Greek Multi30K: Translating Image Captions for Greek Vision-Language Research

Anonymous ACL submission

Abstract

Multimodal Machine Translation (MMT) and vision-language research depend on parallel image-caption corpora, but such resources remain scarce for languages like Greek. We address this gap by releasing **Greek Multi30K**, the first Greek-language version of the widely-used Multi30K image caption dataset. To select the optimal translation system, we conduct a systematic evaluation of 23 system configurations spanning traditional MT (DeepL, Google Translate) and state-of-the-art LLMs (e.g., GPT-4, Claude 3.7 Sonnet, Gemma-3-27B, Llama-Krikri-8B) under zero-shot, few-shot, and post-editing prompting strategies. We assess translation quality using five automatic metrics (BLEU, METEOR, chrF, COMET, TER), an LLM-as-a-Judge framework for adequacy and fluency, and paired non-parametric statistical testing. Our findings show that specialized MT systems still outperform general-purpose LLMs in zero-shot settings, but few-shot prompting substantially closes the gap, most strikingly for the Greek-adapted Llama-Krikri-8B (rank 18.2 $\rightarrow$ 10.4 with three examples). We further observe a systematic divergence between lexical metrics and semantic evaluations on post-edited output. The resulting 31,014-sentence corpus, released in three versions (Baseline, Post-Edited, LLM-Generated), provides a foundational resource for Greek vision-language and MMT research.

1 Introduction

The expansion of Large Language Models (LLMs) has significantly advanced the capabilities of automated translation, often achieving near-human performance. Still though, MT in general is far from a solved problem (Jurafsky and Martin, 2009). It might be considered a solved problem should it ever become impossible to distinguish automated output from human translation (Corston-Oliver et al., 2001). The LLMs progress has been fueled by the availability of large-scale, parallel datasets.

While Large Language Models (LLMs) have achieved near-human performance in many tasks, a disparity exists in the availability of resources for multimodal research, particularly for languages other than English. The Multi30K dataset (Elliott et al., 2016), which contains 31,014 images sourced from Flickr30K dataset (Young et al., 2014) is a cornerstone for Multimodal Machine Translation (MMT), offering parallel image descriptions in English, German, French, and Czech. However, it explicitly lacks a Modern Greek version. This absence creates a bottleneck for Greek NLP, limiting the development of vision-language models that rely on visual context to resolve linguistic ambiguities.

While XM3600 dataset (Thapliyal et al., 2022), represents a broader evaluation benchmark covering 36 languages including Greek, its scope and design differ substantially from Multi30K in ways that limit direct comparability: XM3600 consists of 3,600 geographically diverse images selected from Open Images, whereas Multi30K is grounded in the Flickr30K image set, which, as stated, has become a standard testbed for multimodal machine translation research. Furthermore, the Greek portion of XM3600 dataset is intended purely as an evaluation benchmark, not a training resource, leaving a gap in the resources available for training Greek multimodal systems in a directly comparable setting.

This work makes four contributions. (1) We present a systematic evaluation of LLM-era English-to-Greek MT, covering 23 system configurations including specialized MT engines, frontier LLMs, and Greek-adapted models. (2) We demonstrate that few-shot prompting can substantially elevate small specialized models: the Greek-adapted Llama-Krikri-8B improves by eight rank positions with just three in-context examples. (3) We identify a systematic divergence between lexical n-gram metrics and semantic evaluation (COMET, LLM-as-a-Judge) on paraphrastic post-edited output,

with implications for evaluating fluency-oriented techniques. (4) We release **Greek Multi30K**, a 31,014-sentence parallel corpus in three versions (Google Translate baseline, Llama-Krikri-8B post-edited, and Gemma-3-27B LLM-generated), enabling future Greek vision-language and MMT research.

2 Background and Related Work

Image Caption Translation and Multi30K extensions. The Multi30K dataset (Elliott et al., 2016) extends Flickr30K (Young et al., 2014) with parallel image captions in German, French, and Czech, and has served as the standard benchmark for MMT for nearly a decade. Subsequent efforts have extended it to additional languages, including Hindi (Parida et al., 2019) and others, demonstrating the value of multilingual caption corpora. A complementary resource, XM3600 (Thapliyal et al., 2022), covers 36 languages including Greek, but is constructed from a different image source (Open Images) and is intended as an evaluation-only benchmark rather than a training resource. To our knowledge, no Greek Multi30K resource exists, leaving a gap for training and evaluating Greek vision-language models in the same setting as the established Multi30K literature.

