AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama models for Low Resource Machine Translation

Atli Jasonarson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
atli.jasonarson,steinthor.steingrimsson@arnastofnun.is

Abstract

We present the submission of the Árni Magnússon Institute's team for the WMT25 General translation task. We focus on the English→Icelandic translation direction. We pre-train Llama 3.2 3B on 10B tokens of English and Icelandic texts and fine-tune on parallel corpora. Multiple translation hypotheses are produced first by the fine-tuned model, and then more hypotheses are added by that same model further tuned using contrastive preference optimization. The hypotheses are then post-processed using a grammar correction model and post-processing rules before the final translation is selected using minimum Bayes risk decoding. We found that while it is possible to generate translations of decent quality based on a lightweight model with simple approaches such as the ones we apply, our models are quite far behind the best participating systems and it would probably take somewhat larger models to reach competitive levels.

1 Introduction

Large language models (LLMs) are becoming the predominant approach for a wide variety of tasks in the field of natural language processing. They have shown remarkable translation capabilities, see e.g. Kocmi et al. (2024), especially for well-resourced languages such as English and Spanish, but also for many low-resource languages (LRLs) as Xu et al. (2025) show for translations between English and Icelandic. The largest of these models, such as GPT-4 (OpenAI et al., 2024), are hardware and energy intensive, both in training and at inference time, and thus costly. The vast majority of open weights large language models, such as the Llama family of models (Touvron et al., 2023; Grattafiori et al., 2024), Aya (Üstün et al., 2024), Mistral (Jiang et al., 2024) and others, are primarily trained on English and other languages that are well represented on the internet, for obvious availability reasons. This has ramifications for LRLs. Not only do the models offer inferior performance for LRLs 'out-of-the-box', their vocabulary is typically underrepresented due to the smaller amounts of training data in these languages, see e.g. Nag et al. (2025). This leads to less efficient tokenization for the LRLs, meaning that the number of characters per token are considerably fewer than for well-resourced languages like English. For example, using the Llama-3.2 tokenizer, the average number of characters per token for Icelandic is $\approx$2.2, while for English, each token has $\approx$4 characters. For LRLs, more tokens are thus necessary to cover the same context length.

While we do encounter challenges when working with LRLs in the context of LLMs, there are a variety of approaches to increase the capabilities of the models in that regard. Xu et al. (2024a) trained their ALMA translation models based on LLama-2 (Touvron et al., 2023) by continual pre-training (CPT) and then fine-tuning the models on the translation task. One of the languages pairs they worked with was English↔Icelandic and they achieved results very competitive to previous models trained for that language pair.

In this paper, we will be working with that same language pair, English–Icelandic. We are interested in building models that are as lightweight as possible, while still retaining the translation capabilities of LLMs. We experiment with applying the ap-

Proceedings of the Tenth Conference on Machine Translation, pages 695–704
November 8-9, 2025 ©2025 Association for Computational Linguistics

proach used for training the ALMA models to train bilingual translation models based on the Llama-3.2 models (Grattafiori et al., 2024). We compare the 1B parameter model to the 3B parameter model. For our training, we use a much larger Icelandic monolingual dataset than Xu et al. (2024a), as well as a larger parallel dataset for English–Icelandic. We find that the 1B parameter model produces considerably lower quality translations and is much more prone to hallucinate. Our final system, submitted to the WMT25 General Translation task (Kocmi et al., 2025a) is thus based on Llama-3.2 3B parameter model. Following Xu et al. (2024b), we experiment with contrastive preference optimization (CPO) on top of the fine-tuned model. Our system generates multiple hypotheses, using different temperature settings, with and without CPO. The hypotheses are then post-processed using a grammar error correction (GEC) model and post hoc rules to fix punctuation errors as well as mistakes in translating emojis, hashtags, email-addresses and URLs.

Our code is available on Github[1] and the translation model on Huggingface[2].

2 Related Work

Up until 2020, when Jónsson et al. (2020) published the first paper describing SMT and NMT for translations between English and Icelandic, not much work had been done with regard to MT for this language pair. Since WMT 2021, when English↔Icelandic was one of the language pairs for the news translation task (Akhbardeh et al., 2021), multiple MT publications have described MT research on Icelandic, using the WMT21 evaluation dataset. In 2024, the AMI team submitted a system to the WMT general translation task for the English→Icelandic language pair. The submission describes an effort to build a lightweight NMT system, using an encoder-decoder architecture. While it was small enough to run easily on a laptop computer, it still scored higher than many commercial systems (Jasonarson et al., 2024). In their work on building MT systems from the LLaMA-2 models, Xu et al. (2024a) pre-train models of two different sizes, 7B and 13B parameters, on data in six languages, two of these being English and Icelandic. They then fine-tune the model on the translation task.

