Automated Multi-Language to English Machine Translation Using Generative Pre-Trained Transformers

Elijah Pelofske*[1], Vincent Urias[1], and Lorie M. Liebrock[2]

[1]Sandia National Laboratories
[2]New Mexico Cybersecurity Center of Excellence, New Mexico Tech

Abstract

The task of accurate and efficient language translation is an extremely important information processing task. Machine learning enabled and automated translation that is accurate and fast is often a large topic of interest in the machine learning and data science communities. In this study, we examine using local Generative Pretrained Transformer (GPT) models to perform automated zero shot black-box, sentence wise, multi-natural-language translation into English text. We benchmark 16 different open-source GPT models, with no custom fine-tuning, from the Huggingface LLM repository for translating 50 different non-English languages into English using translated TED Talk transcripts as the reference dataset. These GPT model inference calls are performed strictly locally, on single A100 Nvidia GPUs. Benchmark metrics that are reported are language translation accuracy, using BLEU, GLEU, METEOR, and chrF text overlap measures, and wall-clock time for each sentence translation. The best overall performing GPT model for translating into English text for the BLEU metric is `ReMM-v2-L2-13B` with a mean score across all tested languages of 0.152, for the GLEU metric is `ReMM-v2-L2-13B` with a mean score across all tested languages of 0.256, for the chrF metric is `Llama2-chat-AYT-13B` with a mean score across all tested languages of 0.448, and for the METEOR metric is `ReMM-v2-L2-13B` with a mean score across all tested languages of 0.438.

1 Introduction

Large Language Models (LLMs), specifically transformer based architecture [1], have been shown to be incredibly effective at learning tasks that require significant abstraction. Generative Pre-Trained Transformers (GPT) [2] have been used to demonstrate numerous highly consequential learning and information processing tasks [3], including code generation [4–6], text summarization [7–10], and chemistry experimental design [11].

In this study, we examine the capabilities of GPT models for the task of translating natural language text in an automated black-box fashion. Multi-language translation using GPT models has been investigated before using OpenAI's GPT models [12], and using deep learning [13]. In this study, we evaluate 16 open source GPT models, run locally and offline in order to assess the effectiveness of black-box translation using current local GPT models. We consider the language translation task of going from 50 natural languages into English text, using the dataset of translated TED talk transcripts.

This study is motivated by machine translation being of fundamental interest in computing and information sharing, and given the evident demonstrations of GPT model capabilities, it makes sense to evaluate how well current GPT models perform at this task. Many GPT chat models are available to users as cloud based resources. However, there are significant privacy and security concerns with this model of computation. Therefore, we are interested in using offline, entirely local, GPT inference calls. This also lets us quantify the scale of the computation required for a task such as automated multi-language machine translation - in this case we use single A100 GPU's to perform the inference for each model. Lastly, we aim to evaluate the *automated* machine translation capabilities of the current GPT models - in particular we do not heavily optimize the inference hyperparameters, or the chat prompts. The goal is to measure a reasonably large and language agnostic (e.g., not prompt tuned for each language) benchmark of the translation capabilities of these models. The GPT translation quality is compared against the `Google translate` API, in Python [14].

*E-mail: elijah.pelofske@protonmail.com

Model name	Reference(s)	Context Length	Architecture type	Model Size
zephyr-7b-alpha	[15]	32768 Tokens	mistral	7.24B params
zephyr-7b-beta	[15, 16]	32768 Tokens	mistral	7.24B params
Mistral-7B-Instruct-v0.1	[17]	32768 Tokens	mistral	7.24B params
Turdus	[18]	32768 Tokens	mistral	7.24B params
vicuna-7b-v1.5	[19, 20]	4096 Tokens	llama	7B params
phi-2	[21]	2048 Tokens	phi	2.78B params
phi-1	[22]	2048 Tokens	phi	1.3B params
phi-1_5	[23]	2048 Tokens	phi	1.3B params
ReMM-v2-L2-13B	[24]	4096 Tokens	llama	13B params
wizardLM-7B-HF	[25]	2048 Tokens	llama	7B params
wizardLM-13B-1.0-fp16	[25]	2048 Tokens	llama	13B params
Llama-2-13b-chat-hf	[20]	4096 Tokens	llama	13B params
Llama2-chat-AYT-13B	[20, 26]	4096 Tokens	llama	13B params
TinyLlama-1.1B-Chat-v1.0	[27–29]	2048 Tokens	llama	1.1B params
gpt4all-13b-snoozy	[30]	2048 Tokens	llama	13B params
falcon-7b-instruct	[31, 32]	2048 Tokens	falcon	7B params

Table 1: Summary of the 16 Generative Pre-trained Transformers models used in this study

2 Methods

The GPT models used in this study are summarized in Table 1 - the model weights were downloaded from huggingface [33], where the trained model weights are open sourced. Each of these GPT models have been fine tuned, with varying levels of success, to be prompted in a chat-type mode. These models run using the PyTorch python library [34]. The context window for the GPT models, summarized in Table 1, is not always clearly defined, but for several of the models the context window is given explicitly in the model weights repository. In other cases, the context window is in the metadata under the parameter `max_position_embeddings`, `n_embd`, or is not explicitly stated. No fine-tuning of the model weights is performed, all 16 of these GPT models are evaluated as-is in this black-box benchmarking comparison for language translation. Importantly, the underlying architectures of all of these GPT models rely on a large number of remarkable machine learning developments in recent years, many of which are described in refs. [1, 3, 29, 35–38].

The translation dataset is a set of Ted Talk transcripts aggregated by the study in ref. [39]. Specifically, for 50 of the foreign languages in the transcript dataset, $1,000$ of those sentences are translated into English. Due to the nature of the dataset, the same $1,000$ sentences are not necessarily translated across the 50 foreign languages (many of the transcript translations are incomplete). Then, those translated sentences are compared against the corresponding reference English sentence. This GPT translation is performed on a per-sentence basis because, as detailed in Table 1, each of these models have a maximum token context window that is relatively small compared to the size of a complete document (which could be comprised of tens or hundreds of thousands of tokens). Therefore, we apply the translations for each individual sentence primarily to mitigate the problems that arise if we attempt to generate text that has a longer token length than what the GPT model was designed to process. In order to assess the quality of the translations, four metrics are used; METEOR [40], chrF (CHaRacter-level F-score) [41], BLEU (Bilingual Evaluation Understudy) [42, 43], GLEU (General Language Understanding Evaluation) [44]. The METEOR, GLEU, BLEU, chrF and metrics are computed using NLTK [45], using all default hyper-parameters. The metrics are computed after the reference sentence and the translated sentence have been tokenized, all punctuation is removed, and all text is made lower case in order to strictly evaluate the words used for the translation. All four of these metrics are defined to be in between (or equal to) $[0, 1]$, where 1 indicates the translations completely agree and 0 indicates the translated document shares no overlap with the original reference document. Note that even high-quality human translations do not guarantee a score of 1 for all four metrics; generally, it is a difficult task to capture language translation quality [46, 47]. Additionally, the multi-language dataset that is used in this study is a collection of TED Talk video transcripts, which themselves are not guaranteed to always be accurate. Therefore, when analyzing the translation quality metrics, we should not always expect to be able to reach scores of 1, but rather we should be aiming to get closer to 1 than 0.

Minimal GPT output postprocessing is applied in the form of removing language-agnostic key phrases from the beginning of the generated text, if it matches certain commonly used phrases that are not the actual content of the

translation, such as *This translated text is*. The full list of removed phrases is given in Appendix A.

For each sentence (regardless of the language of the text), the following text prompt is used in order to prompt the GPT model to translate the sentence into English text using a one-shot inference call.

> Translate the following sentence into clearly written English text. Respond only with the translated text; do not write explanations or justifications in your reply.
> Text to be translated

The text that we want translated is put where the phrase `Text to be translated` is in the above prompt example. This prompt is not changed to instruct the GPT model on what the input language is – meaning that this automated translation method has the advantage of being entirely language agnostic, specifically meaning that language detection does not need to be applied so as to have the translation be performed correctly. Or more specifically, this is the prompting method that is applied to the GPT models with the aim of benchmarking how well they perform at the task of automated, and language agnostic, sentence-wise translation. All of the experimental results reported in this study use this fixed prompt so as to simplify the data analysis (and the total required compute time). This prompt was chosen based on minimal small experimentation with prompts that performed reasonably well - but better prompts could likely be found.

The GPT model inference is performed using the Python 3 module *transformers* [33], and each model inference call is performed on a single Nvidia A100 GPU [48] with 82 Gigabytes of memory, with CUDA Version 12.4. The text generation calls are performed using the `pipeline` method in *transformers* [33] using all default parameters, except the inference temperature is set to 0.01 which results in nearly deterministic output where the chosen token at each step of the model is very likely to be the highest probability token. The timing of the inference calls is reported using wall-clock time to generate the translation of each sentence. Importantly, multiple inference calls were performed on several GPUs concurrently - although the computations were independent, the timing statistics that are reported may be slightly greater than what could be achieved on a completed isolated computing platform with no concurrent GPU computations.

Finally, the translation quality from the GPT models is compared against automated translation (performed by supplying only the target language of English) using `Google translate`. This is performed using a python 3 library [14] that calls the `Google translate` API. In cases where the output from the Google API is None, the "translated" text is set to an empty string (this happened for a couple of sentences, but was not very common).

3 Results

Table 2 summarizes the best performing GPT models for translating 50 foreign languages, using the four different translation metrics. Table 2 reports the best mean translation quality per sentence which are given by the rounded value to 3 decimal places. The final aggregate metric of the best performing GPT model across all languages, for each of the language quality metrics, is computed as the mean of the vector of all 50 language scores (this aggregate metric is not weighted by the different amounts of sentences that were translated in the language dataset).

Notably, of the 16 GPT models, only a small subset of these was the best performing for any tuple of language and translation quality metric. Specifically, the models that had the best mean scores (for any combination of language and translation quality measure) were; `ReMM-v2-L2-13B`, `Turdus`, `Llama2-chat-AYT-13B`, `wizardLM-13B-1.0-fp16`, and `zephyr-7b-alpha`. `ReMM-v2-L2-13B` was the best performing model overall. Importantly, for each language, the translation scores shown in Table 2 were computed on the exact same translated sentences, but the best performing GPT model was not always the same across the 4 translation quality metrics.

On average, the best performing of the 16 GPT models were not always able to generate good translations. The languages that the GPT models scored the lowest on were Mongolian, Burmese, Kazakh, Kurdish, Armenian, and Georgian.

Figures 1, 2, 3, 4, and 5 shows detailed performance and wall-clock timing statistics for Spanish, French, Chinese, Arabic, and Hindi – which are the five most commonly spoken natural languages (besides English). These plots are representative of the expected language translation quality for the most commonly used languages. Figures 6, 7, 8 show detailed per-GPT model performance for Mongolian, Kazakh, and Georgian, which were languages for which the GPT models were unable to produce good translations for, on average. The detailed per-GPT translation metrics and timing statistics for translating all of the other 50 languages into English are enumerated in Appendix B.

There are a number of consistent trends seen in the translation quality box-plot figures - namely that the three `phi` models generally have very low accuracy. `Llama-2-13b-chat-hf`, notably, also consistently has very low

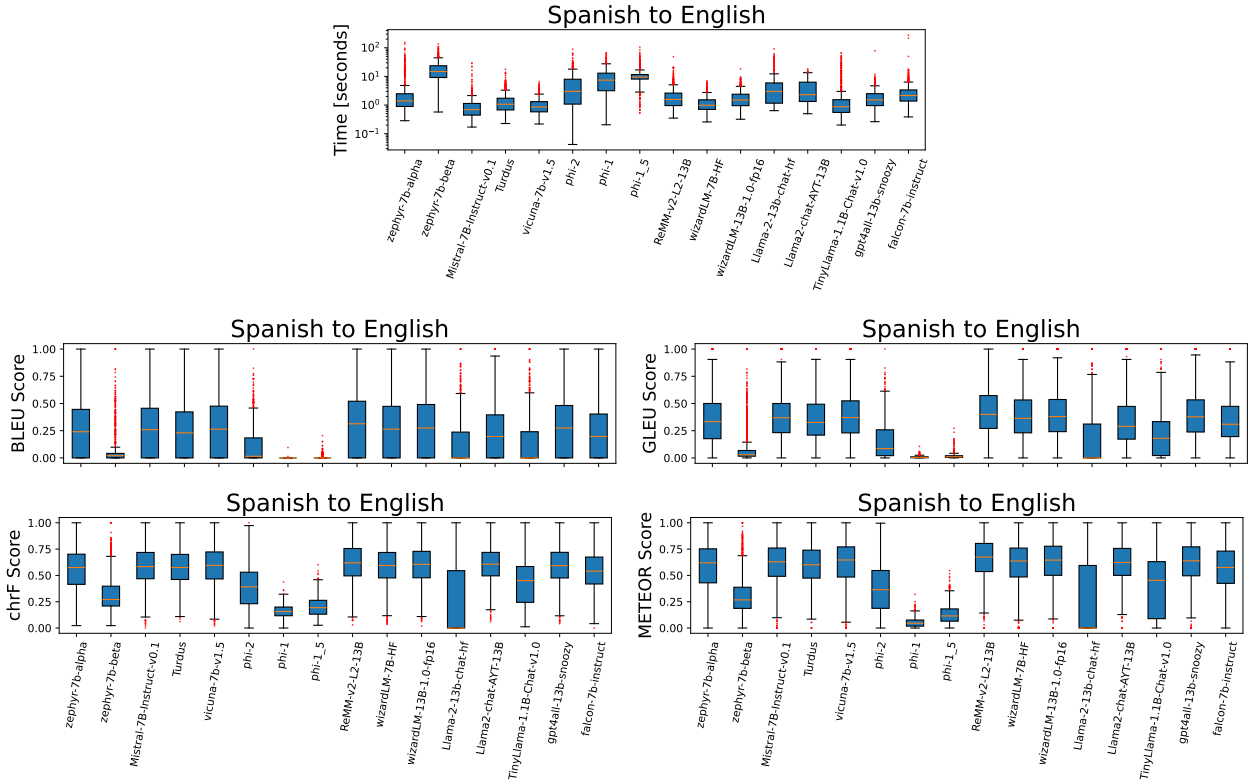

Figure 1: Spanish-to-English dataset per-sentence translation quality and timing statistics for each of the 16 GPT models. Timing is reported in the top sub-plot, on a log scale y-axis, using direct wall-clock compute time to produce the generated text per sentence. Datapoints which are smaller in the time plot mean that the GTP model output took less wall-clock time to generate. The bottom four sub-plots report the distribution of language quality metrics (one datapoint for each sentence), using the four different language quality measures. For all four of the language quality translation plots, scores closer to 1 indicate better translation quality, and scores near 0 indicate bad translation quality. All distributions are shown as box-plot representations, where the red dots indicate outlier points and the blue rectangles indicate the region between the first and third quartile's, the orange line denotes the median.

translation accuracy, which is surprising because nearly all of the best performing models were fine-tuned from Llama-2 models. The mechanism that caused this low accuracy is not clear, but this behavior could be due to the particular prompt that was used and testing other prompts could improve the translation accuracy for future study. In terms of translation speed, the slowest GPT models were `phi-1`, `phi-2`, `phi-1_5`, `zephyr-7b-beta`, and `falcon-7b-instruct`.

Table 3 shows the mean translation quality metrics, for the four language metrics, across all 50 languages being translated into English, using `Google translate`. The same test sentences translated by the GPT models, for each language, were also translated using `Google translate` - therefore the entries in Table 3 should be compared against the best performing GPT models in Table 2. These results show the performance of `Google translate`, using it as a reasonable performance benchmark for automated machine translation of languages. Interestingly, there were exactly two languages where, for at least one of the language metrics (although, in these cases it was for all four language quality metrics), the best performing GPT model had better mean sentence translation quality than google translate. These two languages were French and Chinese. For all other languages, either the best performing GPT model was definitively worse at translating, or was comparable to within a small margin. The languages for which the best performing GPT model and google translate performed marginally the same were German, Spanish, Italian, Russian, Korean, Serbian, Japanese, Ukrainian, Vietnamese, and Bosnian.

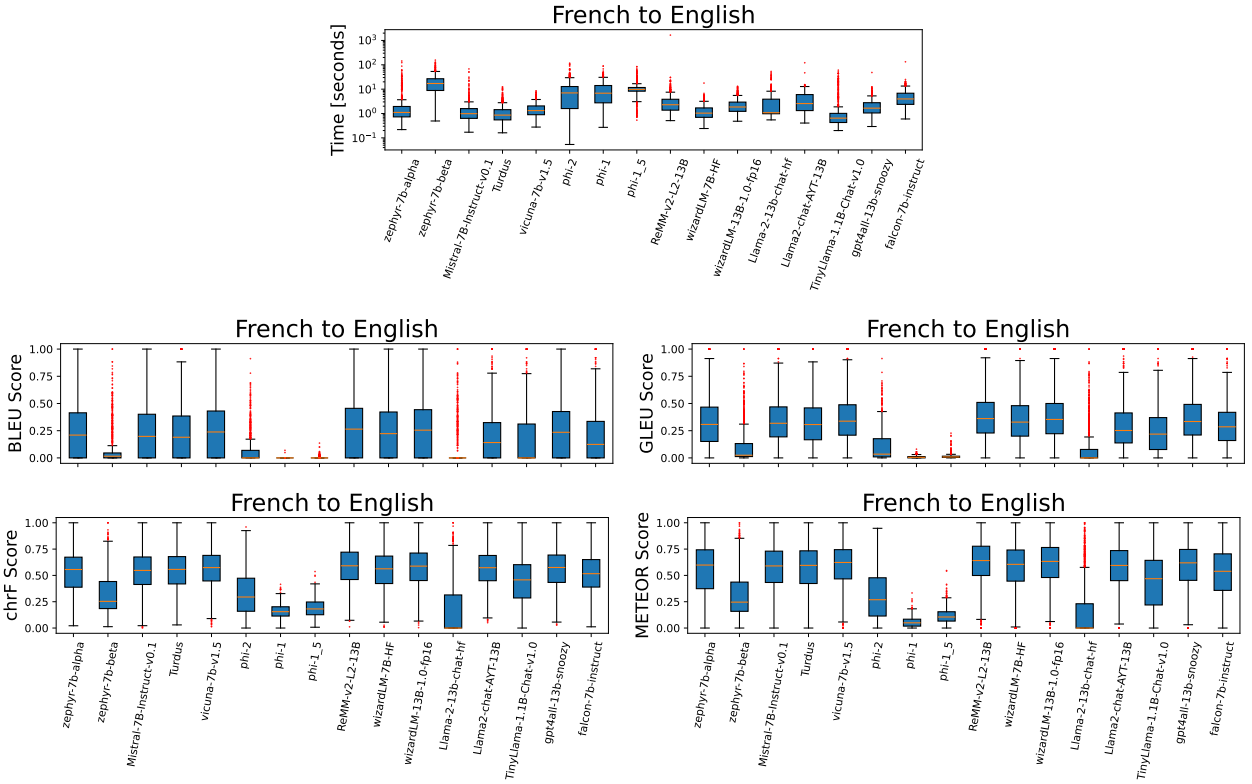

Figure 2: French-to-English dataset per-sentence translation quality and timing statistics for each of the 16 GPT models. Timing is reported in the top sub-plot, on a log scale y-axis, using direct wall-clock compute time to produce the generated text per sentence. Datapoints which are smaller in the time plot mean that the GTP model output took less wall-clock time to generate. The bottom four sub-plots report the distribution of language quality metrics (one datapoint for each sentence), using the four different language quality measures. For all four of the language quality translation plots, scores closer to 1 indicate better translation quality, and scores near 0 indicate bad translation quality. All distributions are shown as box-plot representations, where the red dots indicate outlier points and the blue rectangles indicate the region between the first and third quartile's, the orange line denotes the median.

3.1 Translation Quality Metrics and Example Translations

The following are some examples where the translations produced by the GPT models are reasonable, but the language quality scores are not very close to 1. These examples are shown with the aim of conveying that the overall translation quality for many of the GPT models is quite good even if the mean language quality scores are on average not incredibly close to 1. Importantly, most of the reason for this is that the translation quality metrics are computed for individual sentences, not the entirety of the translated document - and this can lead to unstable measurements of translation quality. However, the mean of the sentence translation quality is a good representation of the overall translation quality – in particular the language quality metrics over the entire translated corpus are very similar (but not necessarily equal) to the mean of the translation metrics across all of the component sentences.

This is an example sentence translation from Spanish into English from the TED talk dataset where the translated sentence has a GLEU score of 0.435, a BLEU score of 0.320, a chrF score of 0.780, and a METEOR score of 0.864. Note that both of these sentences have been tokenized before the score was computed and are shown in their tokenized form.

Reference English sentence: this is a viking lander photograph of the surface of mars
GPT translated sentence from Spanish into English: this is a photograph from the viking lander on the surface of mars

This is another Spanish to English sentence translation where the translated sentence has a GLEU score of

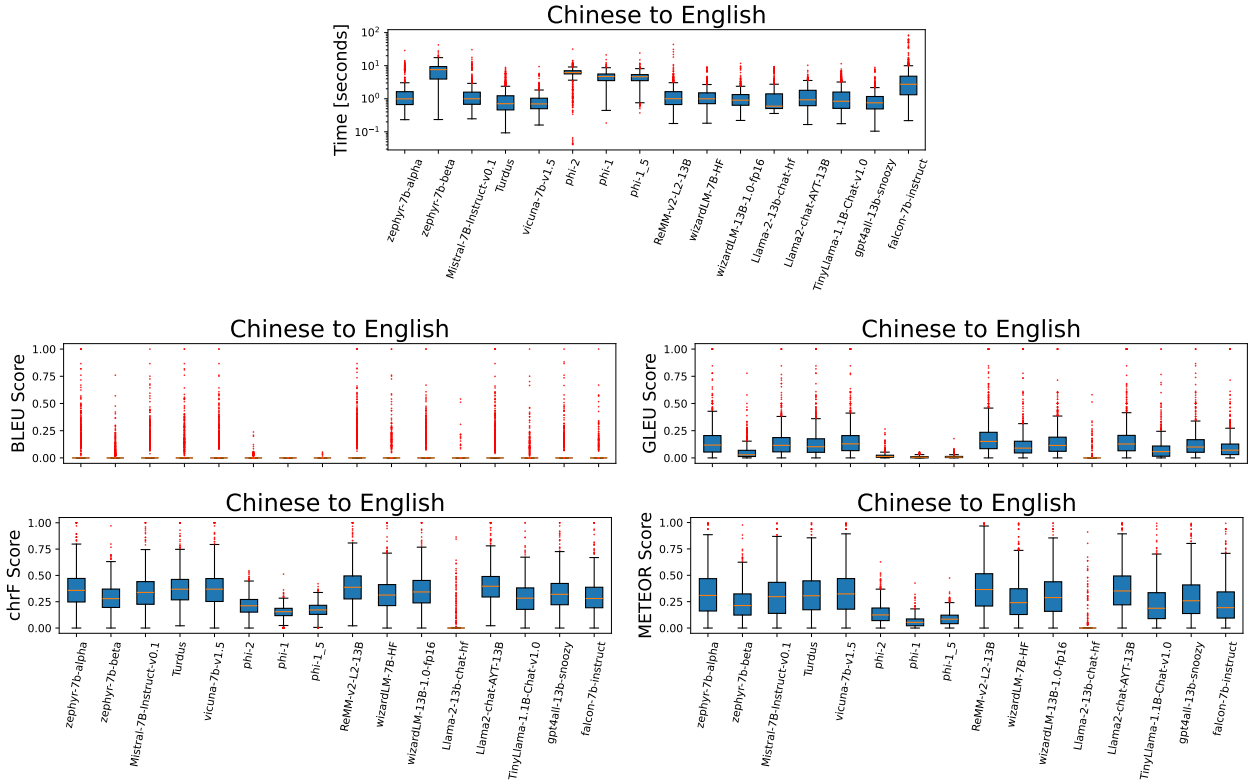

Figure 3: Chinese-to-English dataset per-sentence translation quality and timing statistics

0.427, a BLEU score of 0.379, a chrF score of 0.645, and a METEOR score of 0.725:

> **Reference English sentence:** but there is intriguing evidence that suggests that the early history of mars there may have been rivers and fast flowing water
> **GPT translated sentence from Spanish into English:** there is intriguing evidence suggesting that the early history of mars may have had rivers and streams of water

This is an example sentence translation from Spanish into English which had a GLEU score of 0.481, a BLEU score of 0.429, a chrf score of 0.629, and a METEOR score of 0.735:

> **Reference English sentence:** the answer is no there is no liquid water on the surface of mars today
> **GPT translated sentence from Spanish into English:** there is no water liquid on the surface of mars today

