pairs
Attic	3,089	73,287
Ionic	232	8,615
Doric	244	2,554
Homeric / Epic	272	6,649
Hellenistic Koine	1,491	37,126
Total	**5,328**	**128,231**

Table 7: Sentence-Level Dialect Distribution: Train Set.

Dialect	Rows	Pairs
Attic	134	1,300
Hellenistic Koine	58	700
Total	**192**	**2,000**

Table 8: Sentence-Level Dialect Distribution: Dev Set.

Dialect	Rows	Pairs
Attic	48	1,300
Hellenistic Koine	28	700
Total	**76**	**2,000**

Table 9: Sentence-Level Dialect Distribution: Test Set.

Dialect	Rows	Pairs
Ionic	3	110
Doric	9	60
Homeric / Epic	5	80
Total	**17**	**250**

Table 10: Sentence-Level Dialect Distribution: Stress Set.

Appendix C. Qualitative Translation Examples

Table 11 provides a representative qualitative comparison of model outputs on a sentence from the Test set. The base NMT models often fail or hallucinate unrelated text, while the fine-tuned versions align correctly with the context and vocabulary.

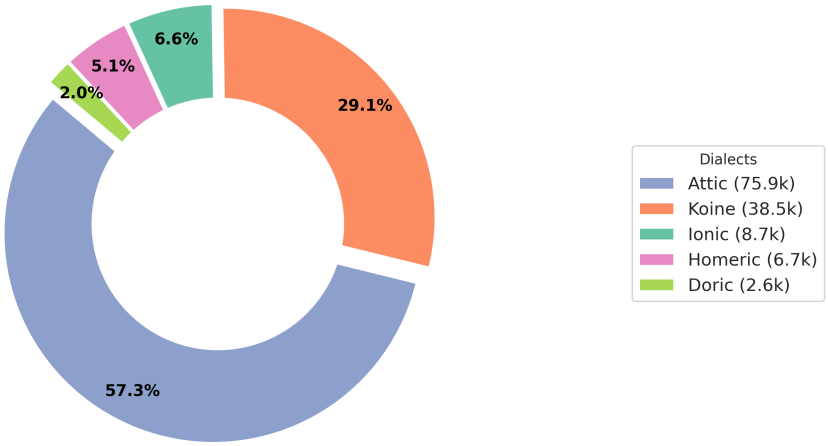

Figure 2: Distribution of Ancient Greek dialects in the sentence-level AG→MG corpus. The dataset is predominantly Attic (57.3%) and Koine (29.1%), reflecting the availability of digital resources.

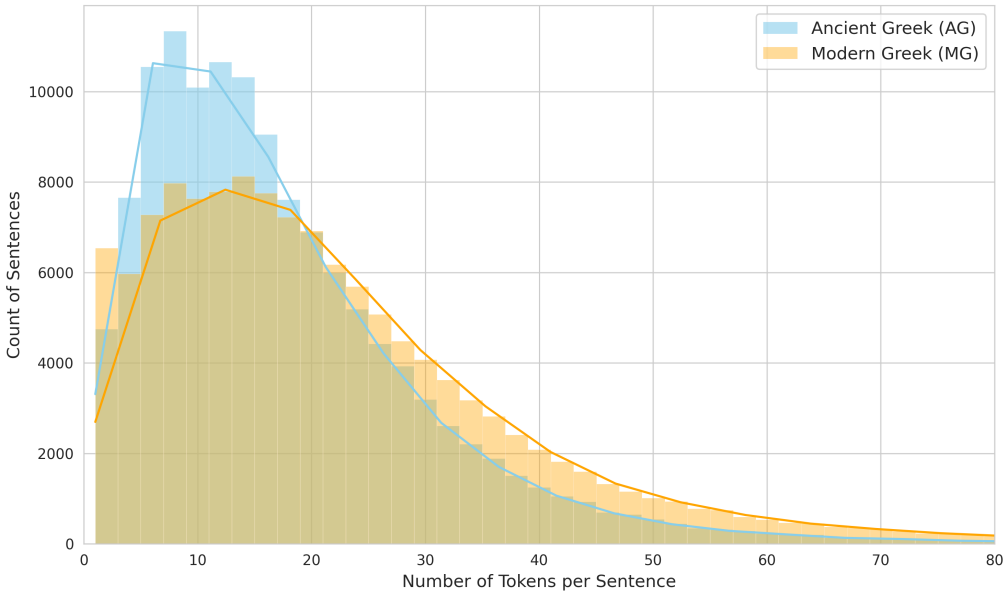

Figure 3: Sentence-Level token count distribution (132.4k pairs). Most sentences are short (10-30 tokens), with Modern Greek translations showing a slightly wider distribution due to their analytic nature.