Greek NLP and MT. Greek is a morphologically rich language with comparatively limited resources for modern neural systems. Recent efforts to develop Greek-adapted language models, including Llama-Krikri (Roussis et al., 2025) and similar initiatives, have begun to close this gap, but parallel image caption corpora remain absent. Evaluation in morphologically rich settings benefits particularly from character-level metrics such as chrF (Popović, 2015), which we include in our evaluation alongside word-level metrics.

MT evaluation. While n-gram metrics like BLEU (Papineni et al., 2002) remain widely used, recent shared tasks (Freitag et al., 2024) have shown that neural metrics like COMET (Rei et al., 2020) correlate substantially better with human judgment. We therefore rely on COMET as our primary indicator of semantic accuracy, complemented by METEOR (Banerjee and Lavie, 2005; Lavie and Agarwal, 2007), chrF (Popović, 2015), and TER (Snover et al., 2006). We further employ an LLM-as-a-Judge protocol, following recent evidence (Zheng et al., 2023; Liu et al., 2023) that strong LLMs can produce reliable adequacy and fluency assessments, though potential biases such as self-preference and limited validation in non-English settings remain open questions (Wang et al., 2023).

Synthetic corpus creation. Recent work has increasingly relied on machine translation and LLM-generated text to create parallel corpora for low- and mid-resource languages (NLLB Team et al., 2022). While this approach scales effectively, the quality and downstream utility of such corpora depend critically on the translation system used. Our study contributes a systematic comparison aimed at selecting the most suitable system for this purpose in the English-to-Greek direction.

3 Experimental Setup

Our experimental design focuses on identifying the most suitable translation system for creating the Greek Multi30K dataset. This involves evaluating a diverse set of models using various prompting engineering techniques, on a controlled sample of 200 sentences, before applying the best methods to the full dataset.

3.1 Data

The primary dataset for this work is the **Multi30K dataset**. For our evaluation, we constructed a reference set of 200 English sentences sampled from the Multi30K training split. To ensure diversity, the sample was stratified along two axes: **length** (short: < 10 tokens; long: > 20 tokens) and **syntactic complexity** (simple: single main clause; complex: containing one or more subordinate clauses or coordinated structures), with approximately one quarter of sentences drawn from each category (categories may overlap). We acknowledge that any manual selection process introduces a degree of subjectivity, and that fully random sampling would yield greater reproducibility at the cost of underrepresenting rarer but linguistically important sentence types. To make this trade-off explicit, we discuss its implications in Section 8.

This sample was then professionally translated into Greek by a human translator to serve as our **golden reference file** for automatic metric calculation.

3.2 Models Evaluated

A range of models was selected to represent the current landscape of translation technology, aiming

to cover:

- **Traditional MT Systems:** DeepL (DeepL GmbH, 2025) and Google Translate (Wu et al., 2016), two leading NMT services. Both were initially accessed through their web interfaces in April 2025. Google Translate (Google, 2025) was subsequently accessed via API in June 2025 to translate the full Multi30K dataset.

- **Large Language Models (LLMs):** A diverse set of recent and powerful models, initially accessed through their web interfaces in April and May 2025 using different prompting techniques. This set includes:

  – State-of-the-art closed LLMs like GPT-4 (OpenAI et al., 2024) and Claude 3.7 (Anthropic, 2024a) Sonnet (Anthropic, 2024b).

  – Representative open models of varying sizes, including the instruct version of Gemma-3-27B (Gemma 3 Team, 2025) accessed via Google AI Studio (Gemma Team, 2025) for initial experiments, and via API in June 2025 for the full dataset translation, Aya Expanse-8B (Dang et al., 2024) accessed via Cohere's Developer Playground, and DeepSeek-V3 (Liu et al., 2024).

  – A model specifically adapted for Greek, Llama-Krikri-8B (Roussis et al., 2025), which was run locally via HuggingFace Transformers in June 2025 to generate the post-edited version of the full dataset.

3.3 Methodology

To assess model performance comprehensively, we employed four distinct prompting strategies. Each of the six LLMs was tested using the first three techniques, creating 18 distinct LLM-based systems for comparison alongside the two traditional MT systems. As for the Greek-specialized Llama-Krikri-8B LLM, we conducted experiments both in English (Llama-Krikri-8B EN) and in Greek (Llama-Krikri-8B EL) for the first three techniques. As shown in Table 1, overall we have 23 models.