This is an example sentence translation from French into English which had a GLEU score of 0.587, a BLEU score of 0.556, a chrF score of 0.718, and a METEOR score of 0.825:

> **Reference English sentence:** i want to talk to you about one of the biggest myths in medicine and that is the idea that all we need are more medical breakthroughs and then all of our problems will be solved
> **GPT translated sentence from French into English:** i want to talk about one of the greatest myths of medicine and that is the idea that all we need are additional medical procedures and then all our problems will be solved

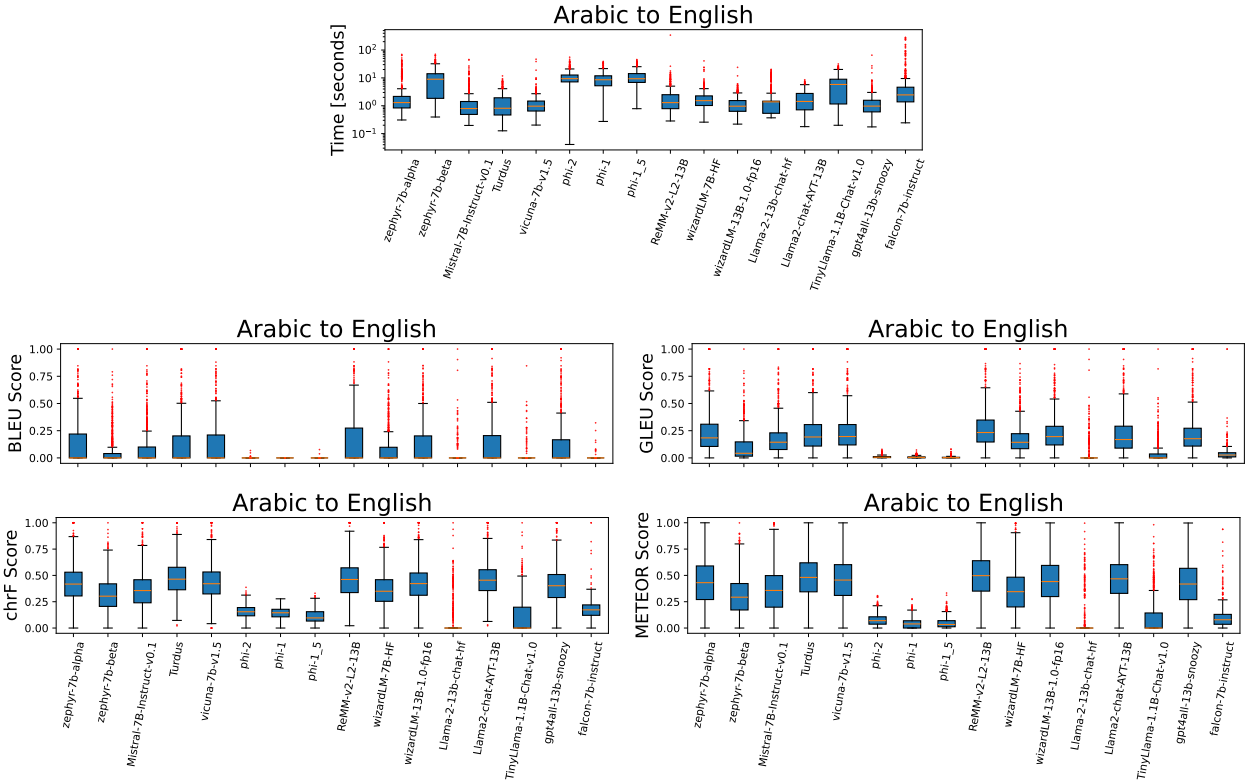

Figure 4: Arabic-to-English dataset per-sentence translation quality and timing statistics

4 Discussion and Conclusion

The translation quality provided by sentence-wise GPT translations showed a clear stratification of the capabilities of the evaluated 16 GPT models. The best performing GPT models, across all 50 foreign languages, for translating into English is `ReMM-v2-L2-13B` and `Llama2-chat-AYT-13B`. This shows that language translation could serve as a clear, and very application-relevant, benchmark for GPT capabilities for processing natural language.

The best performing GPT model translations compare very well against automated machine translation using the Google translate API, although typically Google translate has marginally better scores. Importantly, the GPT model computations offer the security advantage of performing the computations locally, meaning that locally run GPT model automated language translation may be a good alternative depending on the importance of the security and the privacy of handling the information.

The GPT models do not uniformly perform well though – several of the tested models performed noticeably worse than other GPT models, and these trends are consistent across all of the 50 tested languages. Interestingly, there were also some languages that were not able to be translated well by any of the GPT models, for example Mongolian, Kazakh, Burmese, Kurdish, Armenian, and Georgian. This could be due to these languages being relatively low-resource in the training data used when training these GPT models. Notably, there were several GPT models that were consistently the slowest across the different languages; `phi-1`, `phi-2`, `phi-1_5`, `zephyr-7b-beta`, and `falcon-7b-instruct`.

5 Acknowledgments

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The

Language	Best Mean GLEU	Best Mean BLEU	Best Mean chrF	Best Mean METEOR
Arabic	ReMM-v2-L2-13B (0.271)	ReMM-v2-L2-13B (0.157)	Turdus (0.478)	ReMM-v2-L2-13B (0.489)
Azerbaijani	ReMM-v2-L2-13B (0.121)	ReMM-v2-L2-13B (0.035)	Turdus (0.316)	Turdus (0.257)
Belarusian	ReMM-v2-L2-13B (0.2)	ReMM-v2-L2-13B (0.093)	Llama2-chat-AYT-13B (0.392)	ReMM-v2-L2-13B (0.369)
Bulgarian	ReMM-v2-L2-13B (0.363)	ReMM-v2-L2-13B (0.246)	ReMM-v2-L2-13B (0.554)	ReMM-v2-L2-13B (0.582)
Bengali	ReMM-v2-L2-13B (0.121)	ReMM-v2-L2-13B (0.028)	Turdus (0.335)	Turdus (0.282)
Bosnian	ReMM-v2-L2-13B (0.342)	ReMM-v2-L2-13B (0.229)	Llama2-chat-AYT-13B (0.546)	ReMM-v2-L2-13B (0.559)
Czech	ReMM-v2-L2-13B (0.326)	ReMM-v2-L2-13B (0.207)	Llama2-chat-AYT-13B (0.518)	ReMM-v2-L2-13B (0.546)
Danish	ReMM-v2-L2-13B (0.454)	ReMM-v2-L2-13B (0.368)	ReMM-v2-L2-13B (0.632)	ReMM-v2-L2-13B (0.673)
German	ReMM-v2-L2-13B (0.361)	ReMM-v2-L2-13B (0.238)	ReMM-v2-L2-13B (0.555)	ReMM-v2-L2-13B (0.579)
Greek	ReMM-v2-L2-13B (0.294)	ReMM-v2-L2-13B (0.176)	ReMM-v2-L2-13B (0.467)	ReMM-v2-L2-13B (0.499)
Spanish	ReMM-v2-L2-13B (0.443)	ReMM-v2-L2-13B (0.331)	ReMM-v2-L2-13B (0.624)	ReMM-v2-L2-13B (0.658)
Estonian	ReMM-v2-L2-13B (0.134)	Turdus (0.044)	Turdus (0.349)	Turdus (0.287)
Persian	ReMM-v2-L2-13B (0.226)	ReMM-v2-L2-13B (0.112)	Llama2-chat-AYT-13B (0.434)	Llama2-chat-AYT-13B (0.428)
Finnish	ReMM-v2-L2-13B (0.304)	ReMM-v2-L2-13B (0.192)	ReMM-v2-L2-13B (0.523)	ReMM-v2-L2-13B (0.526)
French	ReMM-v2-L2-13B (0.394)	ReMM-v2-L2-13B (0.278)	ReMM-v2-L2-13B (0.585)	ReMM-v2-L2-13B (0.614)
Galician	ReMM-v2-L2-13B (0.317)	ReMM-v2-L2-13B (0.192)	Llama2-chat-AYT-13B (0.53)	Llama2-chat-AYT-13B (0.528)
Hebrew	ReMM-v2-L2-13B (0.267)	ReMM-v2-L2-13B (0.149)	Turdus (0.463)	ReMM-v2-L2-13B (0.477)
Hindi	ReMM-v2-L2-13B (0.191)	ReMM-v2-L2-13B (0.084)	Llama2-chat-AYT-13B (0.396)	ReMM-v2-L2-13B (0.379)
Croatian	ReMM-v2-L2-13B (0.345)	ReMM-v2-L2-13B (0.237)	ReMM-v2-L2-13B (0.541)	ReMM-v2-L2-13B (0.561)
Hungarian	ReMM-v2-L2-13B (0.285)	ReMM-v2-L2-13B (0.171)	ReMM-v2-L2-13B (0.487)	ReMM-v2-L2-13B (0.495)
Armenian	Turdus (0.096)	Turdus (0.019)	Turdus (0.294)	Turdus (0.228)
Indonesian	ReMM-v2-L2-13B (0.29)	ReMM-v2-L2-13B (0.166)	Llama2-chat-AYT-13B (0.495)	ReMM-v2-L2-13B (0.514)
Italian	ReMM-v2-L2-13B (0.375)	ReMM-v2-L2-13B (0.262)	ReMM-v2-L2-13B (0.561)	ReMM-v2-L2-13B (0.59)
Japanese	ReMM-v2-L2-13B (0.186)	ReMM-v2-L2-13B (0.078)	Llama2-chat-AYT-13B (0.403)	Llama2-chat-AYT-13B (0.372)
Georgian	ReMM-v2-L2-13B (0.094)	ReMM-v2-L2-13B (0.017)	Turdus (0.292)	Turdus (0.206)
Kazakh	Turdus (0.053)	Turdus (0.006)	Turdus (0.237)	Turdus (0.138)
Korean	ReMM-v2-L2-13B (0.238)	ReMM-v2-L2-13B (0.116)	ReMM-v2-L2-13B (0.452)	ReMM-v2-L2-13B (0.446)
Kurdish	ReMM-v2-L2-13B (0.044)	Turdus (0.002)	Turdus (0.225)	Llama2-chat-AYT-13B (0.12)
Lithuanian	Turdus (0.12)	zephyr-7b-alpha (0.037)	Turdus (0.328)	Turdus (0.269)
Macedonian	ReMM-v2-L2-13B (0.276)	ReMM-v2-L2-13B (0.152)	Llama2-chat-AYT-13B (0.486)	ReMM-v2-L2-13B (0.479)
Mongolian	ReMM-v2-L2-13B (0.038)	wizardLM-13B-1.0-fp16 (0.002)	Turdus (0.21)	Turdus (0.095)
Malay	ReMM-v2-L2-13B (0.271)	ReMM-v2-L2-13B (0.157)	ReMM-v2-L2-13B (0.472)	ReMM-v2-L2-13B (0.486)
Burmese	ReMM-v2-L2-13B (0.028)	ReMM-v2-L2-13B (0.002)	Turdus (0.228)	Turdus (0.096)
Norwegian	ReMM-v2-L2-13B (0.413)	ReMM-v2-L2-13B (0.299)	ReMM-v2-L2-13B (0.597)	ReMM-v2-L2-13B (0.629)
Dutch	ReMM-v2-L2-13B (0.387)	ReMM-v2-L2-13B (0.277)	Llama2-chat-AYT-13B (0.561)	ReMM-v2-L2-13B (0.59)
Polish	ReMM-v2-L2-13B (0.292)	ReMM-v2-L2-13B (0.177)	Llama2-chat-AYT-13B (0.492)	ReMM-v2-L2-13B (0.494)
Portuguese	ReMM-v2-L2-13B (0.441)	ReMM-v2-L2-13B (0.32)	ReMM-v2-L2-13B (0.618)	ReMM-v2-L2-13B (0.643)
Romanian	ReMM-v2-L2-13B (0.367)	ReMM-v2-L2-13B (0.255)	ReMM-v2-L2-13B (0.562)	ReMM-v2-L2-13B (0.58)
Russian	ReMM-v2-L2-13B (0.305)	ReMM-v2-L2-13B (0.182)	ReMM-v2-L2-13B (0.501)	ReMM-v2-L2-13B (0.516)
Slovak	ReMM-v2-L2-13B (0.287)	ReMM-v2-L2-13B (0.166)	Llama2-chat-AYT-13B (0.493)	ReMM-v2-L2-13B (0.506)
Slovenian	ReMM-v2-L2-13B (0.254)	ReMM-v2-L2-13B (0.14)	ReMM-v2-L2-13B (0.46)	ReMM-v2-L2-13B (0.459)
Albanian	Turdus (0.122)	Turdus (0.041)	Turdus (0.329)	Turdus (0.266)
Serbian	ReMM-v2-L2-13B (0.347)	ReMM-v2-L2-13B (0.233)	ReMM-v2-L2-13B (0.535)	ReMM-v2-L2-13B (0.559)
Swedish	ReMM-v2-L2-13B (0.408)	ReMM-v2-L2-13B (0.294)	ReMM-v2-L2-13B (0.583)	ReMM-v2-L2-13B (0.619)
Thai	ReMM-v2-L2-13B (0.151)	ReMM-v2-L2-13B (0.051)	Turdus (0.338)	ReMM-v2-L2-13B (0.309)
Turkish	ReMM-v2-L2-13B (0.236)	ReMM-v2-L2-13B (0.118)	Llama2-chat-AYT-13B (0.437)	ReMM-v2-L2-13B (0.432)
Ukrainian	ReMM-v2-L2-13B (0.303)	ReMM-v2-L2-13B (0.178)	ReMM-v2-L2-13B (0.501)	ReMM-v2-L2-13B (0.507)
Urdu	ReMM-v2-L2-13B (0.197)	ReMM-v2-L2-13B (0.085)	Llama2-chat-AYT-13B (0.413)	Llama2-chat-AYT-13B (0.385)
Vietnamese	ReMM-v2-L2-13B (0.274)	ReMM-v2-L2-13B (0.155)	ReMM-v2-L2-13B (0.464)	ReMM-v2-L2-13B (0.497)
Chinese	ReMM-v2-L2-13B (0.186)	ReMM-v2-L2-13B (0.068)	Llama2-chat-AYT-13B (0.405)	ReMM-v2-L2-13B (0.372)
All languages	ReMM-v2-L2-13B (0.256)	ReMM-v2-L2-13B (0.152)	Llama2-chat-AYT-13B (0.448)	ReMM-v2-L2-13B (0.438)

Table 2: Best GPT translation metrics for each language, computed by the best mean translation quality over all tested sentence

Language	Mean GLEU	Mean BLEU	Mean chrF	Mean METEOR
Arabic	0.354	0.225	0.559	0.592
Azerbaijani	0.231	0.099	0.431	0.417
Belarusian	0.289	0.169	0.484	0.486
Bulgarian	0.396	0.273	0.584	0.619
Bengali	0.184	0.078	0.365	0.369
Bosnian	0.381	0.26	0.58	0.611
Czech	0.372	0.238	0.559	0.607
Danish	0.527	0.438	0.689	0.744
German	0.367	0.235	0.557	0.588
Greek	0.379	0.25	0.559	0.602
Spanish	0.438	0.324	0.623	0.66
Estonian	0.32	0.198	0.526	0.545
Persian	0.301	0.188	0.494	0.52
Finnish	0.349	0.234	0.555	0.576
French	0.326	0.202	0.552	0.579
Galician	0.367	0.25	0.564	0.586
Hebrew	0.39	0.262	0.569	0.615
Hindi	0.234	0.113	0.437	0.445
Croatian	0.413	0.282	0.599	0.637
Hungarian	0.329	0.21	0.522	0.551
Armenian	0.28	0.158	0.478	0.477
Indonesian	0.327	0.2	0.527	0.567
Italian	0.372	0.255	0.568	0.601
Japanese	0.204	0.078	0.395	0.387
Georgian	0.254	0.119	0.458	0.452
Kazakh	0.206	0.076	0.383	0.379
Korean	0.25	0.126	0.455	0.461
Kurdish	0.21	0.095	0.394	0.407
Lithuanian	0.308	0.176	0.507	0.52
Macedonian	0.369	0.245	0.571	0.597
Mongolian	0.165	0.037	0.338	0.312
Malay	0.309	0.184	0.509	0.537
Burmese	0.078	0.012	0.195	0.152
Norwegian	0.462	0.351	0.635	0.685
Dutch	0.409	0.292	0.589	0.621
Polish	0.313	0.187	0.507	0.521
Portuguese	0.459	0.335	0.64	0.672
Romanian	0.389	0.268	0.581	0.612
Russian	0.298	0.174	0.497	0.507
Slovak	0.361	0.229	0.561	0.594
Slovenian	0.321	0.197	0.525	0.544
Albanian	0.383	0.256	0.575	0.608
Serbian	0.388	0.261	0.57	0.617
Swedish	0.446	0.329	0.618	0.666
Thai	0.189	0.067	0.36	0.361
Turkish	0.323	0.193	0.514	0.546
Ukrainian	0.305	0.17	0.502	0.513
Urdu	0.302	0.18	0.509	0.529
Vietnamese	0.292	0.173	0.483	0.526
Chinese	0.155	0.039	0.358	0.318

Table 3: Mean translation quality metrics from using the Google translate service, taken across all test sentences

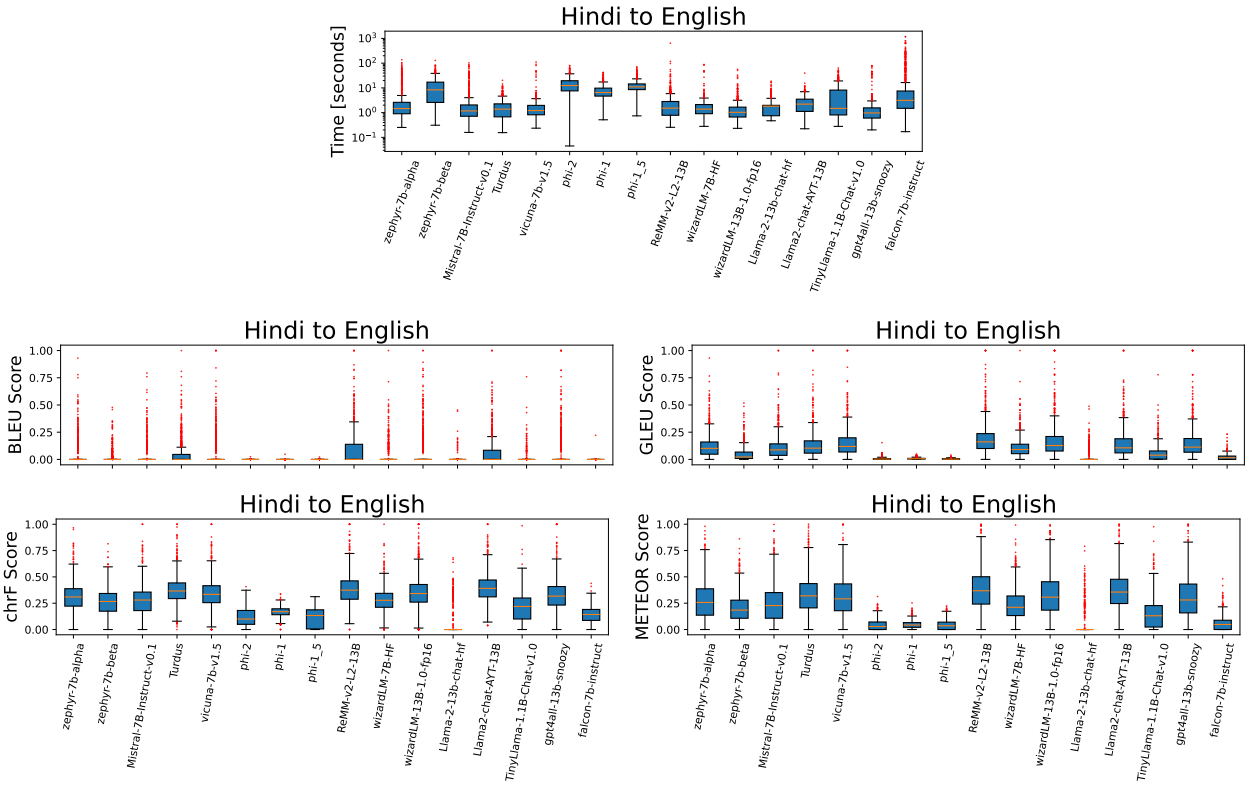

Figure 5: Hindi-to-English dataset per-sentence translation quality and timing statistics

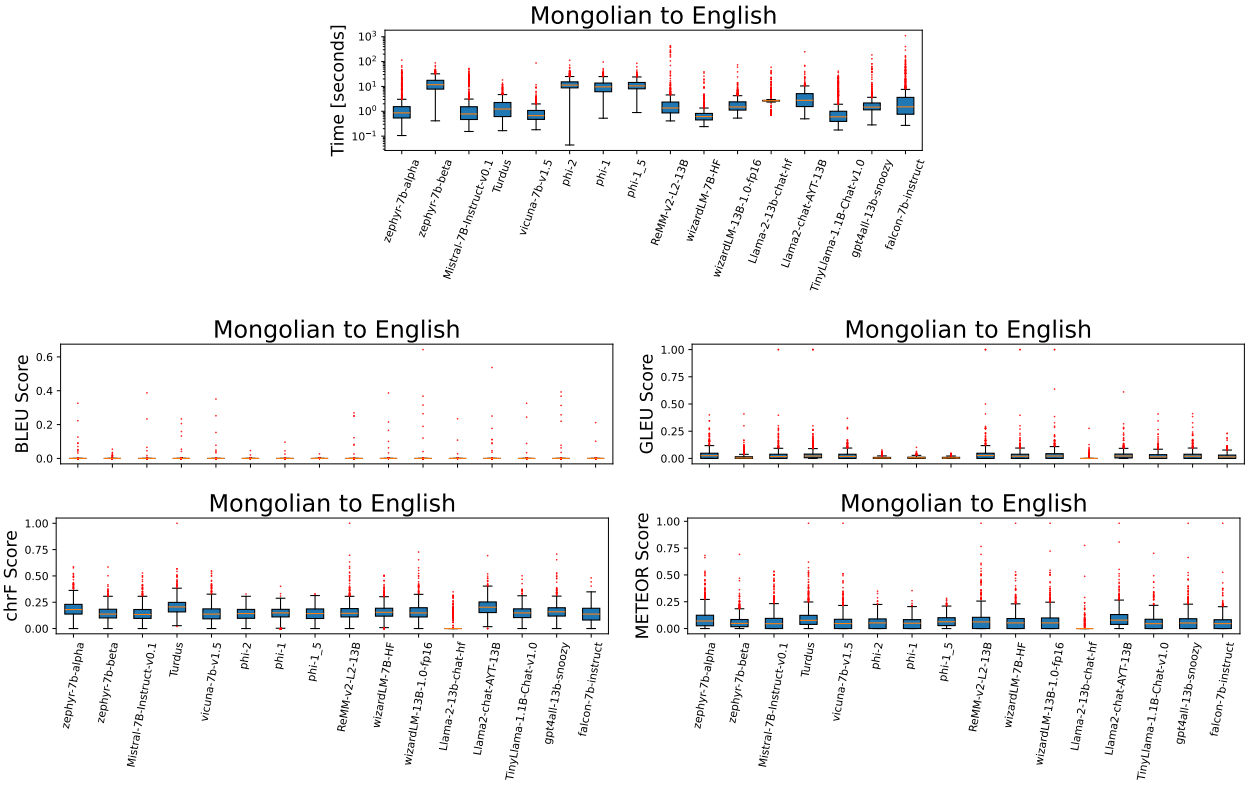

Figure 6: Mongolian-to-English dataset per-sentence translation quality and timing statistics