Model	Text
Source (AG)	Ἐκ δὲ τούτου Παρμενίωνα μὲν πέμπει ἐπὶ τὰς ἄλλας πύλας, αἳ δὴ ὁρίζουσι τὴν Κιλίκων τε καὶ Ἀσσυρίων χώραν, προκαταλαβεῖν καὶ φυλάσσειν τὴν πάροδον, δοὺς αὐτῷ τῶν τε ξυμμάχων τοὺς πεζοὺς καὶ τοὺς Ἕλληνας τοὺς μισθοφόρους καὶ τοὺς Θρᾷκας, ὧν Σιτάλκης ἡγεῖτο, καὶ τοὺς ἱππέας δὲ τοὺς Θεσσαλούς.
NLLB-600M (Base)	Μέχρι να περάσει ο Παρμηνίος από τις άλλες πύλες, που βρίσκονται στην περιοχή της Κιλικνίας και της Ασσυρίας, να προχωρήσει και να φυλάξει το πέρασμα, να κατασκευάσει τα τετράγωνα των πεζοδρόμων, και τα ελληνικά μισθοφόρα και τα θραύσματα, και το Σιτάλκιο και το Σπέπας των Θεσσαλονίκων.
NLLB-600M (LoRA)	Ύστερα έστειλε τον Παρμενίωνα στις άλλες πύλες, που καθορίζουν τη χώρα των Κιλίκων και τωνἈσσυρίων, για να προλάβει και να φυλάξει το πέρασμα, δίνοντάς του και τους πεζούς των συμμάχων του και τωνἙλλήνων τους μισθοφόρους και τους Θράκες, που τους διοικούσε ο Σιτάλκης, και τα άλογα των Θεσσαλών.
NLLB-1.3B (Base)	Ο Παρμαινίος με έστειλε στην Ελληνική πύλη, στις περιοχές της Κιλικίας και της Συρίας, για να προχωρήσω και να προστατέψω το ταξίδι, με τους πεζούς ναυτικούς, τους Ελληνες ναυτικούς, τον Τραύκο, τον Στάλκη, τον Γεώργιο και τον Ιάππο στη Θεσσαλονίκη.
NLLB-1.3B (LoRA)	Ύστερα απ᾽ αυτό έστειλε τον Παρμενίωνα στις άλλες πύλες, που χωρίζουν τη χώρα των Κιλικόνων και τωνἈσσυρίων, για να προλάβει και να φυλάξει το πέρασμα, και του παρέδωσε το πεζικό των συμμάχων, τους Έλληνες μισθοφόρους και τους Θράκες, που τους οδηγούσεό Σίλκης, και τους Θεσσαλούςἱππείς.
M2M-1.2B (Base)	*[Hallucination]* ταν ὁ ἄνθρωπος εἶναι ἄλλος, ὅταν ἔχει ἐντολή νά ὑπάρχει ἀπὸ τήν μαρτία του, ἐκείνος ἐπικρατεί ὅτι ἐξαρτάται ἀπό τόν αυτό του.
M2M-1.2B (QLoRA)	Ύστερ᾽ απ᾽ αυτό έστειλε τον Παρμενίωνα στις άλλες πύλες, που ορίζουν τη χώρα των Κιλίκων και των Ασσυρίων, για να προλάβουν και να φρουρήσουν το πέρασμα, δίνοντάς του από τους συμμάχους τους πεζούς, τους Έλληνες τους μισθοφόρους και τους Θράκες, με αρχηγό τον Σιτάλκη, και το ιππικό τους τους Θεσσαλούς.
M2M-1.2B (FFT)	Ύστερ᾽ απ᾽ αυτό έστειλε τον Παρμενίωνα στις άλλες πύλες, που ορίζουν τη χώρα των Κιλίκων και των Ασσυρίων, για να προλάβει και να φυλάξει το πέρασμα, δίνοντάς του τους πεζούς των συμμάχων, τους Έλληνες, τους μισθοφόρους, τους Θράκες, που τους οδηγούσε ο Σιτάλκης, και τους ιππείς τους Θεσσαλούς.
Krikri-8B-IT (Base)	Μετά το γεγονός αυτό, ο Παρμενίων στάλθηκε στις υπόλοιπες πύλες που οριοθετούν την περιοχή της Κιλικίας και της Ασσυρίας, με εντολή να προλάβει και να φυλάξει τις διαβάσεις, δίνοντάς του τους πεζούς συμμάχους, τους Έλληνες μισθοφόρους και τους Θράκες (με αρχηγό τον Σιτάλκη) καθώς και τους Θεσσαλούς ιππείς.
Krikri-8B-IT (QLoRA)	Ύστερα απ᾽ αυτό ο Αλέξανδρος έστειλε τον Παρμενίωνα να καταλάβει και να φρουρήσει τις άλλες πύλες που χωρίζουν τη χώρα των Κιλίκων και των Ασσυρίων, δίνοντάς του τους πεζούς συμμάχους και τους Έλληνες μισθοφόρους και τους Θράκες, που τους διοικούσε ο Σιτάλκης, καθώς και το θεσσαλικό ιππικό.
Krikri-8B-IT (FFT)	Μετά από αυτό έστειλε τον Παρμενίωνα στις άλλες πύλες, με τις οποίες ορίζεται η χώρα των Κιλίκων και των Ασσυρίων, για να καταλάβει και να φυλάξει τα περάσματα, δίνοντάς του τους πεζούς των συμμάχων, τους Έλληνες και τους Θράκες μισθοφόρους, τους οποίους διοικούσε ο Σιτάλκης, καθώς και το θεσσαλικό ιππικό

Table 11: Qualitative comparison on a Test Set sentence (*The Anabasis of Alexander*, Arrian, Attic dialect). Some base models hallucinate or fail, while fine-tuned models produce fluent Greek.