1. **Without Prompting:** This approach assumes the model has sufficient training on the specific task (like translation) to understand what's expected from just the command word and context. We provided the LLMs with a simple, direct instruction, such as:

   `"Translate the below 200 sentences in Greek language".`

   This tests the model's raw, out-of-the-box translation capability. It enables the model's output to reflect its default translational tendencies more accurately, rather than any stylistic bias introduced by elaborate instructions (Zhang et al., 2025).

2. **Zero-Shot Prompting:** A very common way of using this technique is by giving a concept, describing to the model what is the job needed to be done, in what tone, from which point of view etc. We used a detailed prompt to frame the task, assigning the model a persona and providing explicit guidelines for accuracy, formality, and fluency. The prompt we created is to be used for all models.

   The prompt began:

   "You are a professional translator specializing in English-to-Greek translations. Your task is to translate the following 200 English sentences into Modern Greek. Please adhere to the following guidelines:

   Produce translations that are accurate and faithful to the original meaning.

   Maintain the same level of formality or informality as the English sentence.

   Use natural and fluent Greek that would sound authentic to a native speaker.

   Preserve the original sentence structure when possible, but feel free to adjust slightly for naturalness.

   Translate the sentences carefully and thoughtfully.

   [200 sentences]".

3. **Few-Shot Prompting:** It's a technique where the user provides a Language Model (LM) with a small number of examples (typically 1-5) that demonstrate the task you want it to perform before asking it to complete a similar task (Brown et al., 2020). We used the same detailed prompt as in the zero-shot case but augmented it with three fixed examples from the Multi30K training split (disjoint from the test set), spanning varied length and complexity, to provide in-context learning. The same fixed examples were used across all experiments. Here are the examples:

   `"Example 1`
   `EN``: Boys dancing on poles in the middle of the night.`
   `EL``: Αγόρια χορεύουν σε στύλους στη μέση της νύχτας.`
   `Example 2`

EN: Five individuals in pretending to be zombies and one infant.
EL: Πέντε άτομα που προσποιούνται ότι είναι ζόμπι και ένα βρέφος.
Example 3
EN: A young Caucasian boy climbs through a metal structure of stationary rings outside in a paved area overlooking a body of water.
EL: Ένα νεαρό καυκάσιο αγόρι σκαρφαλώνει μέσα από μια μεταλλική κατασκευή σταθερών δακτυλίων έξω σε μια πλακόστρωτη περιοχή με θέα μια λίμνη νερού".

4. **Post-Editing (PE):** It refers to the process of reviewing and correcting machine-generated output to improve its quality, accuracy, and fluency. As a complementary practice, we took the output from Google Translate and used the Greek-specialized Llama-Krikri-8B LLM to post-edit it, with instructions in English to correct errors and improve naturalness. The prompt was given in zero-shot (without the examples) and in few-shot version:

```
"""You are a professional Greek translator. You will receive an English .txt file along with a Greek translation .txt file generated by a machine (Google Translate). Your task is to post-edit the Greek translation to make it more accurate, natural, and fluent in Greek, while preserving the meaning of the original English sentence. Produce a first draft of the translation. Correct the output (post-edit) of the google_translate_200.txt to produce the best possible translation. Please provide the improved Greek translation only.
Below are some examples:
[The same three examples as provided in the Few-Shot Prompting section were included here.] Now, post-edit the following:
English: {english}
Google Translation: {greek}
Post-edited Greek Translation:
"""
```

As stated above, this resulted in 18 LLM-based systems (6 models x 3 prompt types) plus the 2 traditional MT systems, totaling 23 distinct systems for evaluation. Additionally, 2 post-editing systems involving Llama-Krikri-8B were analyzed separately. At inference time, sentences were submitted in batches of 50 (not all 200 at once) to avoid hallucinations; the prompt template shown above is reproduced verbatim, though the sentence count was adjusted per batch.

3.4 Evaluation

We evaluated the output of each system against the golden reference file using a suite of five widely-used automatic metrics: **BLEU**, **METEOR**, **chrF**, where we report Normalized chr-F for better understanding. Also, **COMET** and **TER** where we report Inverted TER so that higher is better. As for metric implementations we used NLTK for BLEU (sentence-level, averaged) and METEOR (with WordNet stemming/synonyms), sacreBLEU for chrF and TER, and Unbabel wmt22-comet-da for COMET. Our work uses a combination of these metrics to provide a holistic evaluation of the translation quality.