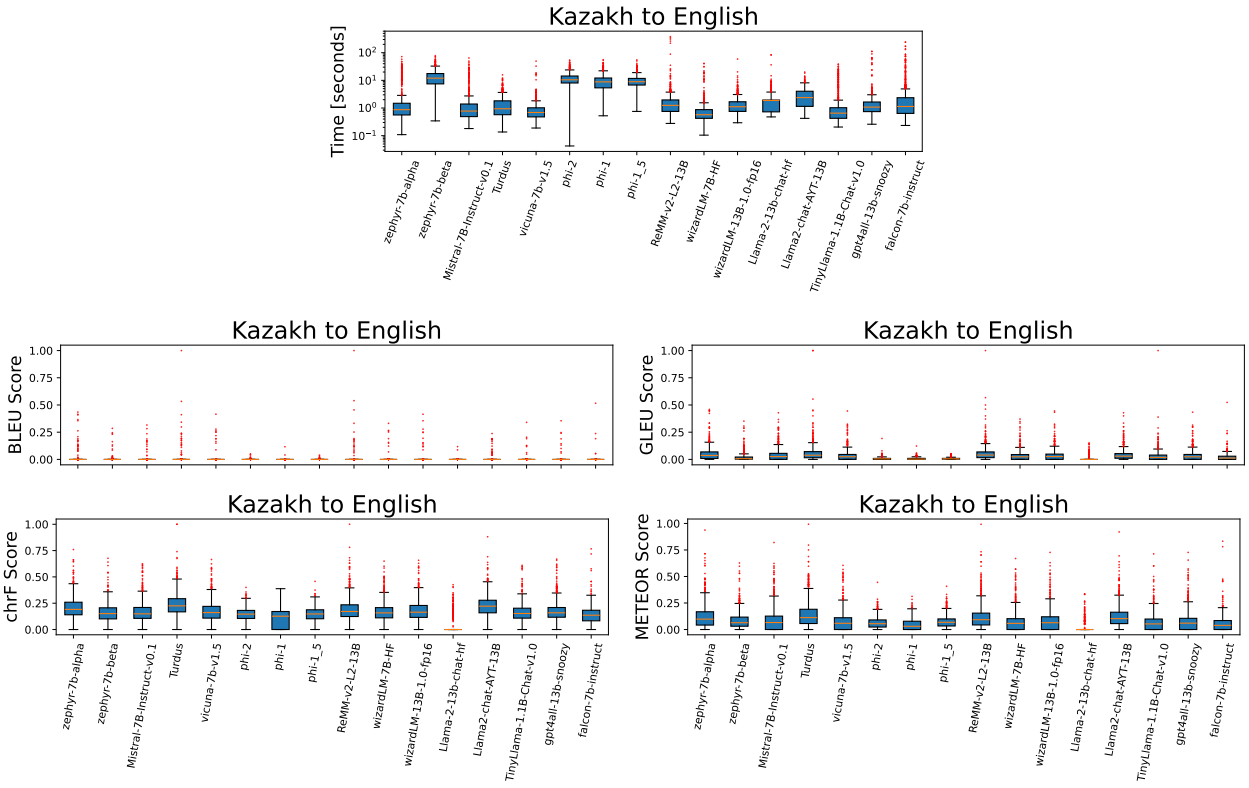

Figure 7: Kazakh-to-English dataset per-sentence translation quality and timing statistics

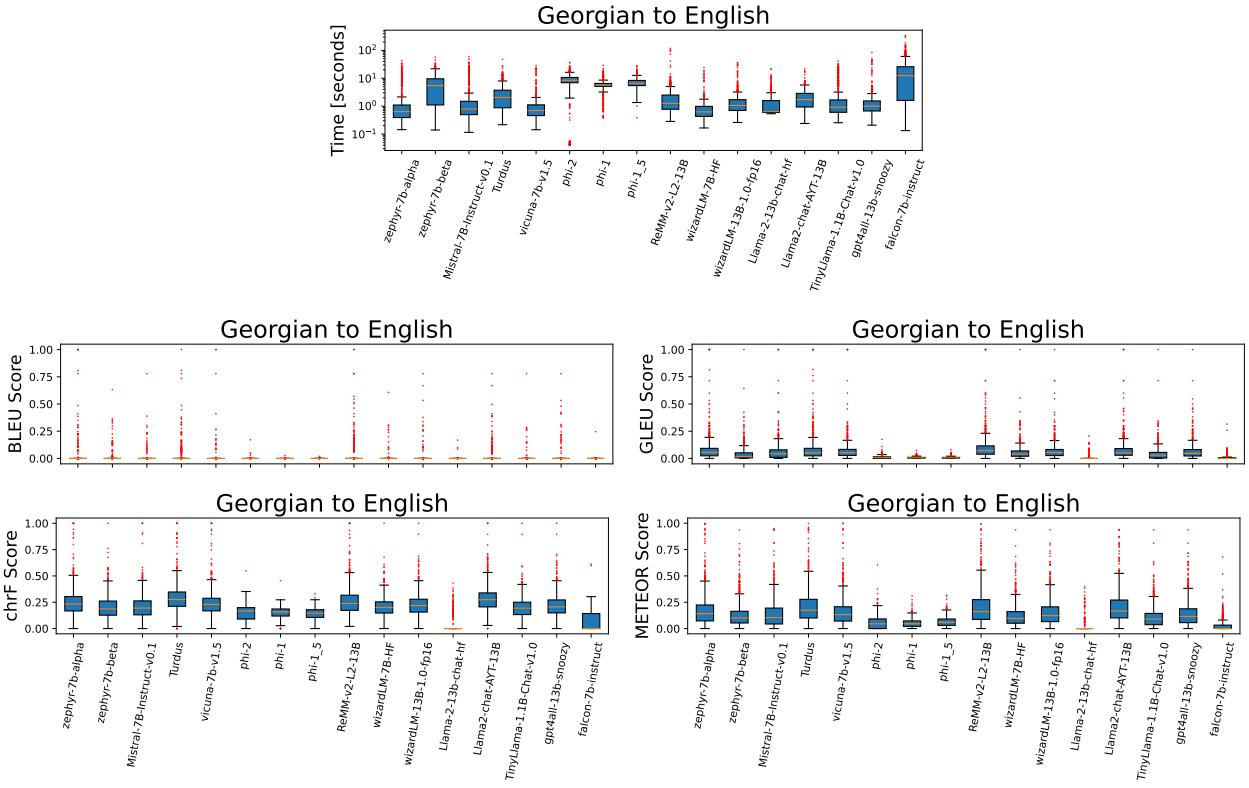

Figure 8: Georgian-to-English dataset per-sentence translation quality and timing statistics

References

[1] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[2] Gokul Yenduri et al. *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. 2023. arXiv: 2305.10435 [cs.CL].

[3] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[4] Aishwarya Narasimhan, Krishna Prasad Agara Venkatesha Rao, and Veena M B. *CGEMs: A Metric Model for Automatic Code Generation using GPT-3*. 2021. arXiv: 2108.10168 [cs.AI].

[5] Li Zhong and Zilong Wang. *Can ChatGPT replace StackOverflow? A Study on Robustness and Reliability of Large Language Model Code Generation*. 2023. arXiv: 2308.10335 [cs.CL].

[6] Theo X. Olausson et al. *Is Self-Repair a Silver Bullet for Code Generation?* 2023. arXiv: 2306.09896 [cs.CL].

[7] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. *News Summarization and Evaluation in the Era of GPT-3*. 2023. arXiv: 2209.12356 [cs.CL].

[8] Sengjie Liu and Christopher G. Healey. *Abstractive Summarization of Large Document Collections Using GPT*. 2023. arXiv: 2310.05690 [cs.AI].

[9] Ahsaas Bajaj et al. *Long Document Summarization in a Low Resource Setting using Pretrained Language Models*. 2021. arXiv: 2103.00751 [cs.CL].

[10] Xiao Pu, Mingqi Gao, and Xiaojun Wan. *Summarization is (Almost) Dead*. 2023. arXiv: 2309.09558 [cs.CL].

[11] Daniil A Boiko et al. "Autonomous chemical research with large language models". In: *Nature* 624.7992 (2023), pp. 570–578.

[12] Amr Hendy et al. *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*. 2023. arXiv: 2302.09210 [cs.CL].

[13] Martin Popel et al. "Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals". In: *Nature communications* 11.1 (2020), p. 4381.

[14] *Google Translate API for Python*. 2024. URL: https://pypi.org/project/googletrans/.

[15] Rafael Rafailov et al. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. 2023. arXiv: 2305.18290 [cs.LG].

[16] Lewis Tunstall et al. *Zephyr: Direct Distillation of LM Alignment*. 2023. arXiv: 2310.16944 [cs.LG].

[17] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL].

[18] UDK dot AI, Daniel Devatman Hromada. *Turdus (Revision 923c305)*. 2024. DOI: 10.57967/hf/1611. URL: https://huggingface.co/udkai/Turdus.

[19] Lianmin Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL].

[20] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[21] *phi-2*. 2024. URL: https://huggingface.co/microsoft/phi-2.

[22] Suriya Gunasekar et al. "Textbooks Are All You Need". In: *arXiv preprint arXiv:2306.11644* (2023).

[23] Yuanzhi Li et al. "Textbooks Are All You Need II: **phi-1.5** technical report". In: *arXiv preprint arXiv:2309.05463* (2023).

[24] *ReMM-v2-L2-13B*. 2024. URL: https://huggingface.co/Undi95/ReMM-v2-L2-13B.

[25] Can Xu et al. *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. 2023. arXiv: 2304.12244 [cs.CL].

[26] Subhabrata Mukherjee et al. *Orca: Progressive Learning from Complex Explanation Traces of GPT-4*. 2023. arXiv: 2306.02707 [cs.CL].

[27] Peiyuan Zhang et al. *TinyLlama: An Open-Source Small Language Model*. 2024. arXiv: 2401.02385 [cs.CL].

[28] Lightning AI. *Lit-GPT*. 2023. URL: https://github.com/Lightning-AI/lit-gpt.

[29] Tri Dao. "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning". In: (2023).

[30] Yuvanesh Anand et al. *GPT4All: Training an Assistant-style Chatbot with Large Scale Data Distillation from GPT-3.5-Turbo*. https://github.com/nomic-ai/gpt4all. 2023.

[31] Guilherme Penedo et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only". In: *arXiv preprint arXiv:2306.01116* (2023). arXiv: 2306.01116. URL: https://arxiv.org/abs/2306.01116.

[32] Ebtesam Almazrouei et al. "Falcon-40B: an open large language model with state-of-the-art performance". In: (2023).

[33] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2020. arXiv: 1910.03771 [cs.CL].

[34] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[35] Jianlin Su et al. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. 2023. arXiv: 2104.09864 [cs.CL].
[36] Tri Dao et al. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. 2022. arXiv: 2205.14135 [cs.LG].
[37] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].
[38] Noam Shazeer. *Fast Transformer Decoding: One Write-Head is All You Need*. 2019. arXiv: 1911.02150 [cs.NE].
[39] Qi Ye et al. "When and Why are pre-trained word embeddings useful for Neural Machine Translation". In: *HLT-NAACL*. 2018.
[40] Satanjeev Banerjee and Alon Lavie. "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*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. URL: https://www.aclweb.org/anthology/W05-0909.
[41] Maja Popovic. "chrF: character n-gram F-score for automatic MT evaluation". In: *WMT@EMNLP*. 2015. URL: https://api.semanticscholar.org/CorpusID:15349458.
[42] Kishore Papineni et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: 2002, pp. 311–318.
[43] Chin-Yew Lin and Franz Josef Och. "ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation". In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: COLING, 2004, pp. 501–507. URL: https://www.aclweb.org/anthology/C04-1072.
[44] Andrew Mutton et al. "GLEU: Automatic evaluation of sentence-level fluency". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 2007, pp. 344–351.
[45] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.
[46] Yonghui Wu et al. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. arXiv: 1609.08144 [cs.CL].
[47] Seungjun Lee et al. "A Survey on Evaluation Metrics for Machine Translation". In: *Mathematics* 11.4 (2023). ISSN: 2227-7390. DOI: 10.3390/math11041006. URL: https://www.mdpi.com/2227-7390/11/4/1006.
[48] Jack Choquette et al. "NVIDIA A100 Tensor Core GPU: Performance and Innovation". In: *IEEE Micro* 41.2 (2021), pp. 29–35. DOI: 10.1109/MM.2021.3061394.

A Key Phrases Removed From GPT Output

Any generated text that starts with any of the following strings has that text removed before the language quality metrics are computed. This list is not complete for the sake of space, but these serve as representative ancillary text that were commonly seen in the GPT model output.

- Translated text:
- Translation:
- The translated text is:
- The sentence translates to:
- The translation is:
- The sentence should be translated to:
- The following sentence is translated into English text:
- Answer: The translation of the sentence is:
- This is the translated text:
- Clear English translation:
- You can translate this sentence into English as:
- This sentence translates to
- Solution: The sentence is translated into clearly written English text as follows:
- The following sentence is translated into clearly written English text:

- This is a clear and accurate translation:
- The sentence has been translated into English as:
- The given sentence translates to:
- A clear English translation of the given sentence is:
- In English, this translates to:
- The following is the translated text:
- The correct translation is:
- This should be translated as:
- Here is the translation of the sentence:
- The text translates to:
- Here is my translation:
- You can translate the sentence as:
- The English translation of the given sentence is:
- This sentence can be translated as:
- This sentence can be translated into English as:
- The following is a clearly written English translation of the provided text:
- The following sentence is translated into English:
- Translate the following sentence into clearly written English text. Respond only with the translated text; do not write explanations or justifications in your reply.

Note that the last entry in the above list is actually just the input prompt; we found some of the GPT models would on occasion re-output the prompt before generating any further text, so this was included in the simple text removal post-processing.

B Complete Translation Quality Measure and Timing Plots

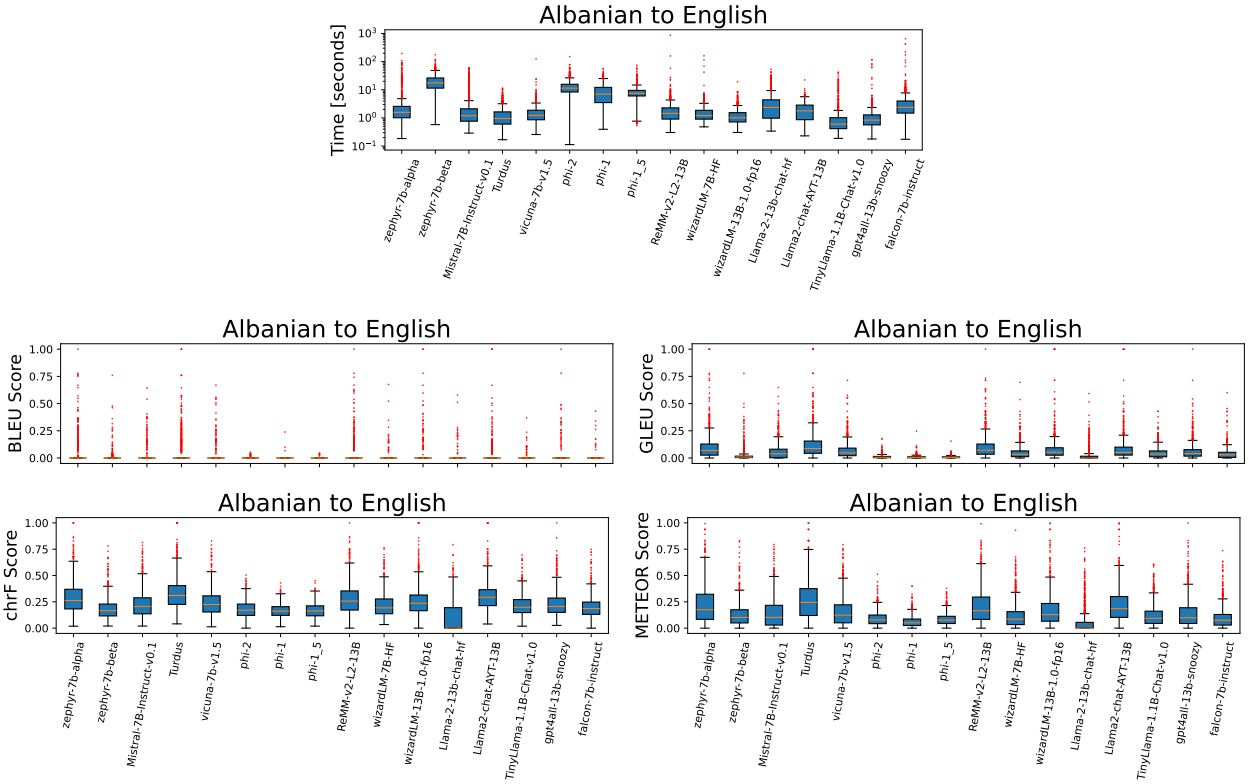

Figure 9: Albanian-to-English dataset per-sentence translation quality and timing statistics

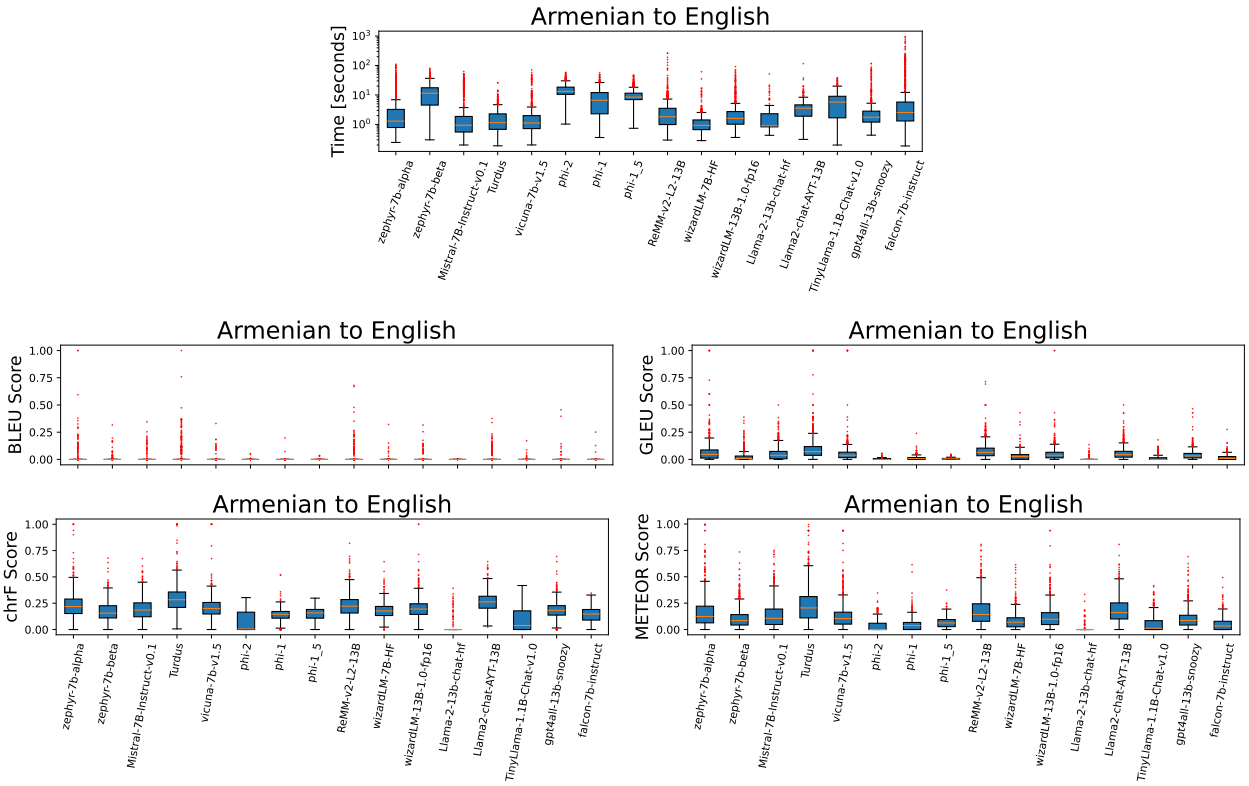

Figure 10: Armenian-to-English dataset per-sentence translation quality and timing statistics

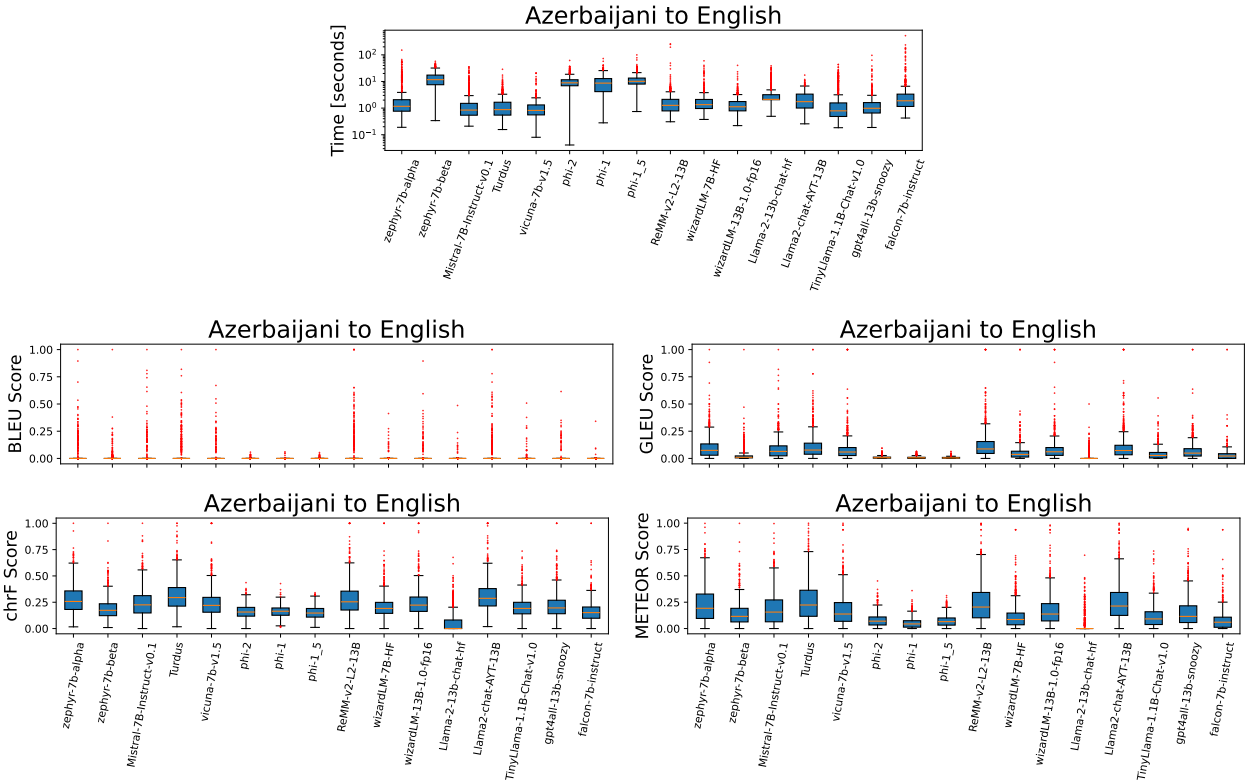

Figure 11: Azerbaijani-to-English dataset per-sentence translation quality and timing statistics

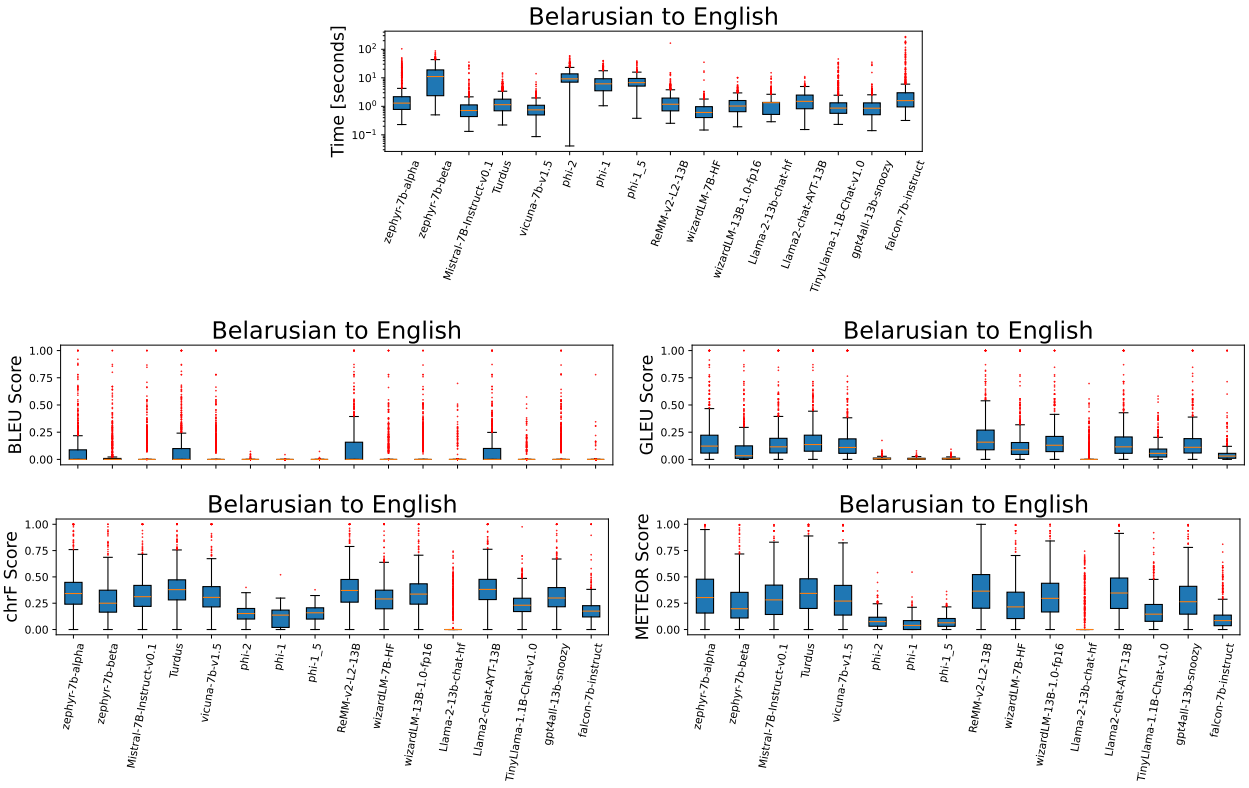

Figure 12: Belarusian-to-English dataset per-sentence translation quality and timing statistics