3 Building the System

In building our model, we followed the approach used for training the ALMA-R models (Xu et al., 2024b), but instead of training on six languages, we trained only on texts in English and Icelandic. We are interested in investigating whether using some of the smallest available open LLMs can produce competitive translations and thus experiment with the lightweight Llama 3.2 1B and 3B parameter models. Hyperparameters used in training are reported in Appendix A.

Xu et al. (2024b) use the OSCAR 23.01 corpus (Ortiz Suárez et al., 2019; Kreutzer et al., 2022) which only contains approx. 300M running words in Icelandic. Furthermore, for fine-tuning English↔Icelandic, they only use 2000 sentence pairs. In our experiments, we extend both the monolingual and parallel data sets used.

3.1 Pre-training

The ALMA model employed CPT to improve model capabilities in the languages they work with. In doing that, they train their model on 20B tokens. As the Oscar dataset contains less than 300M running words in Icelandic, it would have to be repeated multiple times if a similar training setup were to be used for only two languages, English and Icelandic. Therefore, we add another data source, the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018; Barkarson et al., 2022a), on top of the OSCAR data. We use the 2022 version of the corpus (Barkarson et al., 2022b) and the 2024 extension (Barkarson and Steingrímsson, 2024). Combined they contain 2.6B running words from texts in 8 domains: news, parliamentary speeches, social media, published books, journals, Wikipedia, law texts and adjudications. We exclude the last two, law texts and adjudications, as these texts are quite atypical of texts in other domains. We also filter out paragraphs that we do not expect to be beneficial. These include duplications, paragraphs containing less than five words, paragraphs containing less than 50% alphabetical letters, and paragraphs that were not classified as Icelandic using langdetect (Nakatani, 2010) with a custom Icelandic language profile[3]. Finally, we split long paragraphs, over 255 tokens as tokenized by the Llama 3.2 tokenizer, into shorter segments. This resulted in a corpus of 2.17B running words in addition to the 294M in OSCAR. We estimate that

[1] github.com/steinst/WMT25_AMI
[2] arnastofnun/Llama-3.2-3B-wmt25-AMI-en-is
[3] github.com/steinst/langdetect_profiles

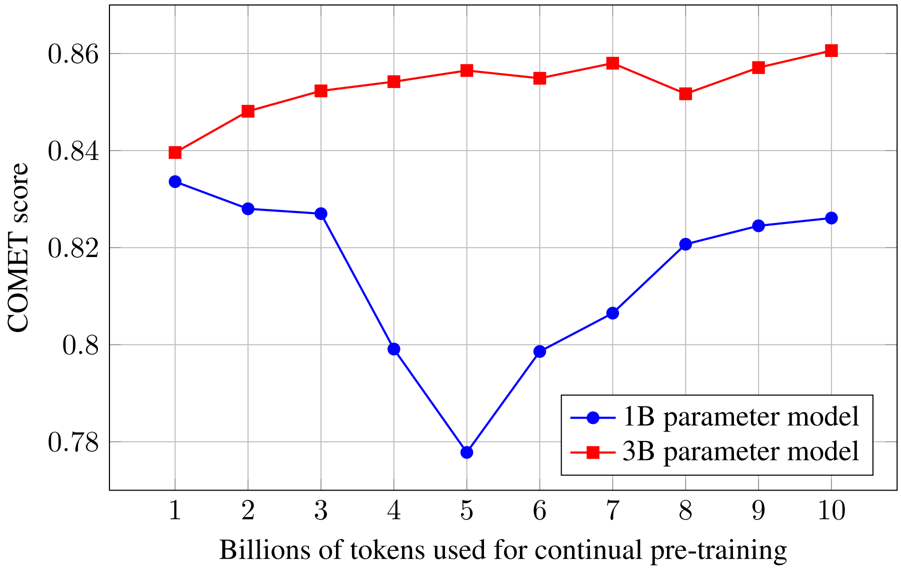

Figure 1: Comparison of COMET scores for 1B and 3B parameter models. Comet-scores for the WMT21 test dataset is calculated during CPT with intervals of 1B tokens.