To complement the automatic metrics, we employed an LLM-as-a-Judge evaluation framework to assess translation quality along two dimensions: adequacy and fluency. Each of the 200 source sentences and their corresponding translations from all 25 systems were submitted to the MiniMax M2.5 model via the OpenRouter API, using a structured prompt that instructed the model to act as an expert bilingual linguist specializing in English and Modern Greek. For each sentence, the judge assigned a score from 1 to 5 for both adequacy (how well the Greek translation captures the meaning of the English source) and fluency (how grammatically correct and natural the translation sounds to a native speaker), along with a brief reasoning.

MiniMax M2.5 model was chosen as an architecturally distinct family from all 23 evaluated systems, to reduce self-preference bias. We acknowledge the lack of Greek-specific human-correlation validation as a limitation.

To determine statistical significance, we treated the evaluation set as paired data, as all systems translated the same source sentences. We employed the Friedman test to detect overall differences across systems. For pairwise comparisons between the models (e.g., DeepL vs. Claude 3.7 Sonnet), we utilized the Wilcoxon Signed-Rank test with Bonferroni correction to account for non-normal distribution in automatic metrics. We report rankings as the arithmetic mean across the five metric-specific ranks for ease of inspection. Since ranks are ordinal, this aggregation should be interpreted as a coarse summary rather than an interval-scaled score; we therefore complement it with per-metric reporting (Tables 1–4) and with pairwise non-parametric significance testing, which does not rely on rank-averaging assumptions.

4 Results and Analysis

The comprehensive evaluation showed clear and consistent results across all metrics and statistical

Model	BLEU	METEOR	Norm. chrF	COMET	Inv. TER	BLEU R.	METEOR R.	chrF R.	COMET R.	TER R.	Avg. Rank
DeepL	0.6713	0.8761	0.8796	0.9502	0.8669	1	1	1	1	1	1
Claude 3.7 Sonnet F-S	0.6207	0.8460	0.8207	0.9402	0.7978	2	2	2	3	2	2.2
Claude 3.7 Sonnet Z-S	0.6170	0.8413	0.8172	0.9406	0.7918	3	3	3	2	3	2.8
Claude 3.7	0.5990	0.8318	0.8054	0.9352	0.7820	4	4	4	4	4	4
gemma-3-27B-it F-S	0.5684	0.8057	0.7823	0.9316	0.7615	6	5	5	6	5	5.4
gemma-3-27B-it	0.5830	0.8040	0.7806	0.9309	0.7557	5	6	6	7	6	6
gemma-3-27B-it Z-S	0.5581	0.7944	0.7682	0.9294	0.7476	7	7	8	8	7	7.4
GPT-4 F-S	0.4962	0.7687	0.7351	0.9343	0.7072	9	9	9	5	9	8.2
Google Translate	0.4983	0.7874	0.7693	0.9198	0.7446	8	8	7	12	8	8.6
Krikri-8B EN F-S	0.4549	0.7365	0.7174	0.9223	0.6668	10	10	10	10	12	10.4
Aya Expanse-8B F-S	0.4065	0.7225	0.7114	0.9190	0.6712	13	11	12	14	11	12.2
GPT-4 Z-S	0.4214	0.7179	0.6853	0.9273	0.6462	12	12	17	9	15	13
Aya Expanse-8B Z-S	0.3824	0.7135	0.7005	0.9197	0.6567	15	13	13	13	13	13.4
DeepSeek-V3 Z-S	0.3797	0.7123	0.7134	0.9174	0.6823	17	14	11	17	10	13.8
Krikri-8B EL Z-S	0.3875	0.6990	0.6909	0.9181	0.6449	14	16	15	15	16	15.2
GPT-4	0.4313	0.7044	0.6419	0.9207	0.6061	11	15	20	11	21	15.6
Krikri-8B EL F-S	0.3558	0.6879	0.6919	0.9178	0.6469	20	19	14	16	14	16.6
DeepSeek-V3	0.3801	0.6858	0.6857	0.9146	0.6429	16	20	16	19	17	17.6
Krikri-8B EN Z-S	0.3758	0.6907	0.6818	0.9158	0.6139	18	17	18	18	20	18.2
Aya Expanse-8B	0.3570	0.6890	0.6737	0.9057	0.6395	19	18	19	21	18	19
Krikri-8B EN	0.2893	0.6383	0.6411	0.9092	0.6162	22	21	21	20	19	20.6
Krikri-8B EL	0.3083	0.6096	0.5550	0.8868	0.5263	21	22	23	22	23	22.2
DeepSeek-V3 F-S	0.2841	0.5617	0.6384	0.8312	0.5937	23	23	22	23	22	22.6

Table 1: Overall performance scores and average rank across all metrics for all 23 models evaluated. For LLMs, the prompt technique is specified. Due to space constraints, Zero-Shot is abbreviated as Z-S, and Few-Shot is abbreviated as F-S.