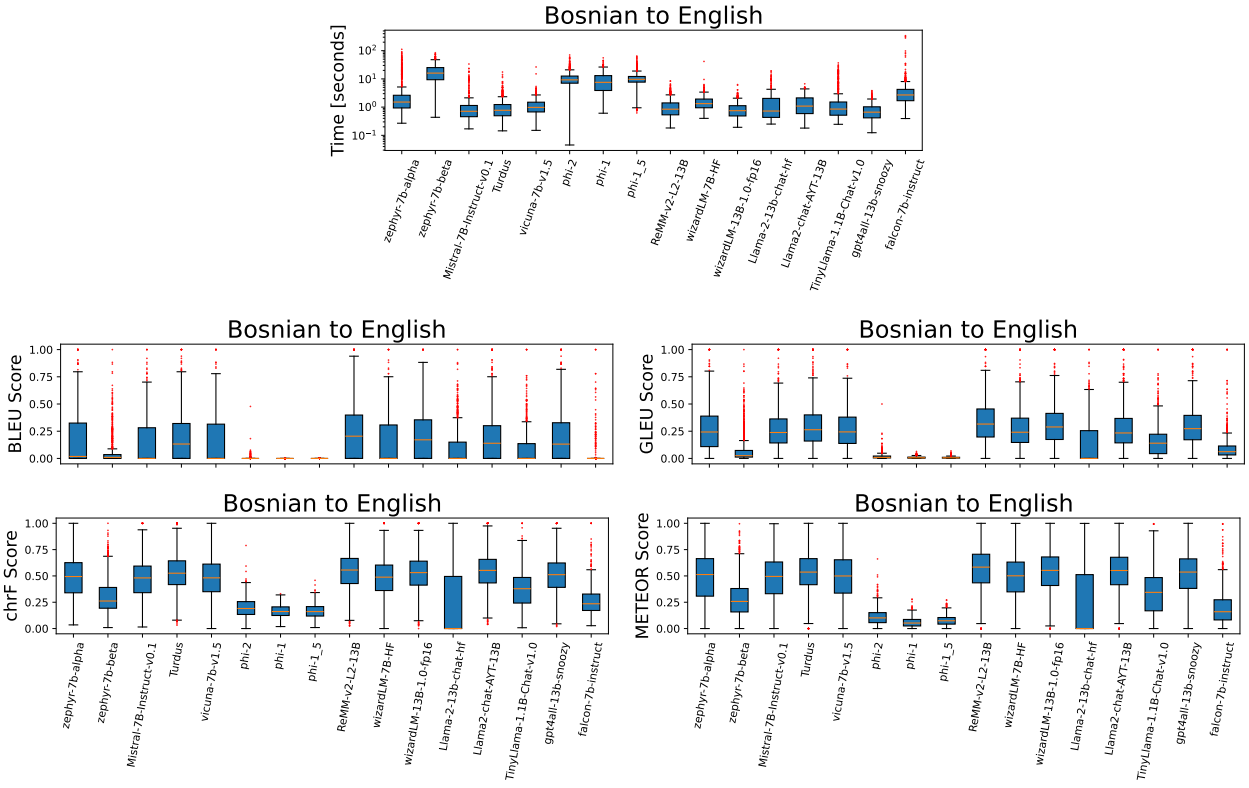

Figure 13: Bosnian-to-English dataset per-sentence translation quality and timing statistics

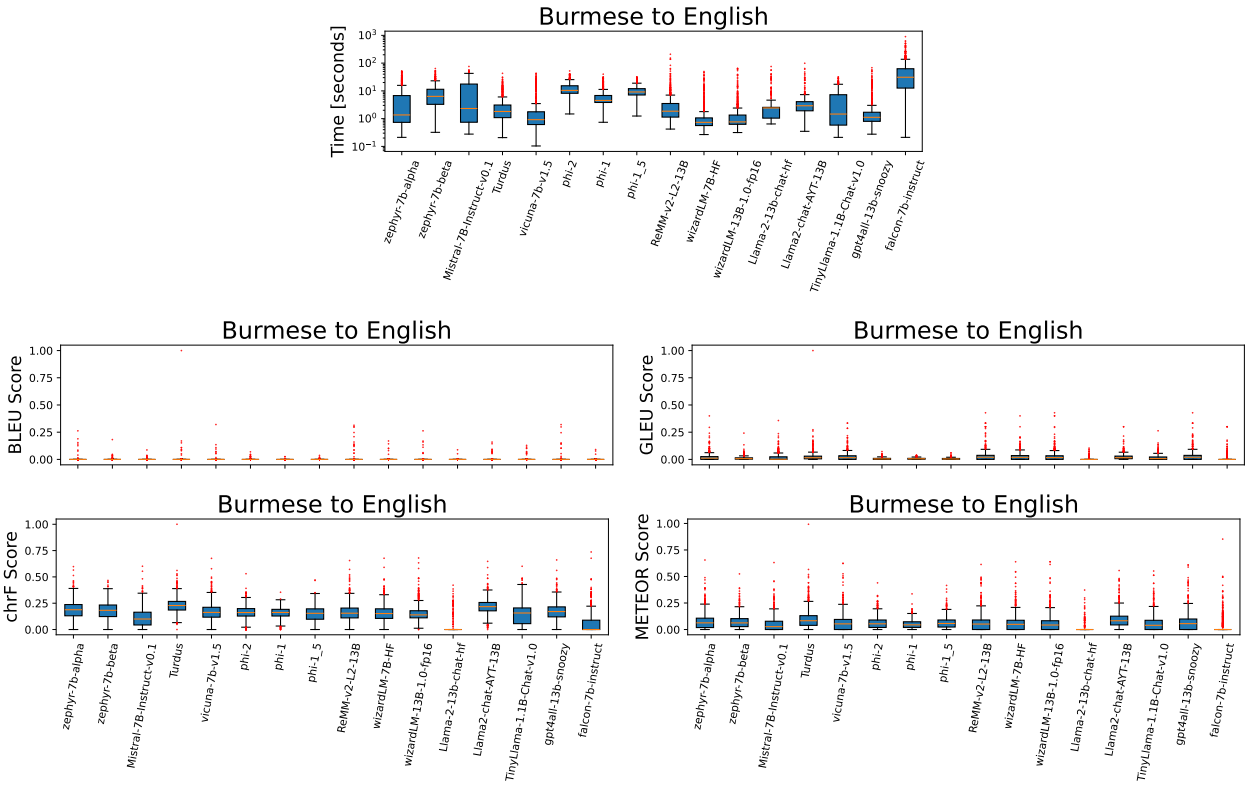

Figure 14: Burmese-to-English dataset per-sentence translation quality and timing statistics

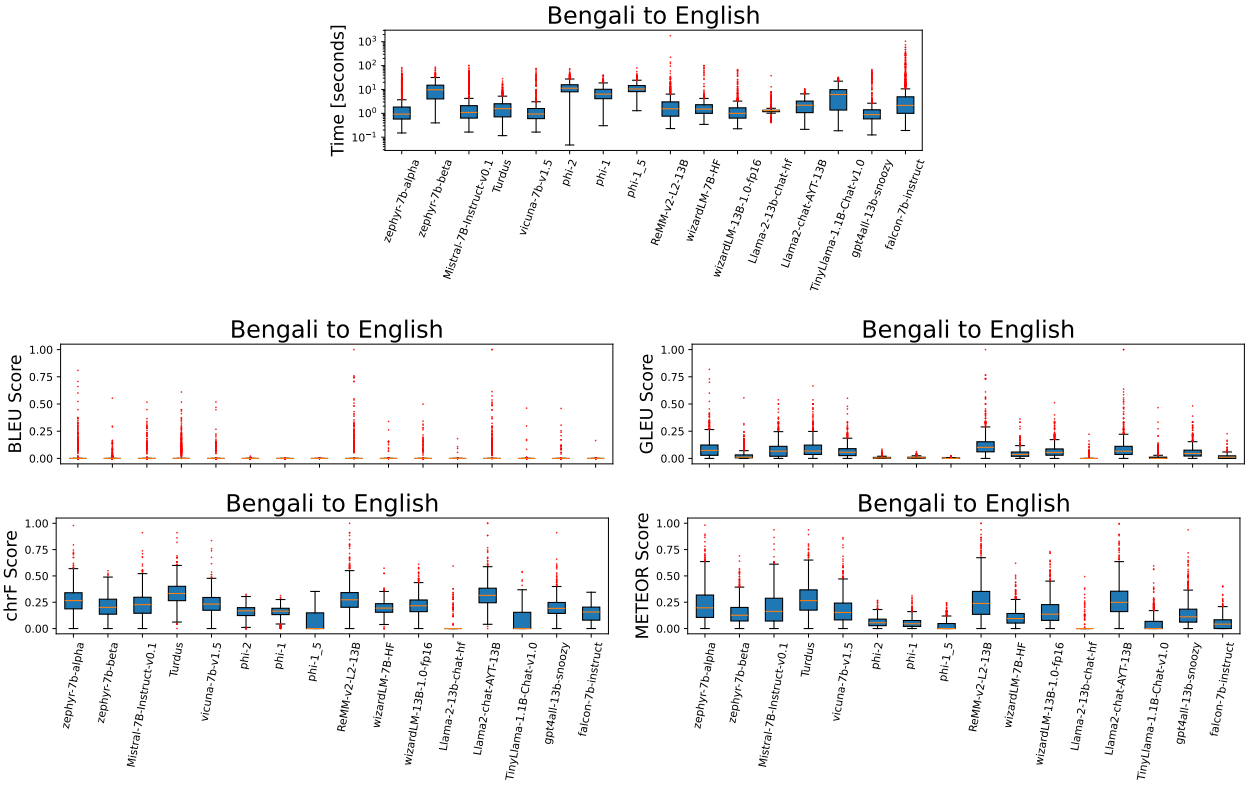

Figure 15: Bengali-to-English dataset per-sentence translation quality and timing statistics

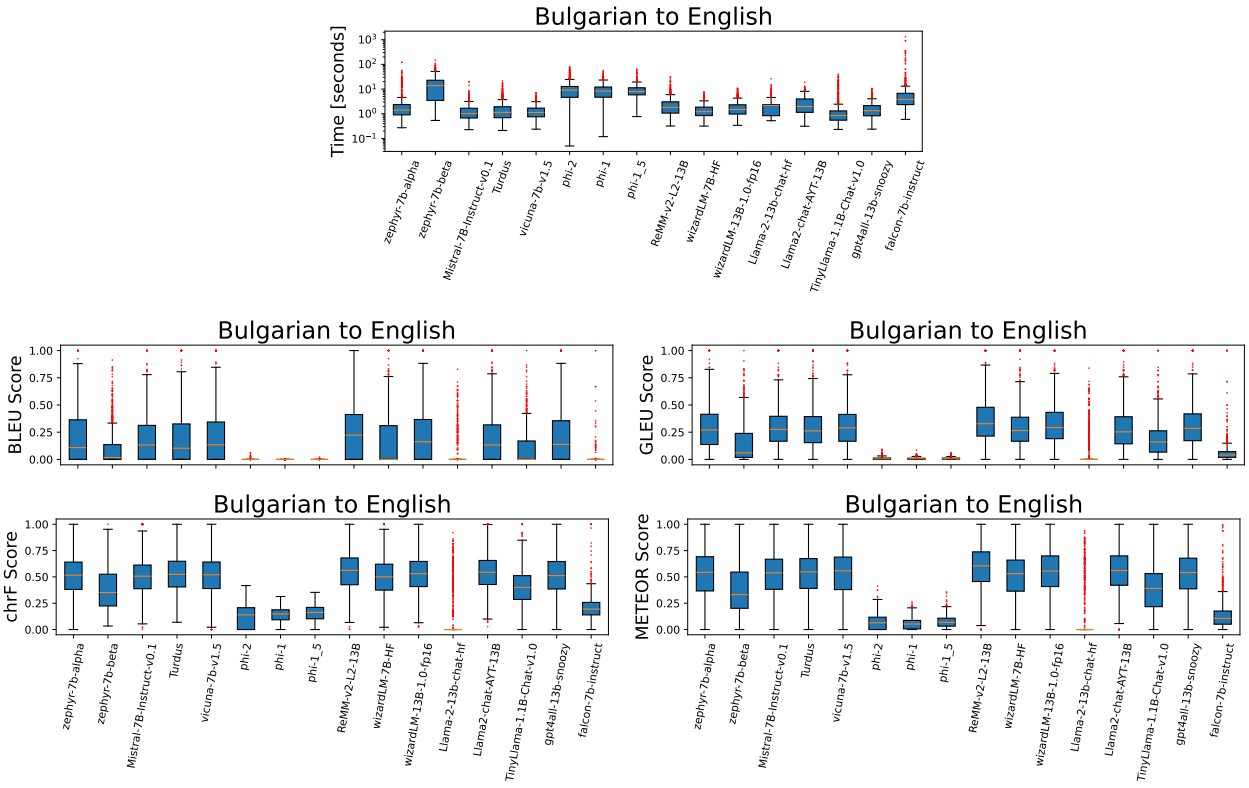

Figure 16: Bulgarian-to-English dataset per-sentence translation quality and timing statistics

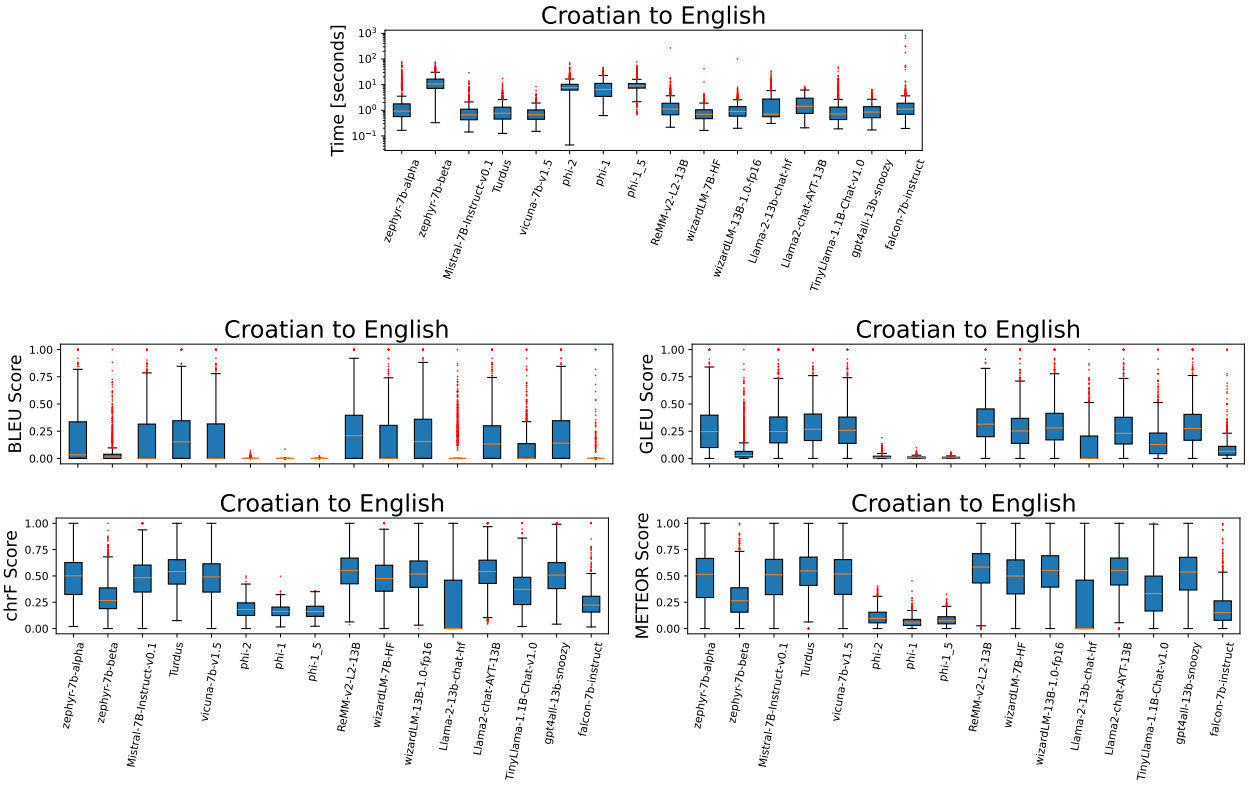

Figure 17: Croatian-to-English dataset per-sentence translation quality and timing statistics

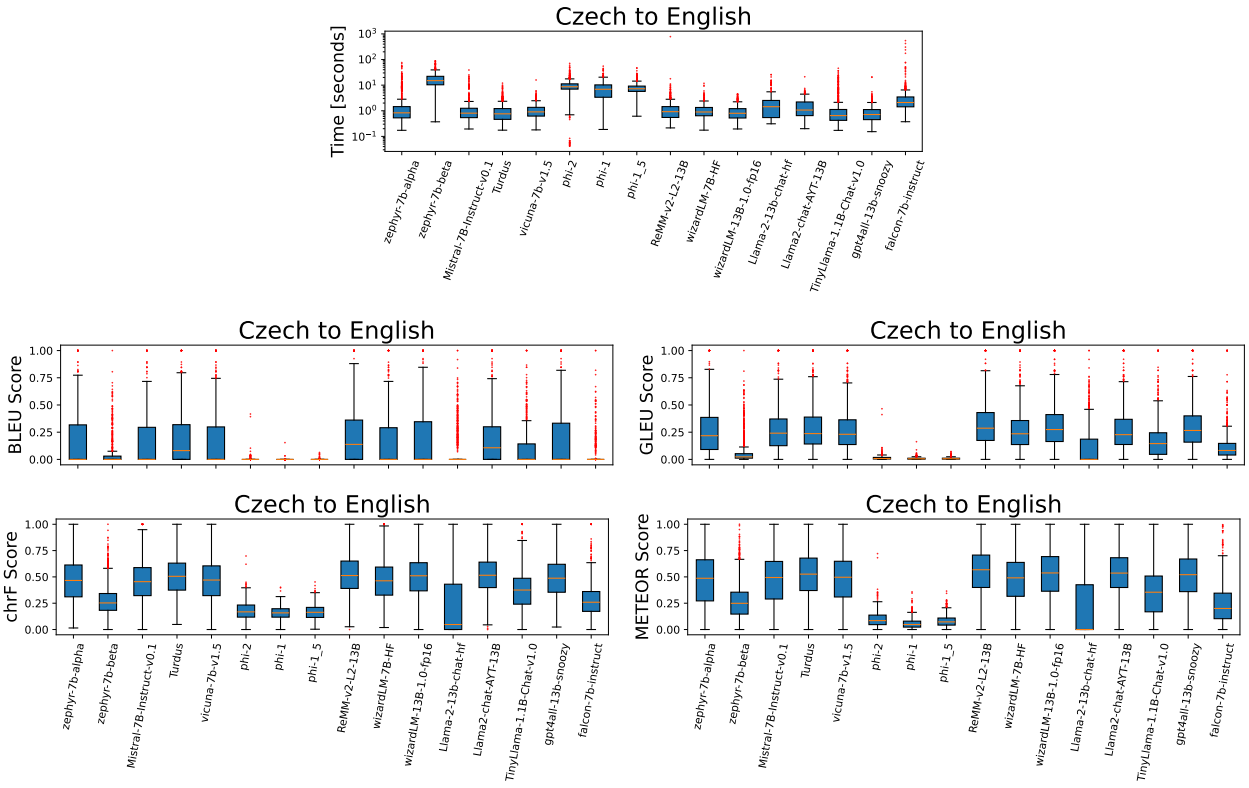

Figure 18: Czech-to-English dataset per-sentence translation quality and timing statistics

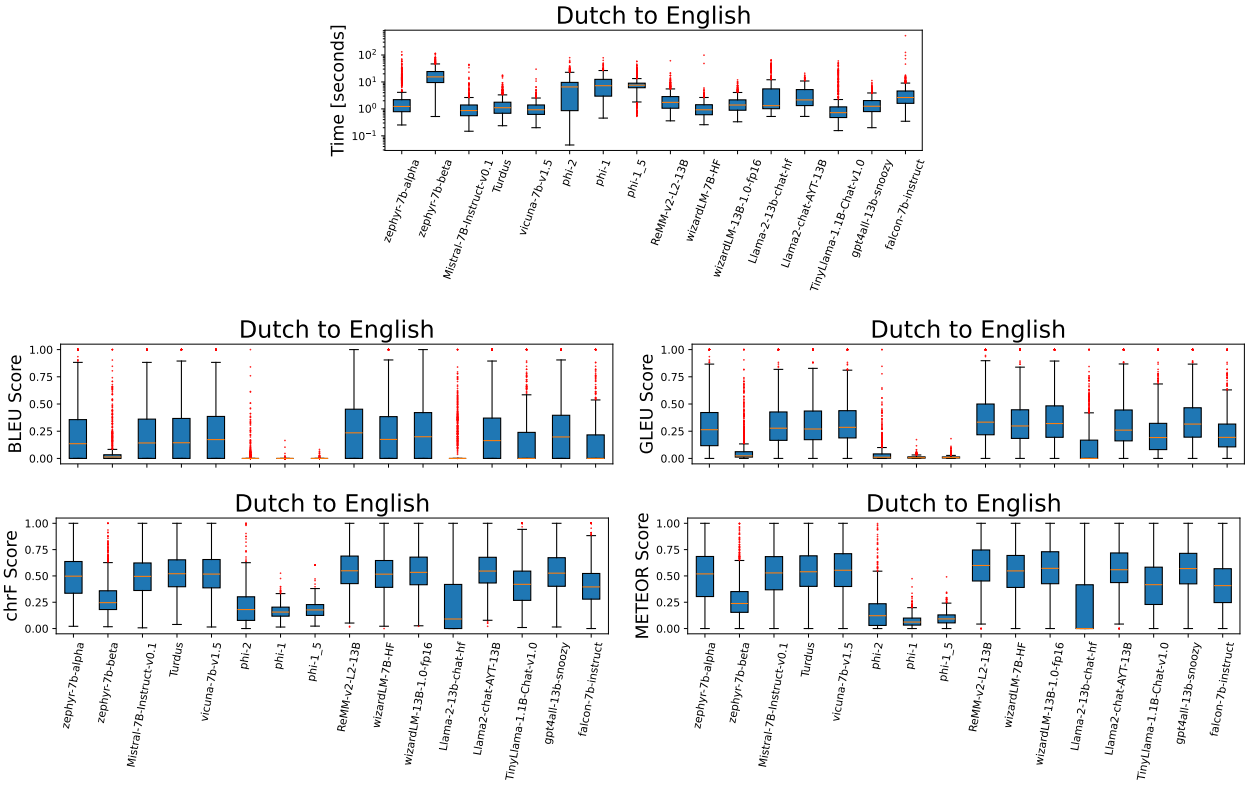

Figure 19: Dutch-to-English dataset per-sentence translation quality and timing statistics