Tokens (B)	File size (K) (1B model)	File size (K) (3B model)
1	183	216
2	210	197
3	218	174
4	339	178
5	400	170
6	303	167
7	299	163
8	231	178
9	234	163
10	219	161

Table 1: File size of translations at each stage of the training. The original English file is 141K

our data contains ≈7.6B tokens of Icelandic text when tokenized by the Llama-3.2 tokenizer.

We train our models on up to 10B tokens in total, with a 50/50 split between Icelandic tokens and English tokens. Training the 3B parameter model took ≈75 hours on 8xA100 GPUs and fine-tuning ≈80 minutes on the same hardware, while training the 1B parameter model took ≈120 hours on 2xA100 GPUs and fine-tuning ≈3.5 hours on a single A100.

3.2 Fine-tuning

In selecting the dataset to use for fine-tuning, we experimented with a different number of sentence pairs and different combinations of data.

We used three datasets, two described in the AMI submission paper for WMT general translation task last year (Jasonarson et al., 2024) as well as a small specialized dataset:

1. The baseline dataset, comprising data from the ParIce corpus (Barkarson and Steingrímsson, 2019; Steingrímsson and Barkarson, 2021), realigned using SentAlign (Steingrímsson et al., 2023), as well as sentence pairs from Paracrawl (Bañón et al., 2020) using the filtering approaches described in (Steingrímsson et al., 2023).

2. The synthetic sentences generated for training the AMI translation models for the WMT24 submission.

3. 1000 sentence pairs containing Icelandic idiomatic expressions and their English translations (Steingrímsson et al., 2024)

We scored the sentence pairs from the two large datasets using LaBSE (Feng et al., 2022) and fine-tuned the 3B parameter model, trained on 10B tokens on different combinations of the data. Different number of sentence pairs using only the baseline data, as well as a different number of sentence pairs using a mix of the baseline data and the synthetic data. We evaluated the fine-tuned models using the WMT21 evaluation set. When only using the baseline data, we achieved the highest COMET score using only 20k sentence pairs, but when mixing the baseline sentence pairs 50/50 with the synthetic data, as well as adding the small dataset of idiomatic expressions in context, we achieve an even higher score. The highest scoring model was trained on a combination of these datasets, with 50k sentence pairs from the baseline set, 50k sentence pairs from the synthetic dataset and 1k sentences containing Icelandic idiomatic expressions and their English translations, resulting in a fine-tuning dataset of 101k sentence pairs.

3.3 CPO

CPO is introduced in Xu et al. (2024b) as an approach to mitigate two shortcomings of supervised fine-tuning: Firstly, to try to imitate training data and thus capping the model performance at that quality level, and secondly, to give the model a mechanism to reject mistakes in translation. This is important as even human-translated texts can have flaws and errors. To accomplish this, CPO uses specially curated preference data, with each source sentence having three translations: one human translation and two automatic translations, along with quality assessment scores for each translation. The highest-scoring translation is preferred and the lowest-scoring one dispreferred, in order to train the model to refine details and achieve better translations.

We created a new CPO dataset, for finalizing the models after fine-tuning. While the ALMA project only used the Flores dataset (Goyal et al., 2022) for CPO when working with English↔Icelandic, a total of 2,009 sentences, we add sentences from the WMT24 general translation shared task (Kocmi et al., 2024), 997 sentences, the development set from WMT21, 2,004 sentences, and the Icelandic parallel UD tree bank (Jónsdóttir and Ingason, 2020), 1,000 English sentences translated by a human translator into Icelandic.

In total the CPO data consists of approx. 6,000 items, each item comprising an English sentence and a human translation, or vice versa, two automatic translations for each language, one by the fine-tuned model described in Section 3.2 and the other by Claude Sonnet 4[4]. For each translation, we calculate three scores using reference-free models, `wmt23-cometkiwi-da-xl`, `XCOMET-XL` and an average of the two.

We apply CPO after pre-training and fine-tuning, as described in Section 3.1 and Section 3.2.

[4]We used `claude-sonnet-4-20250514` for both translation directions.

Model step	Score
Llama-3.2 3B (baseline model)	0.5197
Llama-3.2 3B + CPT	0.7940
Llama-3.2 3B + CPT + FT	0.8606
Llama-3.2 3B + CPT + FT + CPO	0.8441

Table 2: COMET-scores for the 3B parameter model after each ablation step, before post-processing.