Model	BLEU	METEOR	Norm. chrF	COMET	Inv. TER	BLEU R.	METEOR R.	chrF R.	COMET R.	TER R.	Avg. Rank
Google Translate	0.4983	0.7874	0.7693	0.9198	0.7446	1	1	1	3	1	1.4
GT Post-Edit Krikri-8B Z-S	0.4754	0.7664	0.7529	0.9264	0.7096	2	2	2	2	2	2
GT Post-Edit Krikri-8B F-S	0.4725	0.7643	0.7502	0.9279	0.7038	3	3	3	1	3	2.6

Table 2: Overall performance scores and average rank across all metrics for 3 systems evaluated. Post-Edit prompting technique. Due to space constraints, Zero-Shot is abbreviated as Z-S, Few-Shot is abbreviated as F-S and Google Translate is abbreviated as GT.

Rank	Model	Condition	Avg Adequacy	Avg Fluency	Combined Avg
1	Gemma-3-27B	Few-Shot	4.428	4.728	4.578
2	DeepL	No Prompting	4.371	4.756	4.563
3	Claude 3.7 Sonnet	Zero-Shot	4.359	4.732	4.545
4	Gemma-3-27B	Zero-Shot	4.340	4.722	4.531
5	Claude 3.7 Sonnet	No Prompting	4.288	4.732	4.510
6	Claude 3.7 Sonnet	Few-Shot	4.304	4.691	4.497
7	GPT-4	Few-Shot	4.224	4.755	4.490
8	Krikri-8B EN	Few-Shot	4.279	4.640	4.459
9	GT Post-Edit Krikri-8B	Zero-Shot	4.120	4.738	4.429
10	GT Post-Edit Krikri-8B	Few-Shot	4.160	4.690	4.425

Table 3: Top 10 Model-Condition Pairs by LLM-as-a-Judge Evaluation (Averaged over 200 sentences)

tests. A summary of the overall model rankings, averaged across all metrics, is presented in Table 1. Also, a summary of the post-edited model rankings is presented in Table 2.

4.1 Performance Tiers

The results of Table 1 consistently stratified the models into three performance tiers.

- **Tier 1:** DeepL was the undisputed leader, significantly outperforming all other models across every metric and statistical tests. It was followed by the Claude 3.7 Sonnet family, as one of the top-performing LLMs in this task, which demonstrated robust and high-quality translation.
- **Tier 2:** This tier included the gemma-3-27B-it family, GPT-4 (especially with few-shot (F-S) prompting), and Google Translate. Notably, the freely available Google Translate performed competitively, outranking several advanced LLMs, particularly on lexical metrics like BLEU and chrF.

- **Tier 3:** This tier comprised the remaining models, including Aya Expanse-8B and DeepSeek-V3. The Greek-specialized Llama-Krikri-8B model, while in this tier, showed remarkable improvement with few-shot (F-S) prompting, achieving a top-10 rank and indicating its potential with proper guidance.

4.2 Impact of Prompt Engineering

Prompting strategies had a clear and positive impact on performance for most LLMs.

- **Few-shot (F-S) prompting** consistently yielded the best results for nearly every LLM family (Claude 3.7 Sonnet, gemma-3-27B-it, GPT-4, Llama-Krikri-8B, Aya Expanse-8B, except DeepSeek-V3), consistent with prior findings that in-context examples improve instruction-tuned models. For instance, few-shot prompting significantly boosted the Greek-specialized Llama-Krikri-8B model (using English prompts), improving its average rank from 18.2 (Zero-Shot) to 10.4 (Few-Shot), placing it in the top 10. Notably, this improvement was substantially larger when Krikri-8B was prompted in English than in Greek, despite being a Greek-adapted model: English few-shot (rank 10.4) outperformed Greek few-shot (rank 16.6). We hypothesize this reflects the asymmetric distribution of instruction-following data in the model's training mixture: instruction patterns learned from Llama's English pretraining may transfer more reliably than instructions in Greek, while the three in-context examples (which contain Greek targets) provide sufficient language-side anchoring. This suggests that for Greek-adapted models, the optimal prompting strategy may decouple instruction language from target language.