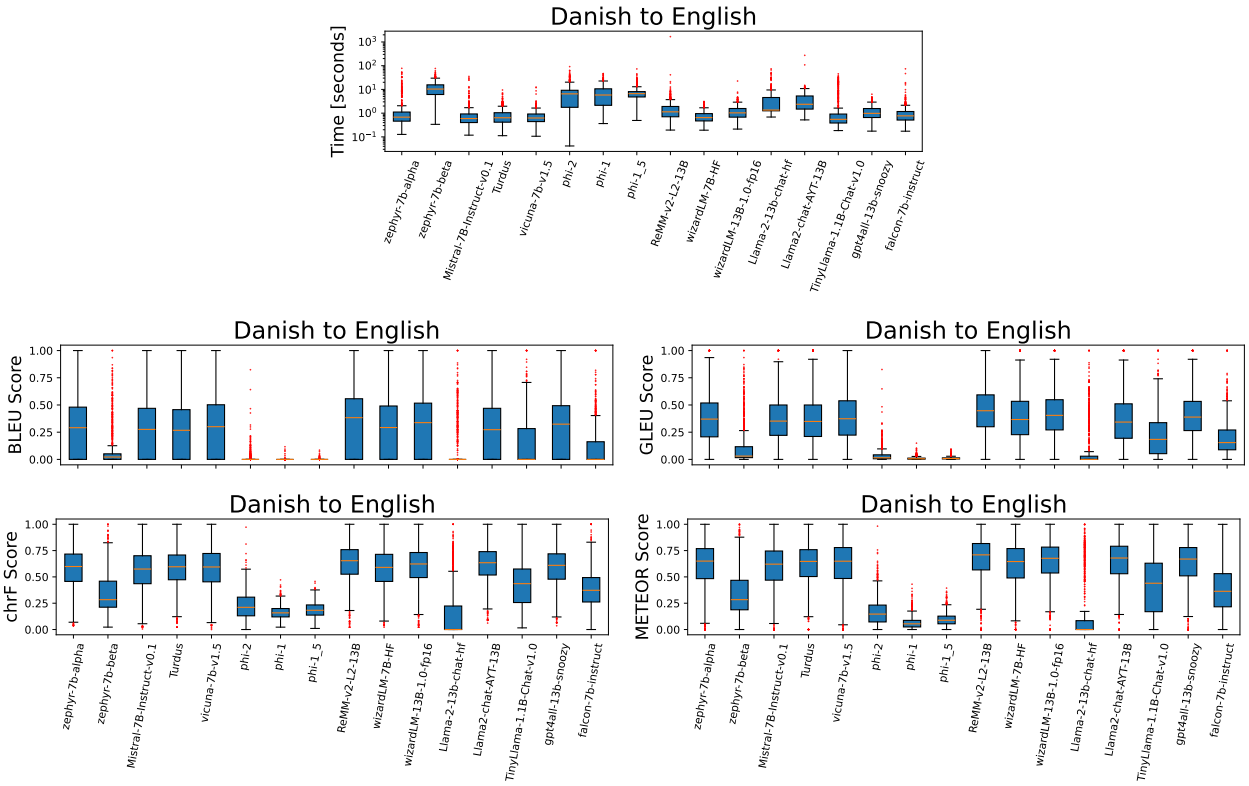

Figure 20: Danish-to-English dataset per-sentence translation quality and timing statistics

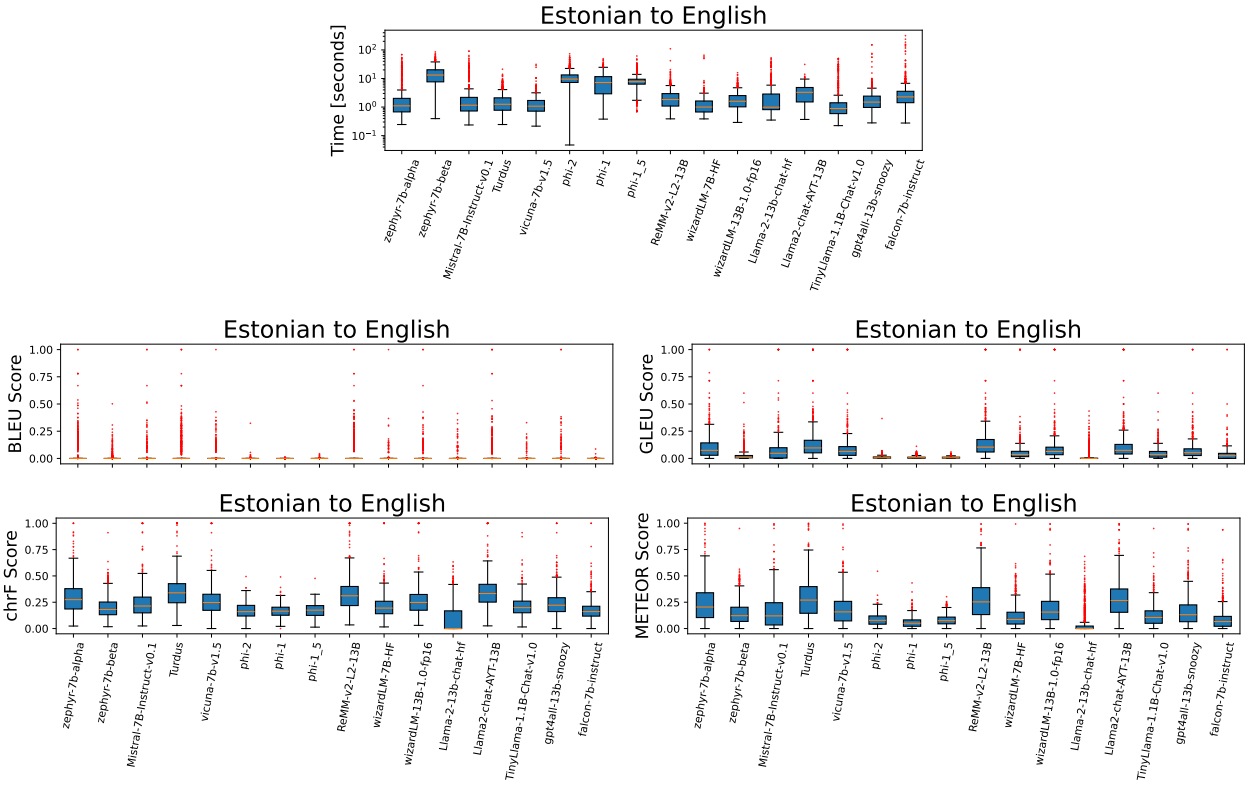

Figure 21: Estonian-to-English dataset per-sentence translation quality and timing statistics

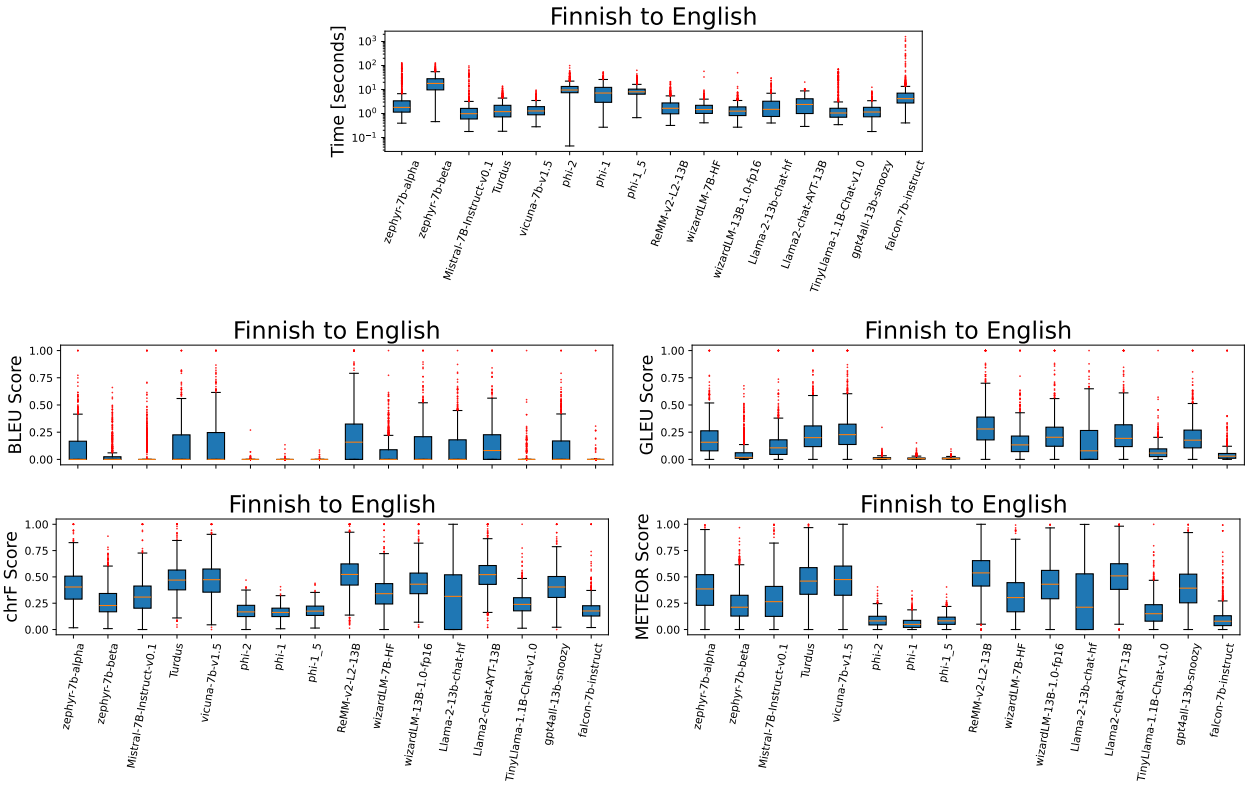

Figure 22: Finnish-to-English dataset per-sentence translation quality and timing statistics

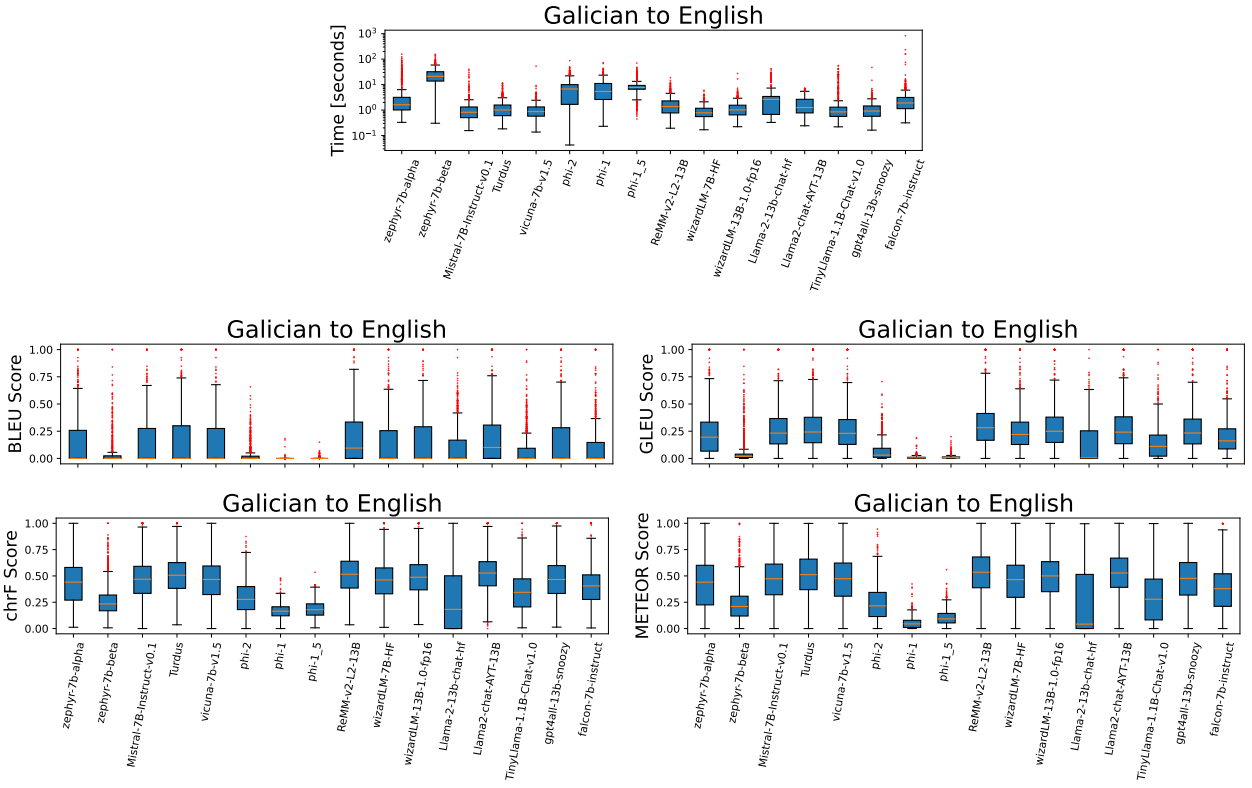

Figure 23: Galician-to-English dataset per-sentence translation quality and timing statistics

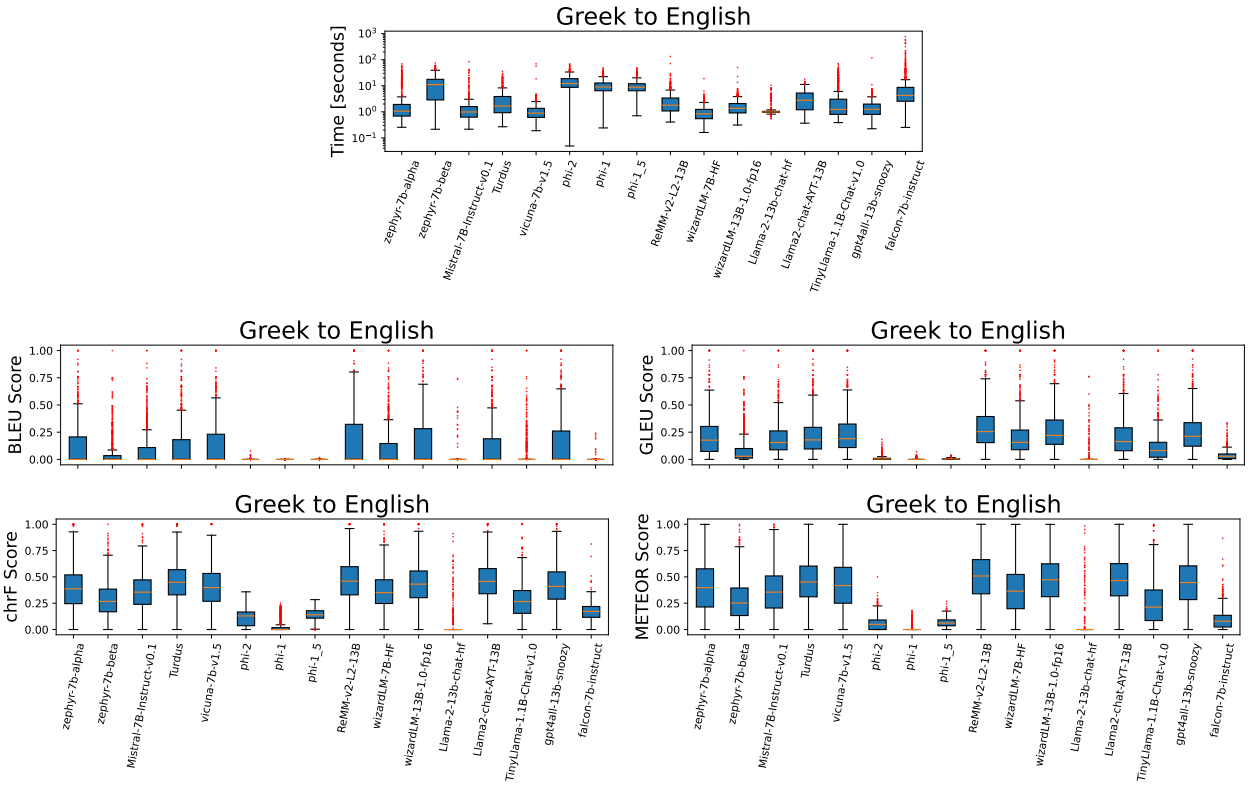

Figure 24: Greek-to-English dataset per-sentence translation quality and timing statistics

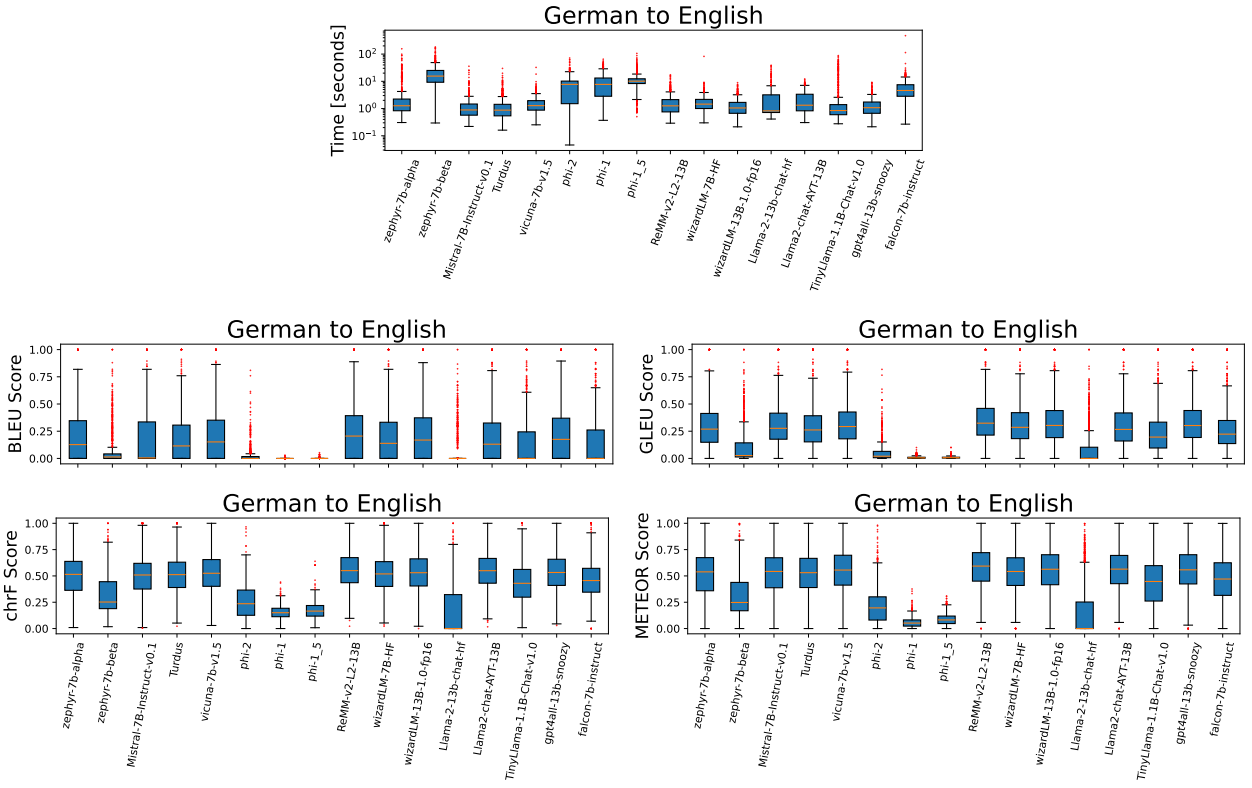

Figure 25: German-to-English dataset per-sentence translation quality and timing statistics

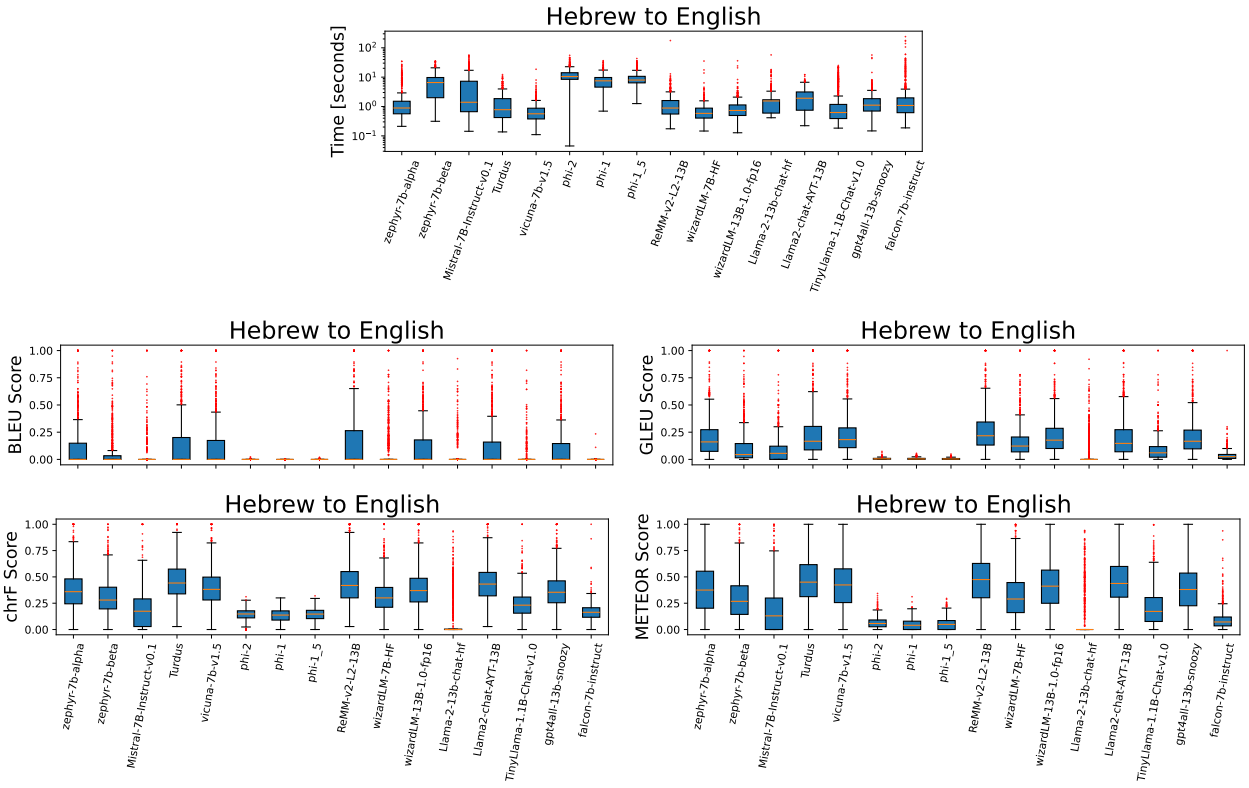

Figure 26: Hebrew-to-English dataset per-sentence translation quality and timing statistics

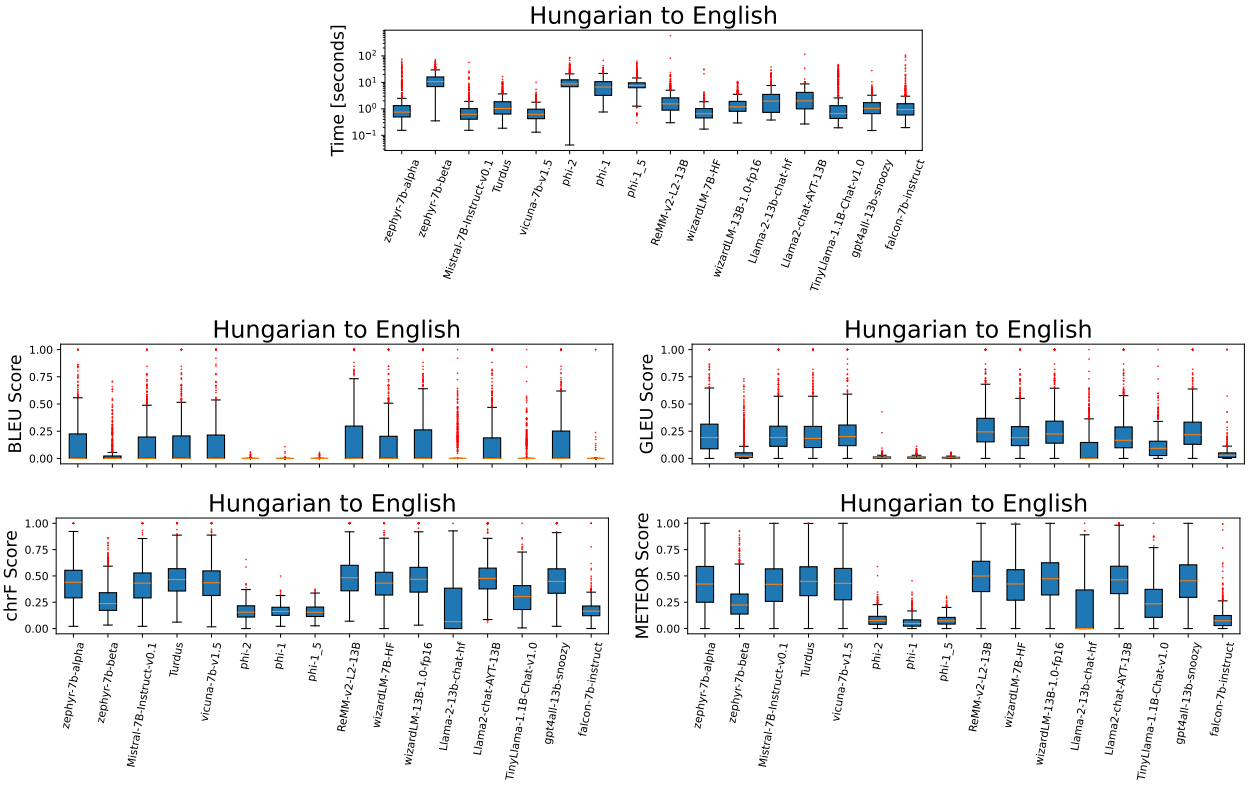

Figure 27: Hungarian-to-English dataset per-sentence translation quality and timing statistics