- The **post-editing** experiment, as shown in Table 2, was particularly insightful. While post-editing the Google Translate output with Llama-Krikri-8B (both zero-shot (Z-S) prompting and few-shot (F-S)) *lowered* scores on lexical metrics like BLEU, it significantly *increased* the COMET score. This suggests that the post-edited translations, while diverging from the exact wording of the reference, were semantically closer and more natural, something that other metrics fail to capture. This highlights a potential limitation of relying solely on lexical metrics when evaluating techniques aimed at improving fluency and adequacy, even if via paraphrasing. Notably, the few-shot post-edited version achieved the 9th highest COMET score overall (0.9279), outperforming many direct LLM translations including the original Google Translate output (0.9198), as shown in Table 4. The magnitude of this COMET gain (+0.008) is modest in absolute terms but consistent across both prompting variants (Z-S and F-S), suggesting a systematic rather than random effect of post-editing on semantic adequacy.

4.3 LLM-as-a-Judge Evaluation

The results from the LLM-as-a-Judge evaluation reveal nuanced differences compared to the automatic metrics. As shown in Table 3, the Gemma-3-27B model with Few-Shot prompting ranks #1 by combined average (4.578). DeepL ranks closely at #2 (4.563) and notably achieves the highest Fluency score (4.756) among the top 10 systems, consistent with its strong handling of Greek morphology. Furthermore, the Claude 3.7 Sonnet family secures three spots in the top tier, reinforcing its robustness as a highly competitive LLM for translation. This slight discrepancy between the traditional metrics (where DeepL was the undisputed #1) and the LLM-as-a-Judge rankings is a notable finding that quantitatively aligns with our observations in Section 4.2. It highlights a known limitation of lexical n-gram metrics: they tend to favor literal, reference-close translations and penalize semantically sound paraphrasing. In contrast, the LLM-as-a-Judge framework functions closer to human evaluation and explicitly evaluates contextual meaning (Adequacy) and natural Modern Greek phrasing (Fluency). These results suggest that while specialized MT systems like DeepL excel in grammatical fluency, properly prompted LLMs (like Gemma-3-27B) can match or exceed specialized MT on semantic adequacy as judged by LLM evaluators.

We note one important caveat to interpreting this result. The LLM judge (MiniMax M2.5) and the systems being judged are both transformer-based language models, raising the possibility of subtle preference bias toward LLM-generated text over the output of specialized MT systems with different decoding characteristics. Prior work has documented related effects, including self-preference and family-preference biases in LLM judges (Wang

et al., 2023). While we mitigated direct self-preference by selecting a judge architecturally distinct from all 23 evaluated systems, we cannot rule out broader stylistic preferences favoring LLM-typical fluency patterns. The discrepancy between automatic metrics and the LLM judge should therefore be read as suggestive rather than conclusive, and motivates future work with native-speaker validation of the judge's scores for Greek.

4.4 Statistical Significance

Our statistical analysis confirmed that the observed differences were statistically significant. As stated, we used non-parametric tests (Friedman test and Wilcoxon Signed-Rank test with Bonferroni correction) due to the nature of the data. The results showed that DeepL significantly outperformed 21 of the other 24 systems (the 22 remaining systems in Table 1 plus the 2 post-editing systems in Table 2; $p < 0.001$). The only exceptions were the three Claude 3.7 Sonnet variants (Claude 3.7 Sonnet F-S, Claude 3.7 Sonnet Z-S, Claude 3.7 Sonnet) whose pairwise comparisons with DeepL did not reach statistical significance after Bonferroni correction ($p = 0.001831, p = 0.002699, p = 0.000201$ accordingly). Moreover, while these uncorrected p-values were small, they did not survive the strict Bonferroni-corrected threshold of p<0.000167, suggesting that the performance difference between DeepL and Claude 3.7 Sonnet, though present, is comparatively small.

5 The Greek Multi30K Resource

Based on our evaluation, we generated three versions of the Greek Multi30K dataset, spanning freely-accessible commercial output, open-weights post-editing, and fully open-weights LLM generation. The dataset consists of a training set of 29,000 sentences, a validation set of 1,014 sentences, and a test set of 1,000 sentences, each translated from the original English descriptions. Prioritizing open and reproducible resources for the Greek NLP community, we release the following three versions:

1. **Google Translate Version:** Translated using the free Google Translate API. This version serves as a strong and accessible baseline.

2. **Post-Edited Version:** The Google Translate output was post-edited using the Llama-Krikri-8B LLM using few-shot prompting through API. This version prioritizes semantic accuracy and naturalness.

3. **Gemma 3 Version:** Translated using the gemma-3-27B-it model with few-shot prompting via the official API. Based on our evaluation, this represents a high-quality, LLM-generated version and was used as a pseudo-reference to compare the above two Multi30K dataset translations (the Google Translate Version and the Post-Edited Version).

5.1 Qualitative Analysis

A qualitative check of the translated outputs revealed common error patterns in Machine Translation (MT). The Google Translate version often struggled with verb-noun agreement and idiomatic expressions. For instance, the English phrase 'getting a tattoo' was often literally translated by Google Translate as ΄παίρνει ένα τατουάζ΄ ('takes a tattoo'), whereas the more idiomatic Greek is ΄κάνει ένα τατουάζ΄ ('does a tattoo'), which was better handled by the post-edited and Gemma 3 versions. Those two versions produced more fluent and natural-sounding Greek, though they occasionally made errors with compound words or kept brand names in English inconsistently.

5.2 Data Availability

The three versions of the Greek Multi30K dataset are available for research purposes in an anonymized repository: https://huggingface.co/datasets/dataset-submission/dataset-submission. Upon acceptance, this will be updated with a public link on GitHub/Hugging Face.

6 Conclusion

We presented a systematic evaluation of 23 English-to-Greek translation system configurations and used those findings to construct Greek Multi30K, a 31,014-sentence parallel image-caption corpus.

Our evaluation showed three consistent performance tiers, with specialized MT (DeepL) significantly outperforming all other systems on automatic metrics ($p < 0.001$), followed by the Claude 3.7 Sonnet family. Few-shot prompting consistently improved LLM performance, most dramatically for the Greek-adapted Llama-Krikri-8B (average rank 18.2 $\rightarrow$ 10.4 with three examples). Post-editing Google Translate output with Llama-Krikri-8B lowered lexical metric scores but increased COMET, revealing a systematic divergence between lexical and semantic evaluation on paraphrastic output, a divergence further confirmed

System	COMET
DeepL	0.9502
Claude 3.7 Sonnet Z-S	0.9406
Claude 3.7 Sonnet F-S	0.9402
Claude 3.7 Sonnet	0.9352
GPT-4 F-S	0.9343
gemma-3-27B-it F-S	0.9316
gemma-3-27B-it	0.9309
gemma-3-27B-it Z-S	0.9294
GT Post-Edit Krikri-8B F-S	0.9279
GPT-4 Z-S	0.9273
GT Post-Edit Krikri-8B Z-S	0.9264
Krikri-8B EN F-S	0.9223
GPT-4	0.9207

System	COMET
Google Translate	0.9198
Aya Expanse-8B Z-S	0.9197
Aya Expanse-8B F-S	0.9190
Krikri-8B EL Z-S	0.9181
Krikri-8B EL F-S	0.9178
DeepSeek-V3 Z-S	0.9174
Krikri-8B EN Z-S	0.9158
DeepSeek-V3	0.9146
Krikri-8B EN	0.9092
Aya Expanse-8B	0.9057
Krikri-8B EL	0.8868
DeepSeek-V3 F-S	0.8312

Table 4: COMET scores ranked: 25 models - All prompting techniques. Due to space constraints, Zero-Shot is abbreviated as Z-S, Few-Shot is abbreviated as F-S and Google Translate is abbreviated as GT.

by our LLM-as-a-Judge results, where Gemma-3 surpassed DeepL on combined adequacy and fluency.

We release Greek Multi30K in three versions, Google Translate (baseline), Llama-Krikri-8B-post-edited, and Gemma-3-generated, prioritizing open and reproducible resources to serve different community needs. Beyond the dataset, our findings on prompt engineering for Greek-adapted LLMs and on metric divergence inform evaluation practice for low- and mid-resource MT. We hope this resource fosters new work in Greek vision-language modeling, image captioning, and multimodal MT.

7 Ethics Statement

The source images and captions in Multi30K originate from Flickr30K and may reflect societal biases, like gender, cultural, geographic, and demographic, present in user-generated web content. Machine-translated versions, including ours, may preserve or amplify such biases. We encourage users of Greek Multi30K to consider these factors when training or evaluating downstream systems, and to apply appropriate bias-detection methodologies. No personally identifying information was collected in the construction of this dataset; the captions describe images already published on Flickr under appropriate licenses. The LLM translations were generated using commercial and open APIs in accordance with their respective terms of service.