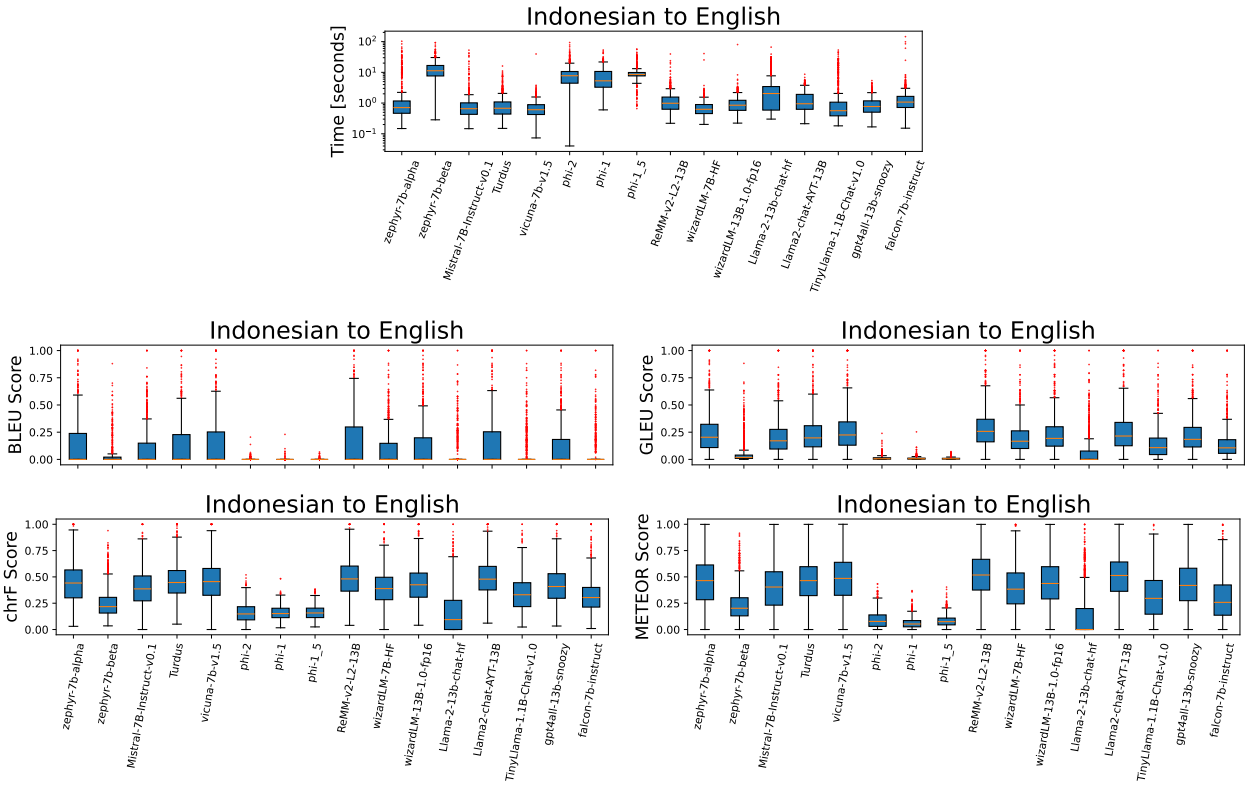

Figure 28: Indonesian-to-English dataset per-sentence translation quality and timing statistics

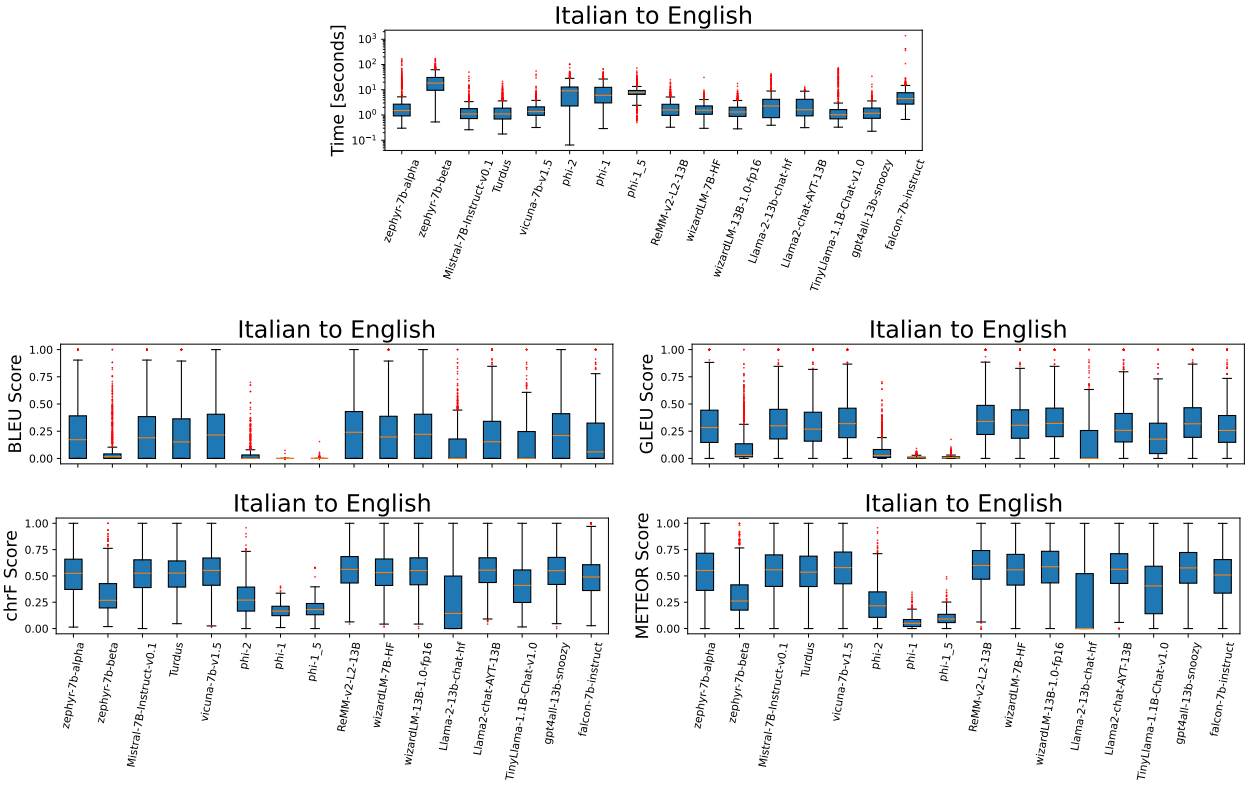

Figure 29: Italian-to-English dataset per-sentence translation quality and timing statistics

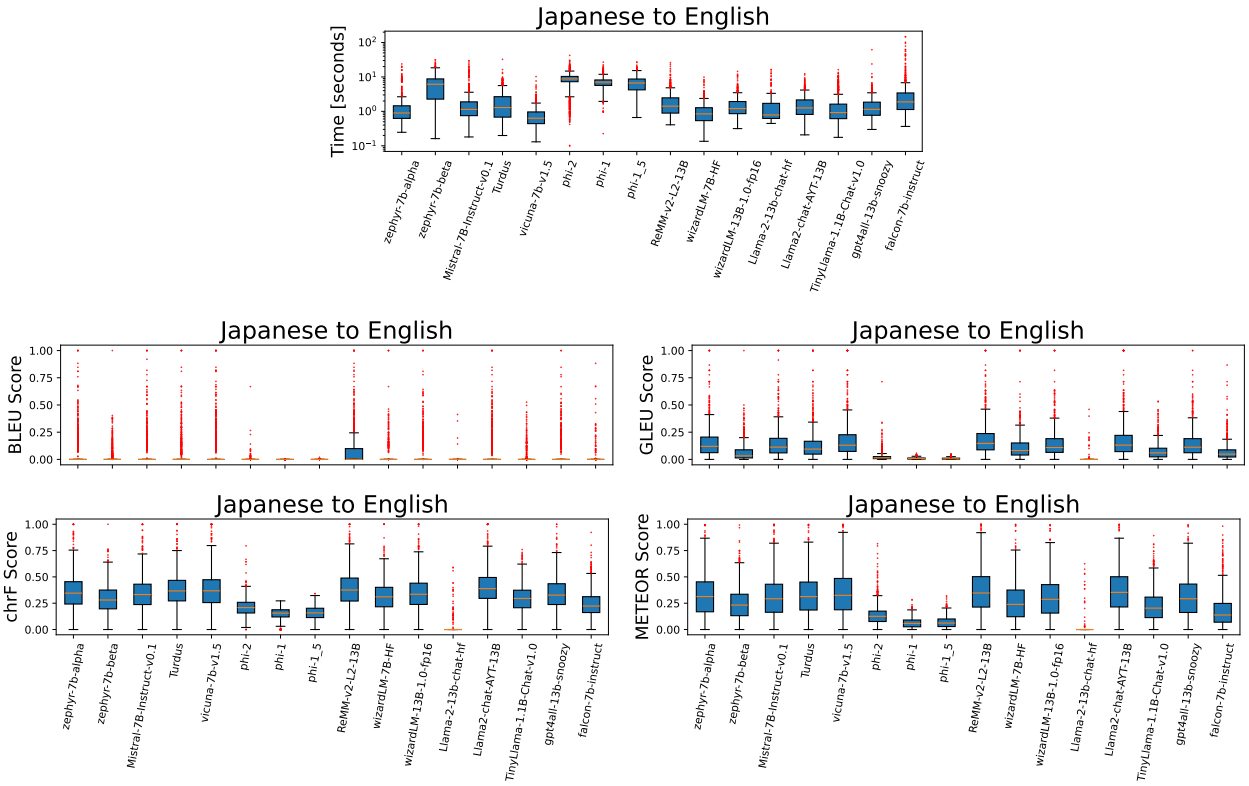

Figure 30: Japanese-to-English dataset per-sentence translation quality and timing statistics

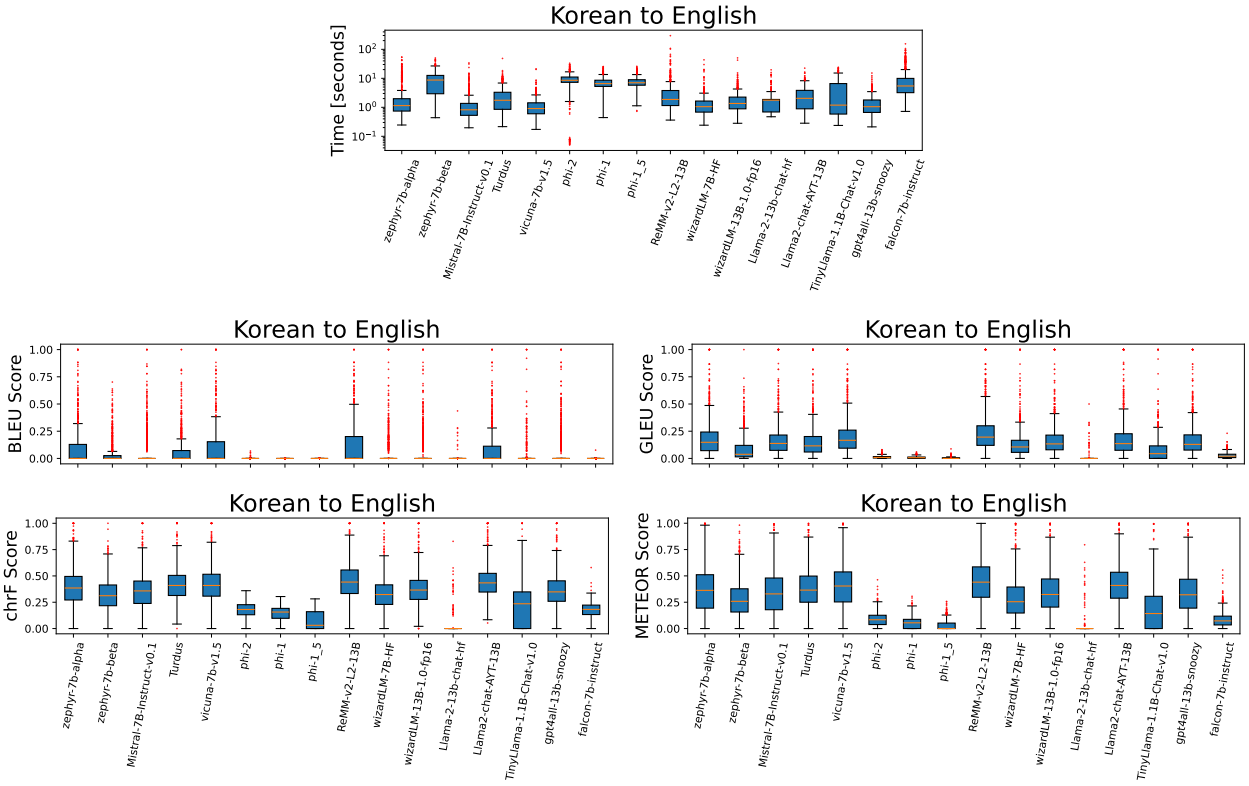

Figure 31: Korean-to-English dataset per-sentence translation quality and timing statistics

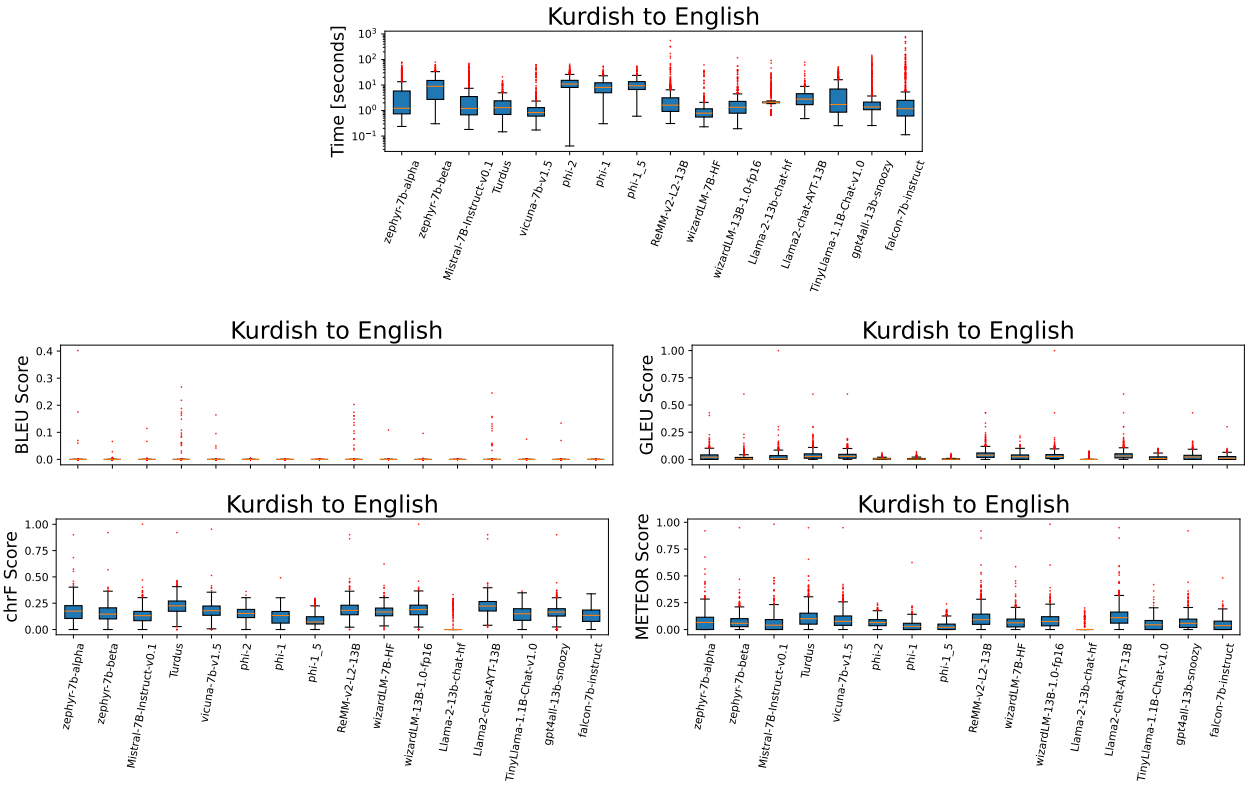

Figure 32: Kurdish-to-English dataset per-sentence translation quality and timing statistics

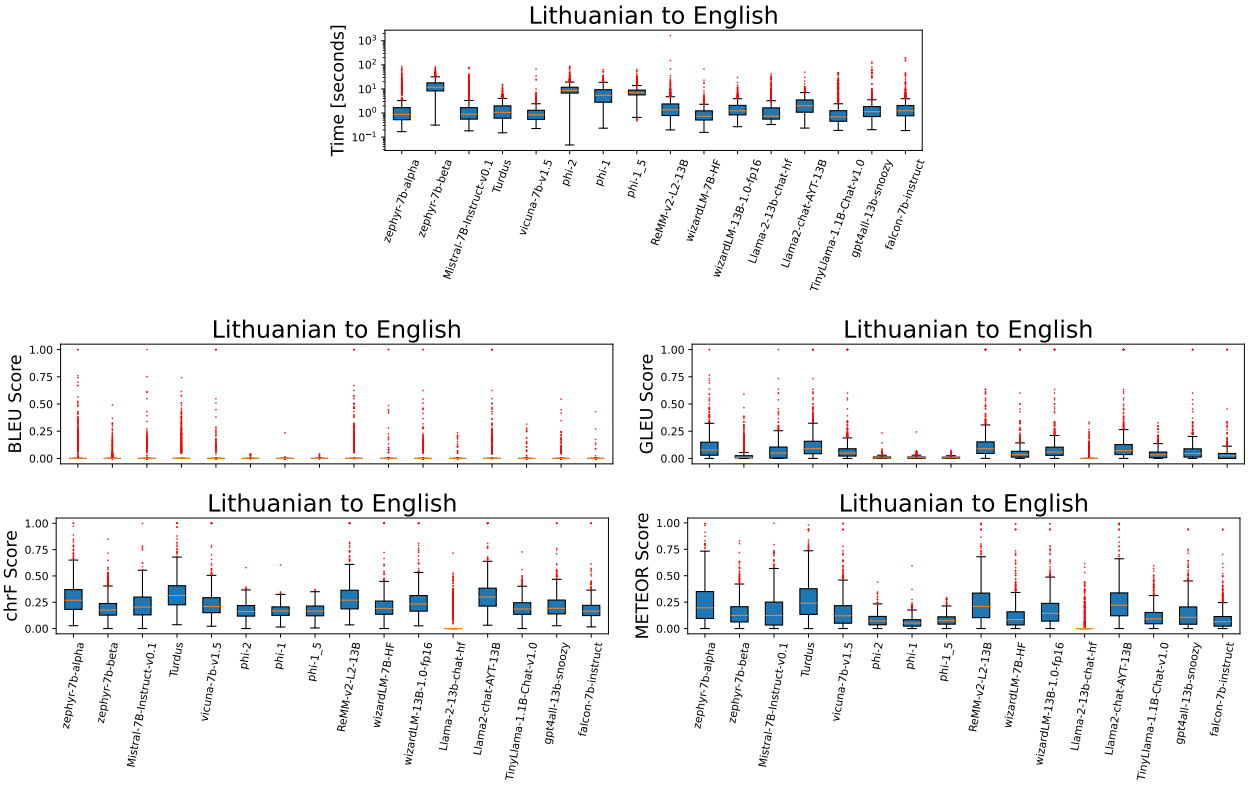

Figure 33: Lithuanian-to-English dataset per-sentence translation quality and timing statistics

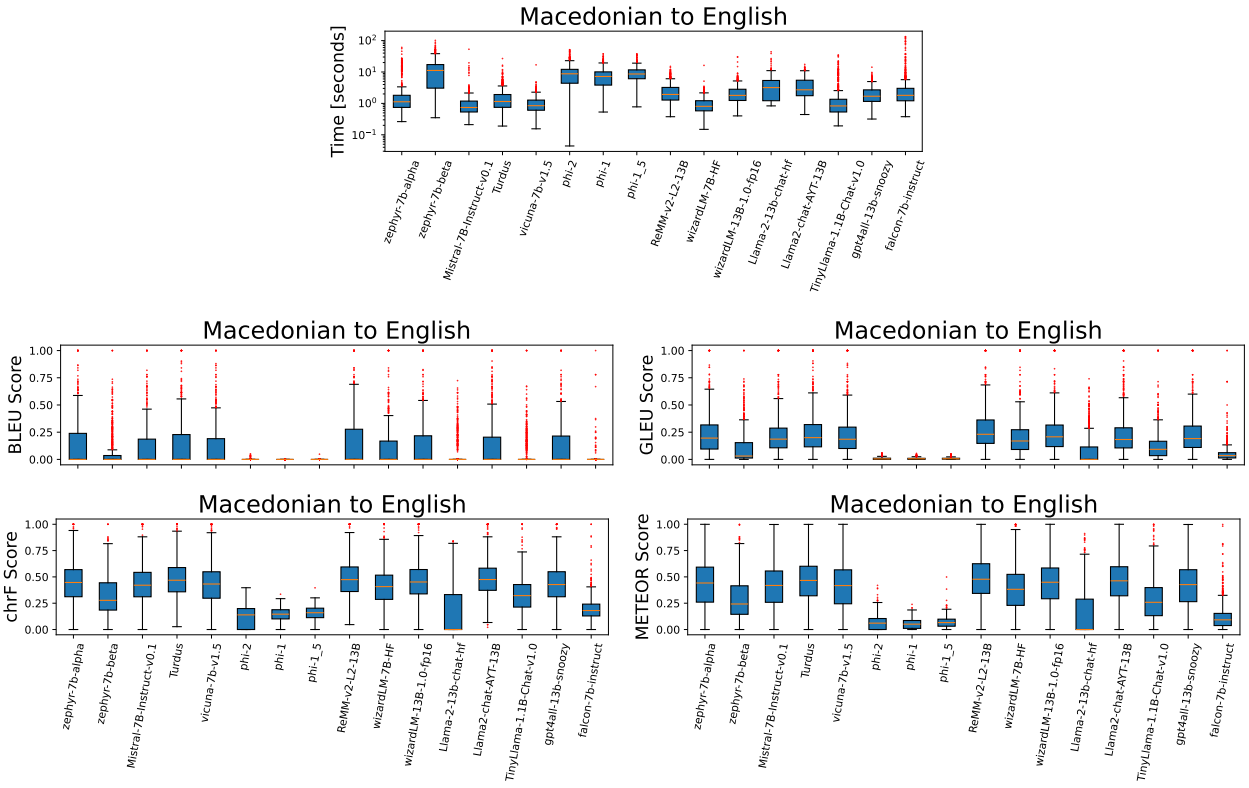

Figure 34: Macedonian-to-English dataset per-sentence translation quality and timing statistics

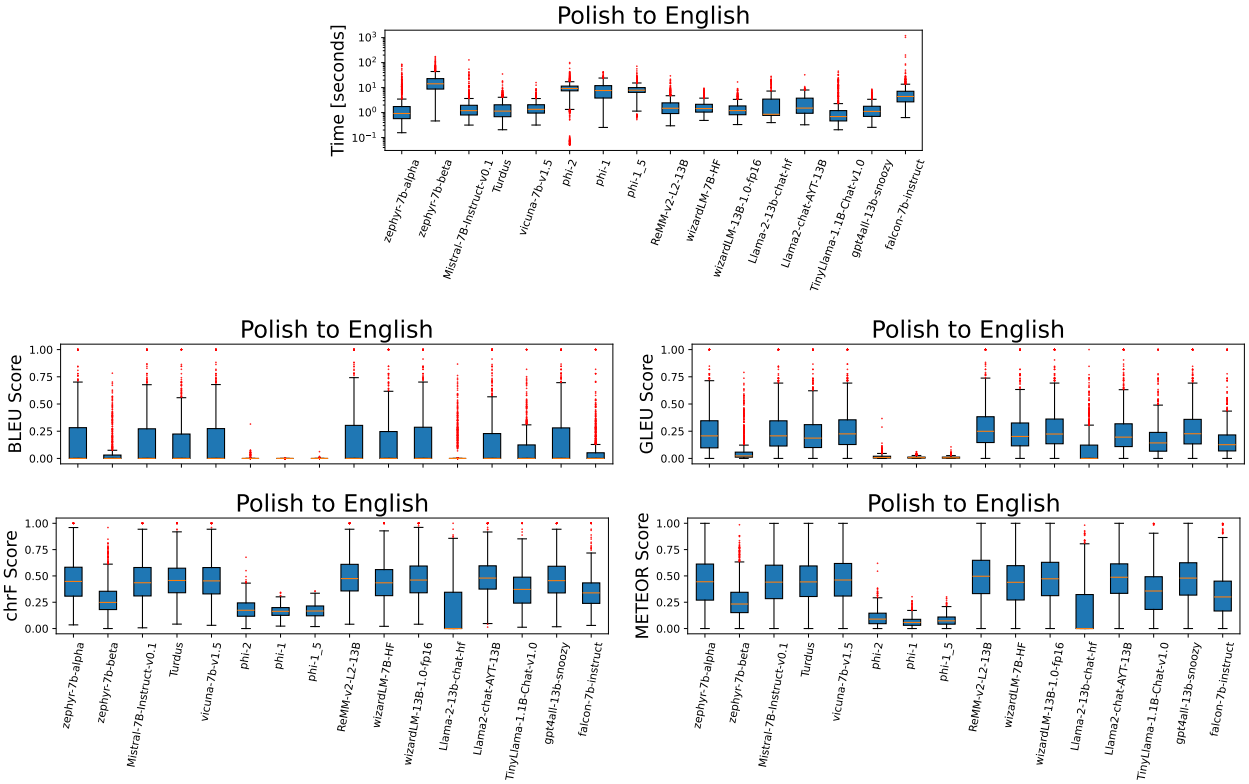

Figure 35: Polish-to-English dataset per-sentence translation quality and timing statistics