8 Limitations

Our study, while comprehensive, is subject to limitations that should be considered when interpreting the results and using the released dataset.

8.1 Evaluation Methodology

Our evaluation relies on automated metrics and an LLM-as-a-Judge protocol; both are proxies for human judgment. While COMET and LLM-as-Judge frameworks have shown strong correlation with human evaluation in English-centric settings (Freitag et al., 2024; Zheng et al., 2023), their reliability for Modern Greek specifically has not been validated against native-speaker judgments in this work. The released dataset has not undergone systematic human validation; future work should include native-speaker spot-checking of a stratified sample from each release version.

Our system selection was based on a stratified 200-sentence sample, manually curated along length and complexity axes (Section 3.1). While this design ensures coverage of diverse sentence types, the manual selection introduces a degree of subjectivity, and the rankings derived from this subset were extrapolated to translate the full 31,014-sentence dataset. A stratified random sampling protocol would offer stronger reproducibility guarantees in future work.

Additionally, our use of mean rank as an aggregation across metrics, while informative, treats ordinal positions as if they were interval-scaled. We mitigate this by reporting per-metric scores and pairwise statistical tests, but a fully ordinal-aware

aggregation (e.g., Borda count) would be preferable.

8.2 Resource and Generalizability

The released dataset does not include a version translated by DeepL, despite DeepL emerging as the strongest system in our evaluation. This omission reflects a deliberate design choice: we prioritized release versions that the community can freely use, redistribute, modify, and reproduce. DeepL's terms of service permit academic use of its API but restrict redistribution of translations as standalone datasets, which is incompatible with our goal of releasing an openly licensed, reusable resource. Our three released versions (Google Translate, Post-Edited via Llama-Krikri-8B, and Gemma-3-27B) span freely accessible commercial output, open-weights post-editing, and fully open-weights LLM generation respectively. We argue this open-resource focus has greater long-term value for the community than including the highest-scoring system under restrictive terms, though we acknowledge that DeepL's strong performance suggests a meaningful upper bound on what is currently achievable for English-to-Greek caption translation.

The source data itself presents limitations. The image descriptions in the Multi30K dataset are sourced from Flickr and may contain societal biases (e.g., gender, cultural) inherent in the original user-generated content. These biases may be reflected or even amplified in the machine-translated Greek versions. Users of this resource should be aware of this potential.

Finally, our findings on system performance are specific to the domain of short, descriptive image captions. The conclusion that specialized MT systems like DeepL outperform general-purpose LLMs may not generalize to other domains, such as legal, medical, or literary texts, which feature different vocabularies and more complex sentence structures.

References

Anthropic. 2024a. The Claude 3 model family: Opus, Sonnet, Haiku. Claude-3 model card.

Anthropic. 2024b. Claude 3.7 sonnet system card. https://www.anthropic.com/claude-3-7-sonnet-system-card. Accessed May 2025.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 148–155, Toulouse, France.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepL GmbH. 2025. Deepl translator. https://www.deepl.com/translator. Retrieved April 2025.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English–German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frédéric Blain, Tom Kocmi, Jinyi Wang, David Ifeoluwa Adelani, Micol Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? Results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, FL, USA.

Gemma 3 Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team. 2025. Gemma 3 model card. https://ai.google.dev/gemma/docs/core/model_card_3. Accessed June 2025.

Google. 2025. Google translate. https://translate.google.com. Retrieved June 2025.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edition. Prentice Hall, Upper Saddle River, NJ.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic.

A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, and C. Deng. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, Singapore.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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, Philadelphia, PA, USA.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. In *Computación y Sistemas*, volume 23, pages 1499–1505.

Maja Popović. 2015. chrf: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online.

Dimitris Roussis, Leonidas Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vasilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2025. Krikri: Advancing open large language models for Greek. *arXiv preprint arXiv:2505.13772v2*.

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, Cambridge, MA, USA.

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, and 1 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144v2*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Z. Zhang, S. N. Syed Abdullah, M. A. R. Abdullah, and L. Zhou. 2025. Google Translate or ChatGPT-4? A multi-metric evaluation of Chinese-to-English technical translation. *Forum for Linguistic Studies*, 7(9):770–788.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, LA, USA.