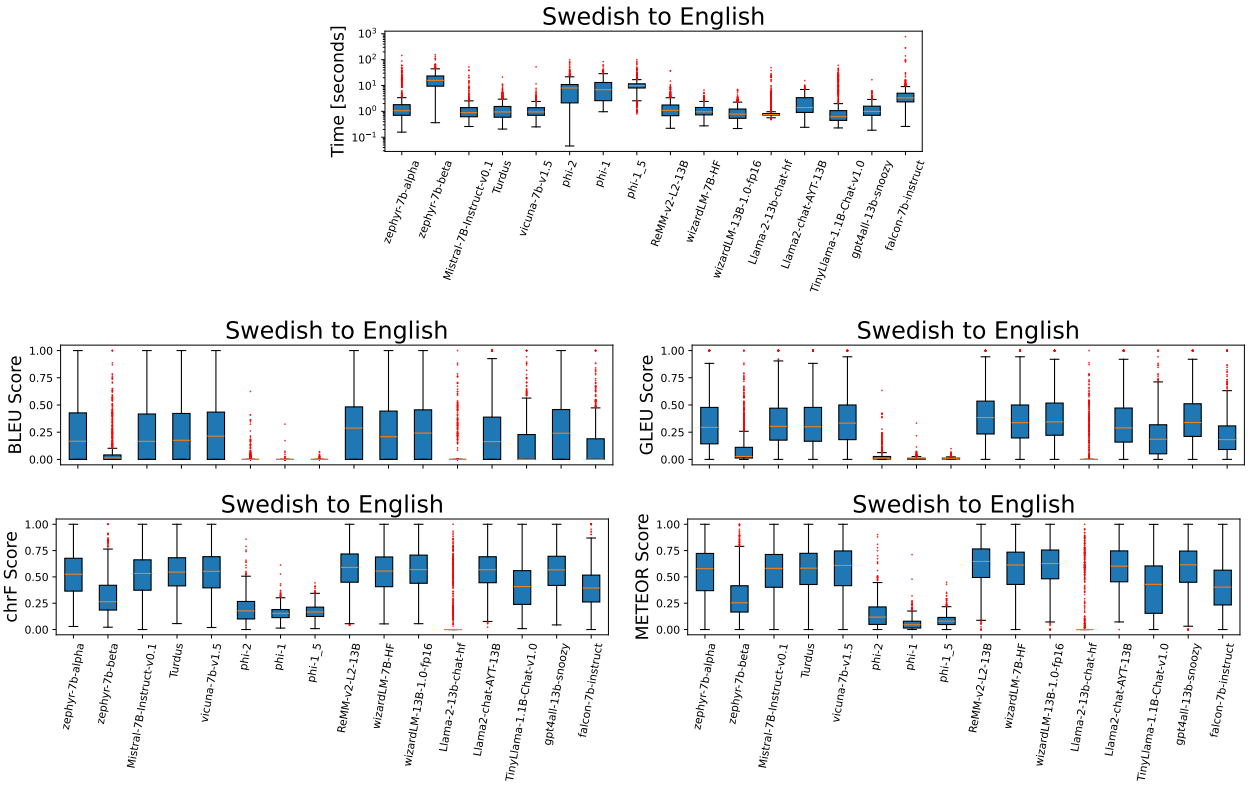

Figure 36: Swedish-to-English dataset per-sentence translation quality and timing statistics

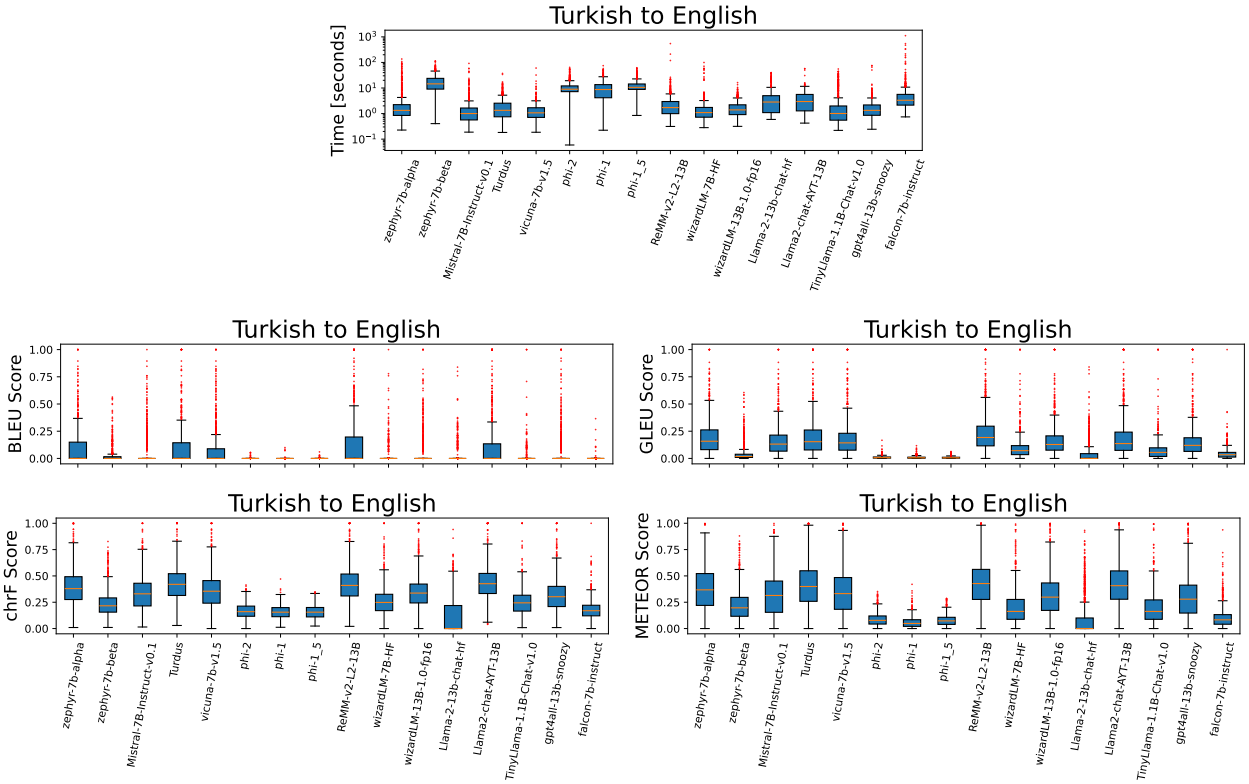

Figure 37: Turkish-to-English dataset per-sentence translation quality and timing statistics

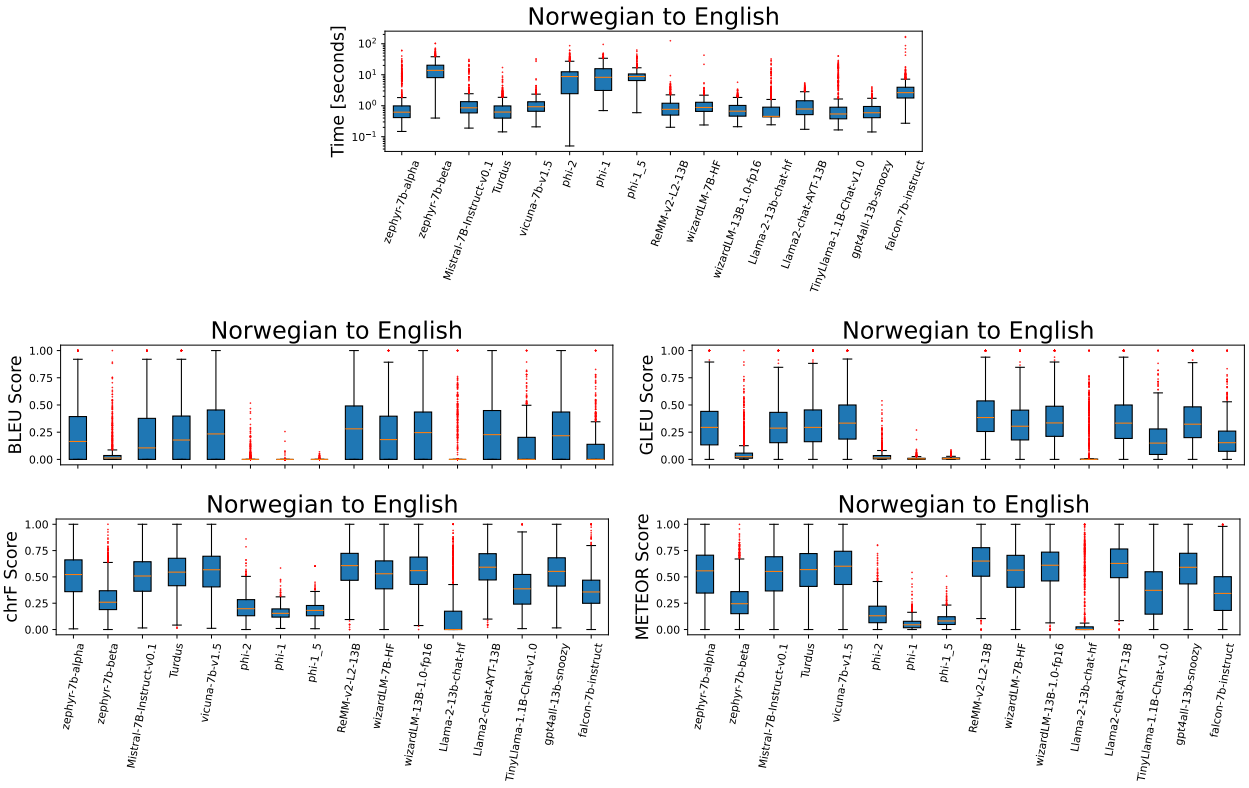

Figure 38: Norwegian-to-English dataset per-sentence translation quality and timing statistics

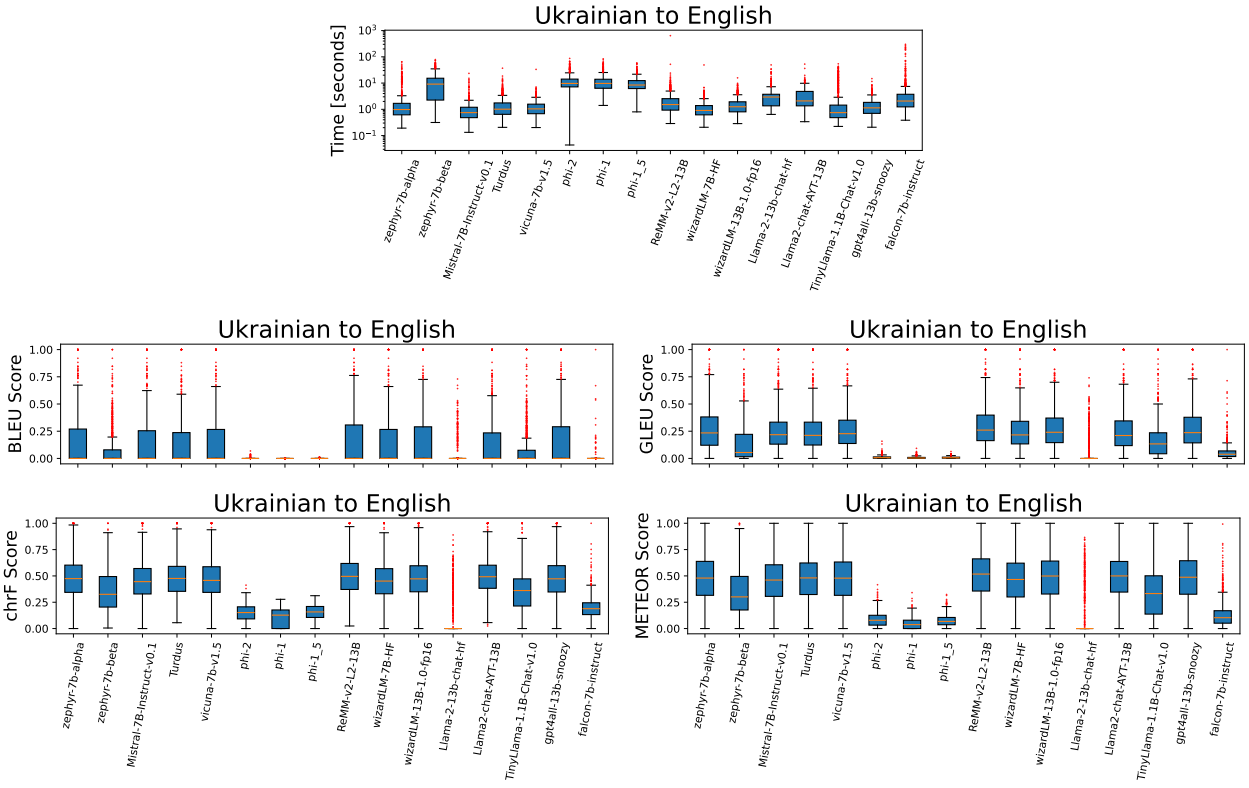

Figure 39: Ukranian-to-English dataset per-sentence translation quality and timing statistics

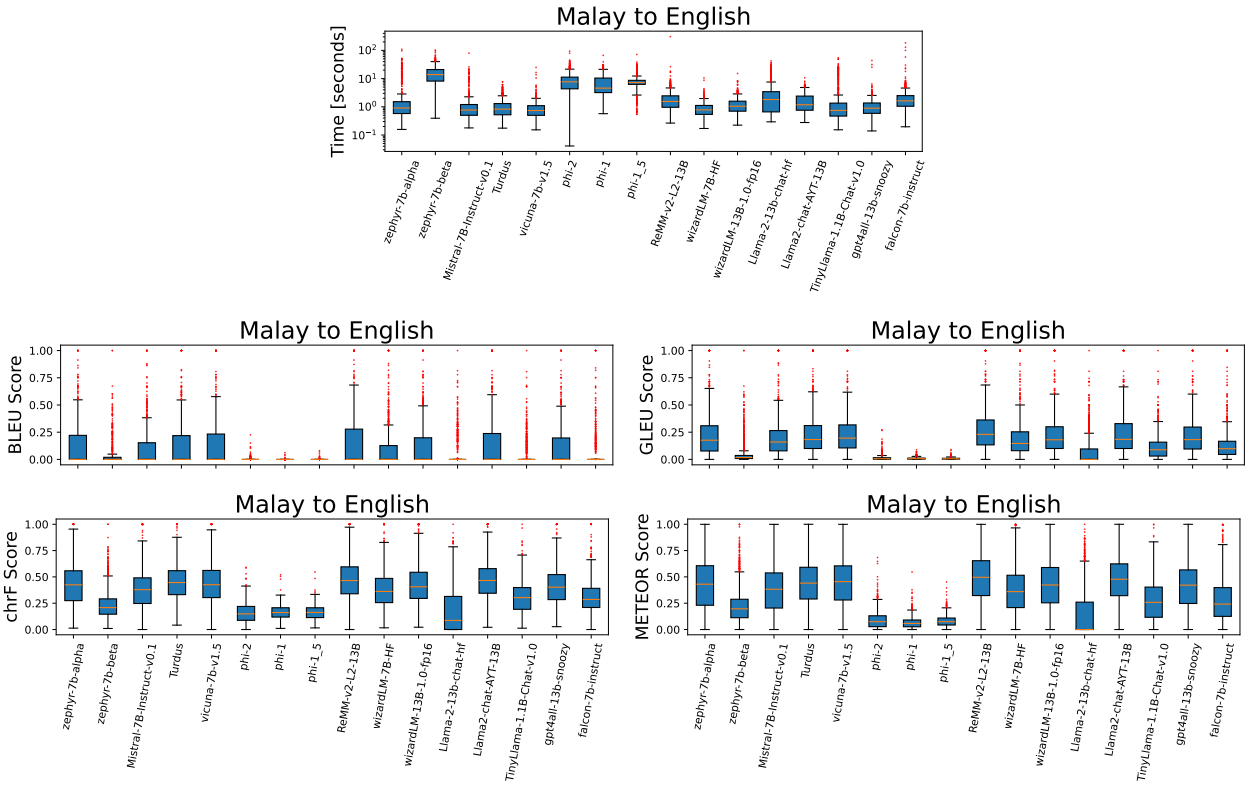

Figure 40: Malay-to-English dataset per-sentence translation quality and timing statistics

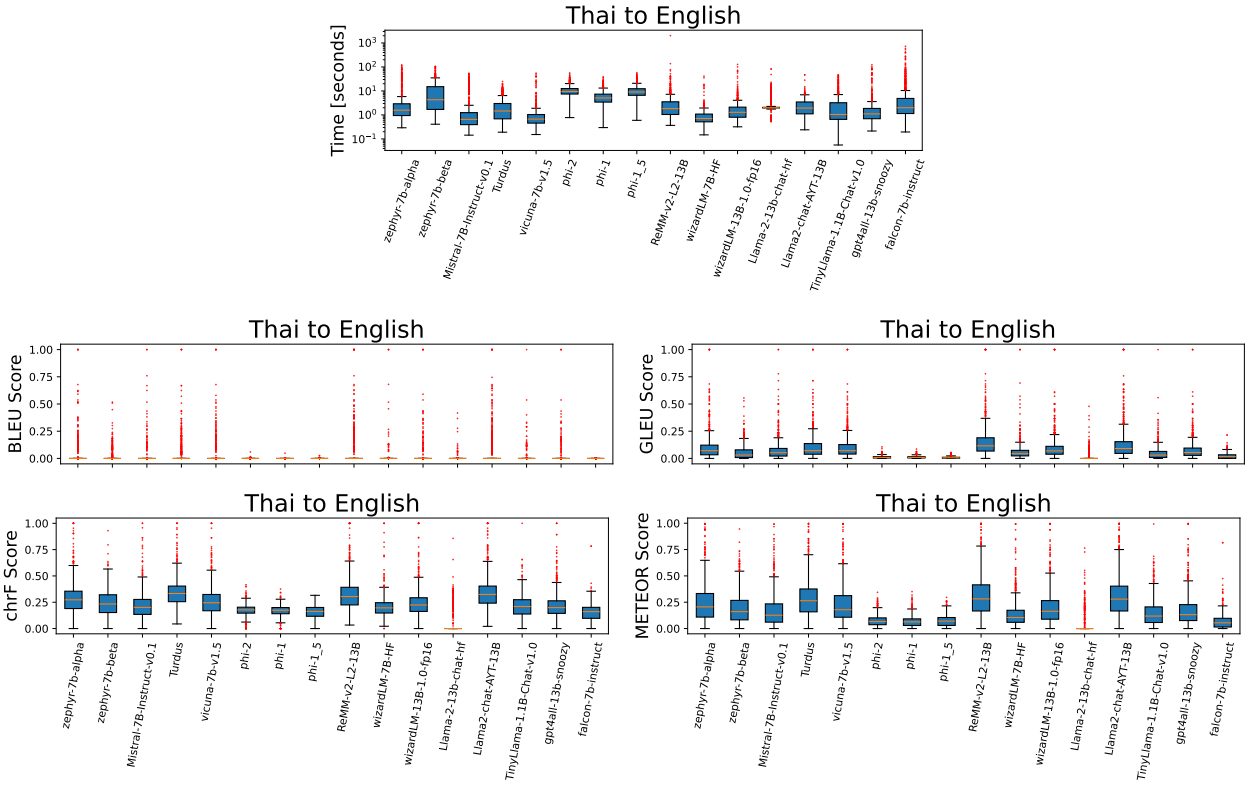

Figure 41: Thai-to-English dataset per-sentence translation quality and timing statistics

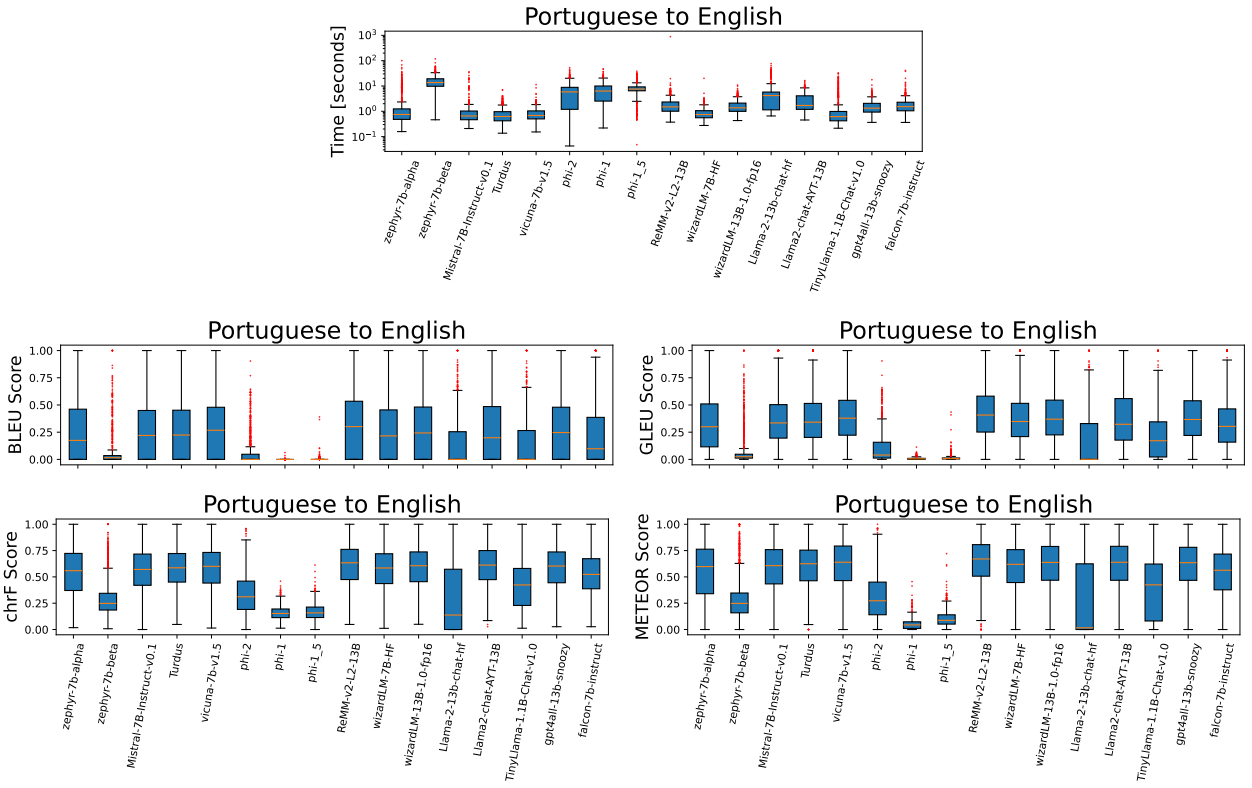

Figure 42: Portuguese-to-English dataset per-sentence translation quality and timing statistics

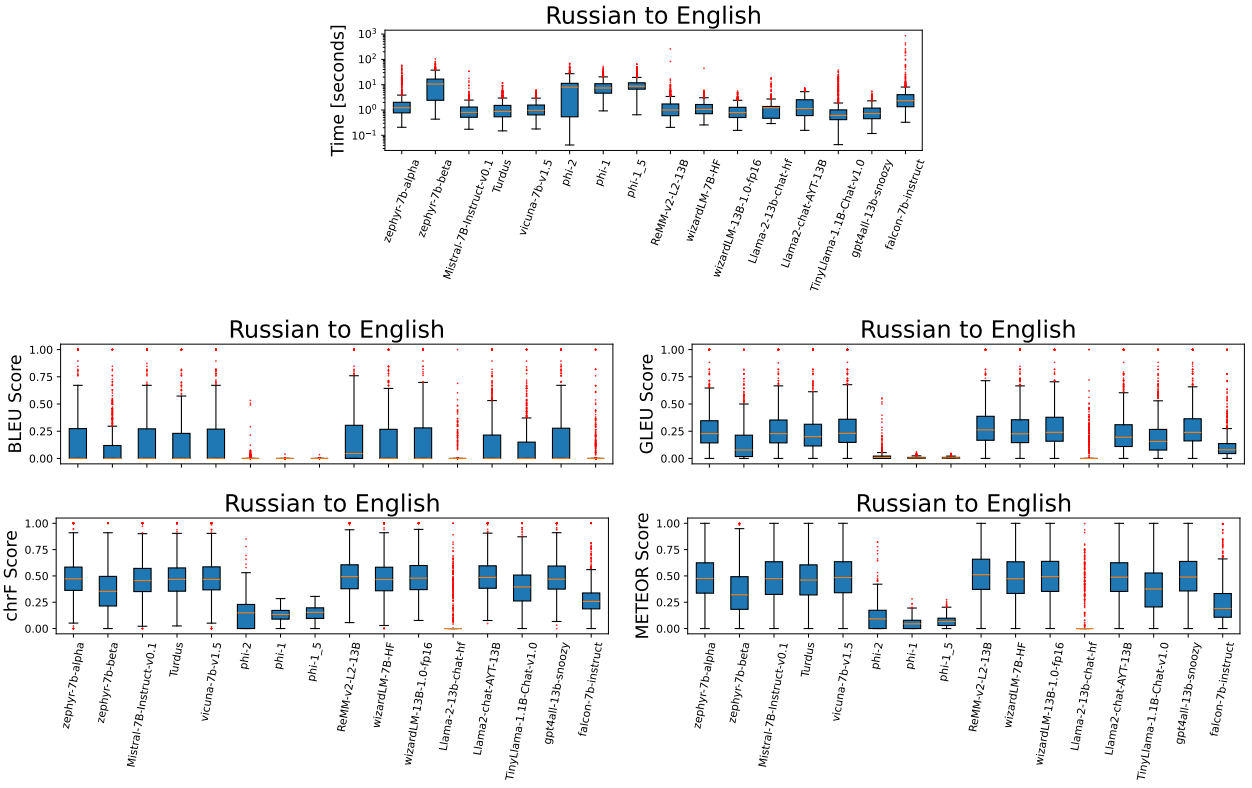

Figure 43: Russian-to-English dataset per-sentence translation quality and timing statistics

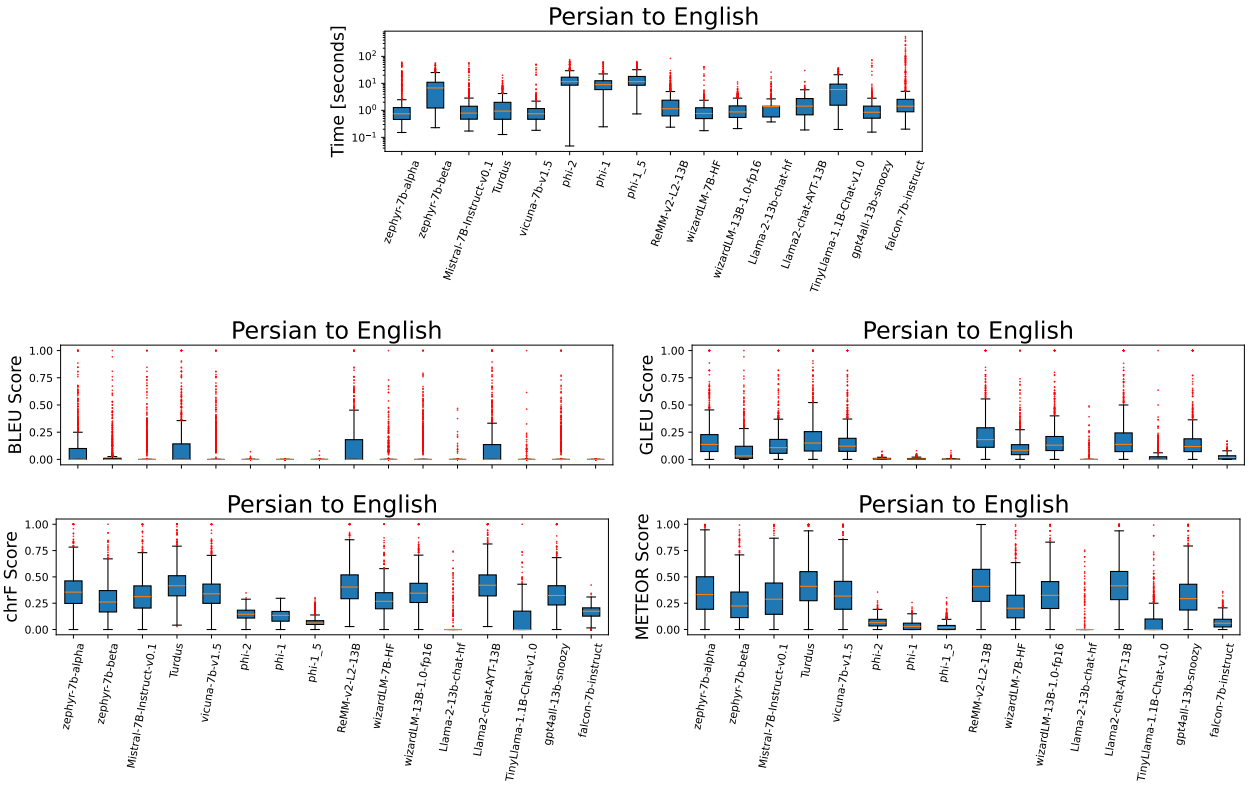

Figure 44: Persian-to-English dataset per-sentence translation quality and timing statistics

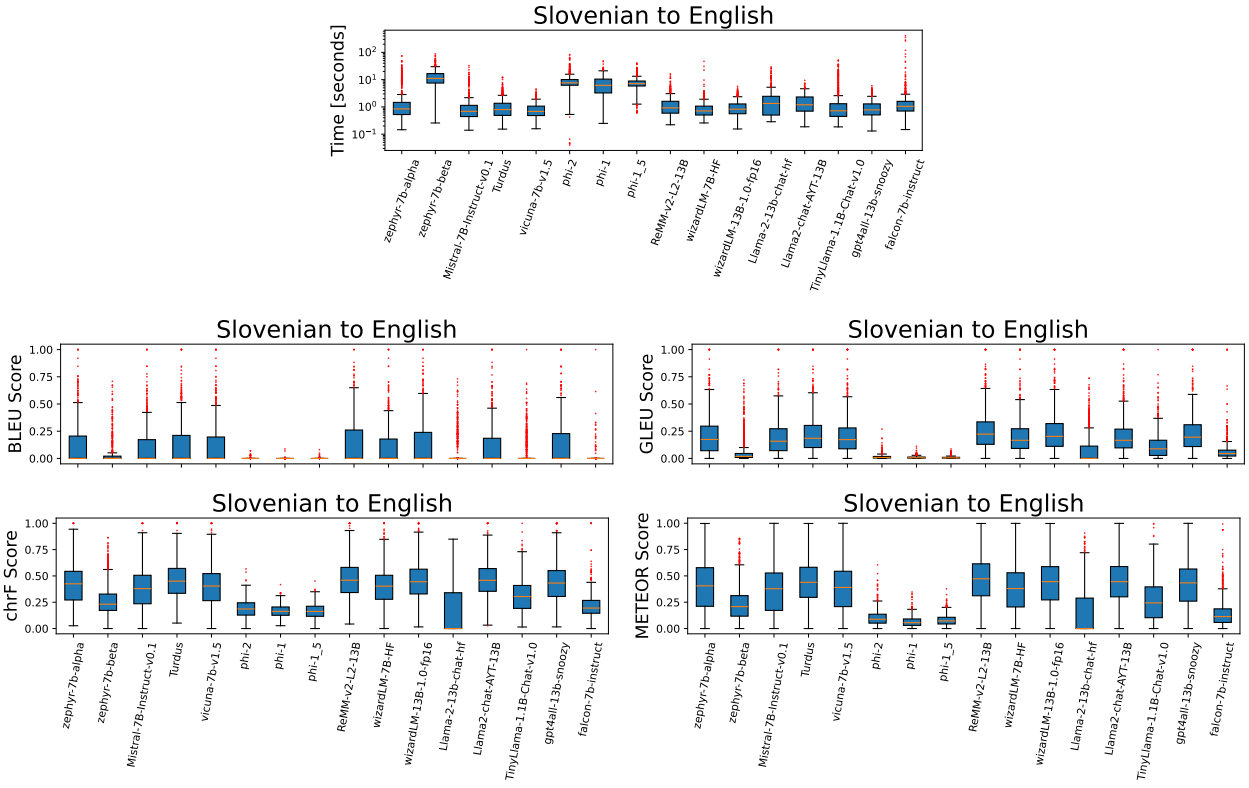

Figure 45: Slovenian-to-English dataset per-sentence translation quality and timing statistics

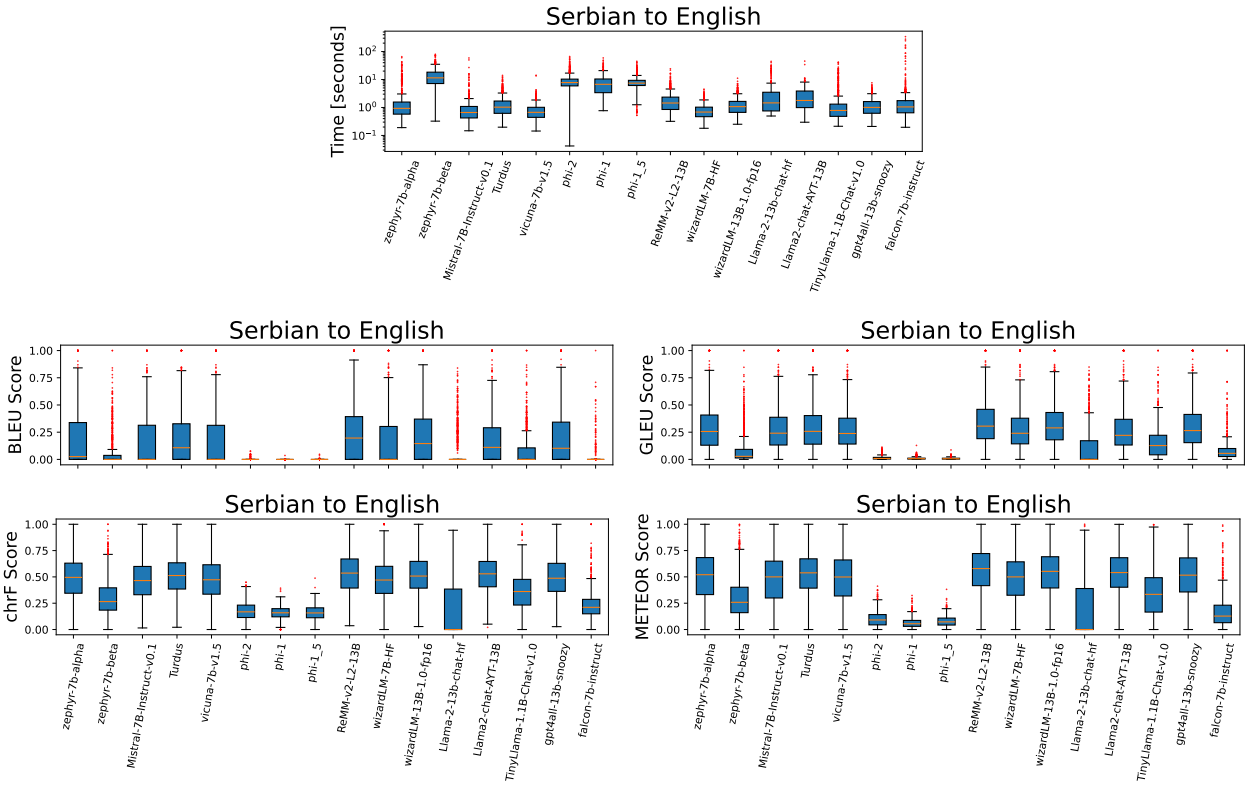

Figure 46: Serbian-to-English dataset per-sentence translation quality and timing statistics

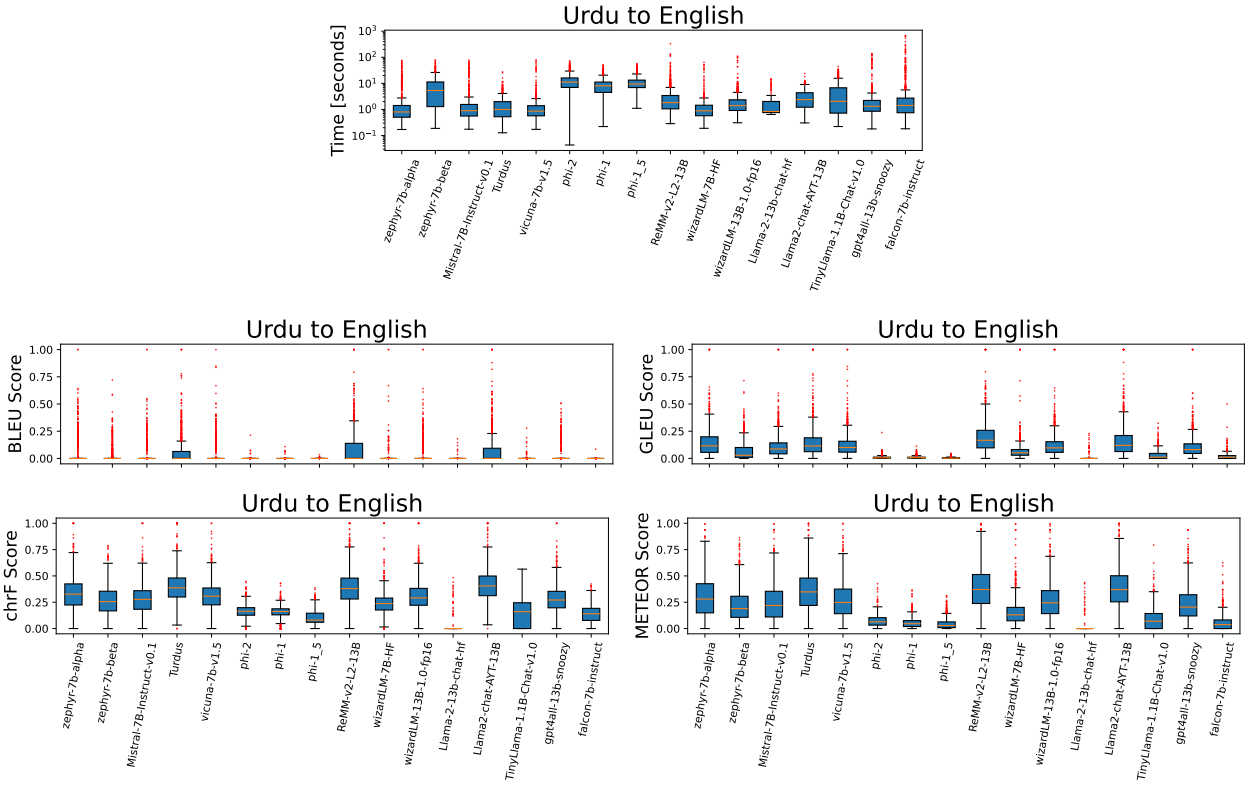

Figure 47: Urdu-to-English dataset per-sentence translation quality and timing statistics

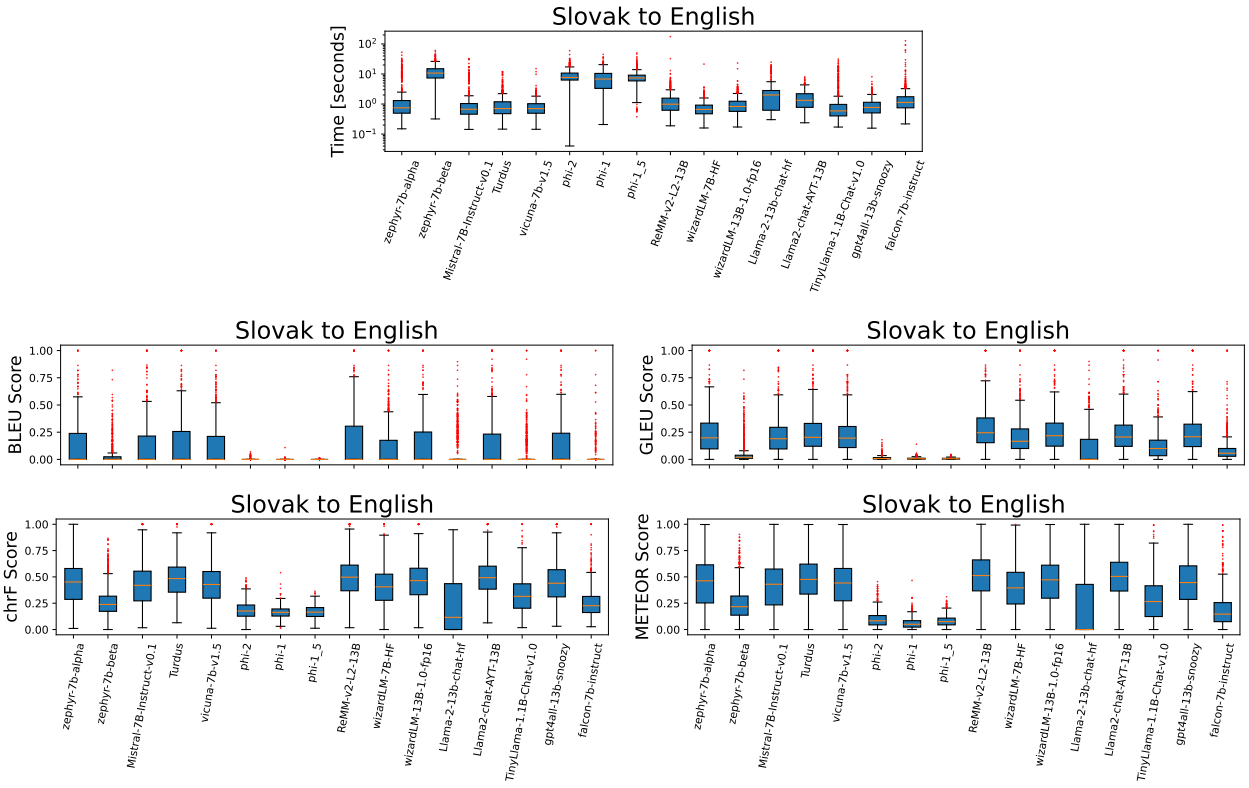

Figure 48: Slovak-to-English dataset per-sentence translation quality and timing statistics

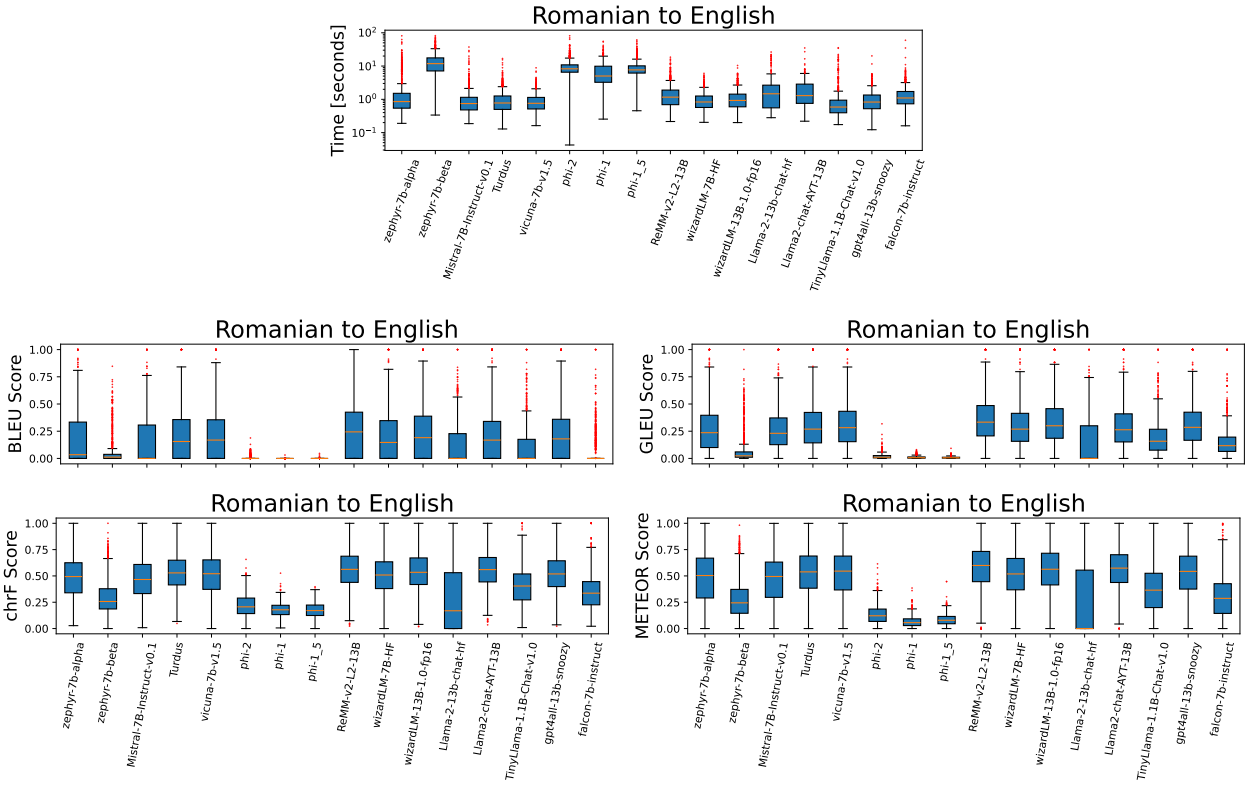

Figure 49: Romanian-to-English dataset per-sentence translation quality and timing statistics

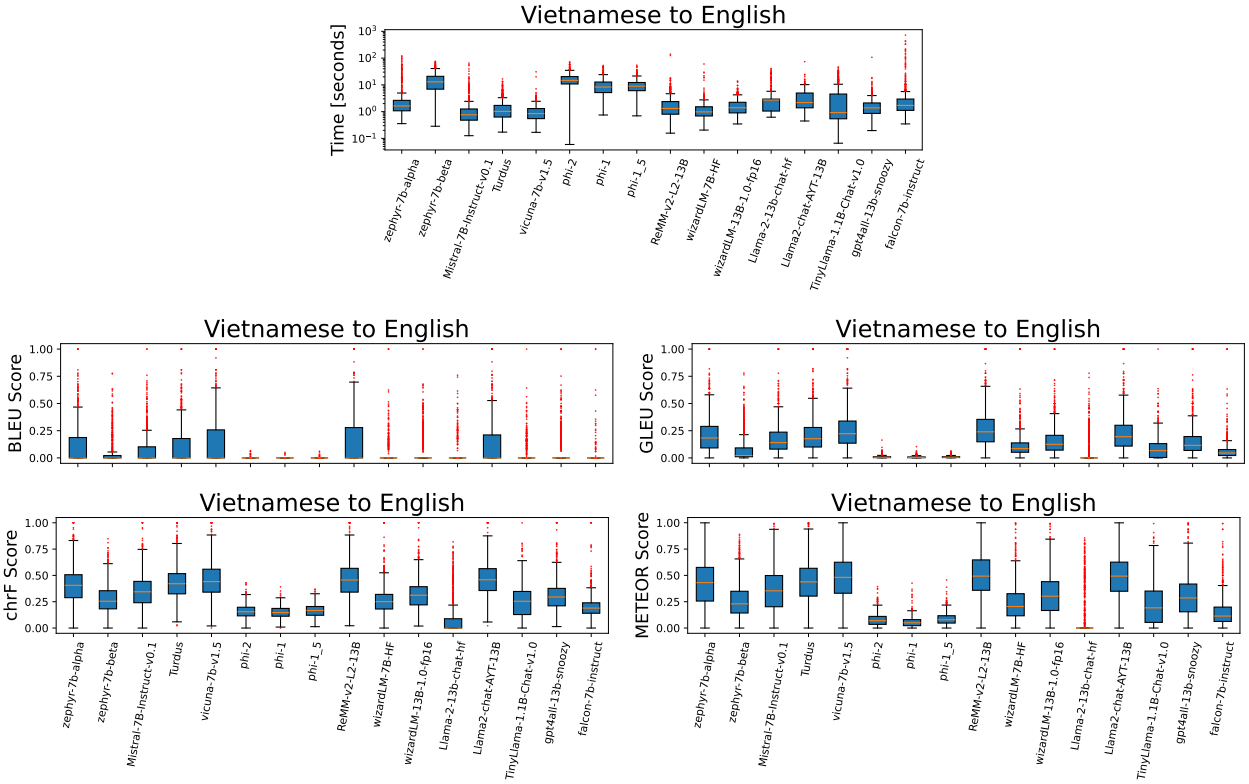

Figure 50: Vietnamese-to-English dataset per-sentence translation quality and timing statistics