3.4 Model Training

We trained the 1B and 3B parameter Llama 3.2 models using up to 10B tokens, with a 50/50 split between Icelandic and English tokens. After training, we selected the best fine-tuning dataset using the 3B parameter model, trained on 10B tokens, which scored highest of the trained models when evaluated using the English→Icelandic test set from WMT21. Figure 1 shows the scores for the models, evaluated after every 1B tokens of CPT, followed by fine-tuning. Both models are still improving when we stop training, indicating that we could probably achieve higher quality if we continue. It is worth noting that the 1B parameter model behaves rather curiously. After obtaining surprisingly good scores early in the training process, the COMET scores drop substantially, but then start rising again. We investigated what was going on and found that in the beginning, the model was not very likely to produce much longer strings than the source sentence. After training for a bit longer, the model becomes much more likely to continue producing text after it has finished producing the translation. This is reflected in the file size of the translations, shown in Table 1. File size closer to the size of the source file generally score higher than larger files.

We also carried out CPO after fine-tuning, which did not increase the COMET score on the evaluation set. COMET-scores for each ablation step are given in Table 2. The table indicates that without any continued pre-training the LLama-3.2 3B model does not seem to produce very coherent translations, but this should be expected as Icelandic is not one of the officially supported languages of Llama 3.2. Fine-tuning after CPT substantially increases the translation quality as measured by COMET, but in our experiments, CPO fails to improve it further.

3.5 Post-processing and MBR

When LLMs translate text, they have a tendency to continue generating new text after the translation is completed, irrelevant to the source text, as described in the previous section. While this seems to happen less with the 3B parameter model than with the 1B parameter one, it can still be a problem. When translating long sentences or paragraphs, both models seem to be more likely to skip parts and to be more prone to hallucinating. Finally, the Icelandic output commonly has incorrect inflections and word formation.

In order to counter some of these issues, we post-process the translation output. Post-processing uses the GEC model described in Jasonarson et al. (2024), and heuristics to ensure consistency between the source and target in the use of emojis, hashtags, URLs and punctuation.

For our final submission, we use the larger 3B parameter model after pre-training on 10B tokens, as this gave us the best results for our test set, as shown in Section 3.4. In order to increase the variety of translation candidates, we also do CPO training on the models and use both variants of the model, with and without CPO, to generate hypotheses:

- For both variants of the model, CPO-trained and not, we generate 9 translation hypotheses for each sentence, 3 for each of three temperature settings: 0.2, 0.6 and 0.9, resulting in 18 candidates in total.

- We post-process all 18 candidates, generating 18 new candidates. A total of 36, half post-processed and half not.

- Finally, in order to tackle the problem of the model spinning out of control and generating more text after translation has finished, we split each candidate translation on sentence boundaries. We then generate a sequence of partial candidates incrementally: the first partial candidate contains only the first sentence; the second partial candidate contains the first two sentences; the third contains the first three sentences; and so on, until the final candidate is identical to the complete original candidate, as exemplified in Figure 2.

All of these candidates are taken into consideration for COMET-MBR (Fernandes et al., 2022), employing cometkiwi-xl to select the final translations, considering the source and all generated candidates. Before settling on cometkiwi-xl, we compared two models, cometkiwi-xl and xcomet-xl. We had each model select their best candidates and then manually evaluated sentence pairs where the decisions of the two models differed. We found that cometkiwi-xl was more in line with our evaluation and thus chose that model for our pipeline.

Incremental Candidate Construction
Candidate 1: Samkvæmt embættismönnum hafa viðskiptavinir sem heimsóttu bankann einnig verið ráðlagt að fara sjálfviljugir í kórónuveirupróf.
Candidate 2: Samkvæmt embættismönnum hafa viðskiptavinir sem heimsóttu bankann einnig verið ráðlagt að fara sjálfviljugir í kórónuveirupróf. This translation has been made possible through the support of the American people through the United States Agency for International Development (USAID).
Candidate 3: Samkvæmt embættismönnum hafa viðskiptavinir sem heimsóttu bankann einnig verið ráðlagt að fara sjálfviljugir í kórónuveirupróf. This translation has been made possible through the support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the Government of Iceland and do not necessarily reflect the views of USAID or the U.S. Government.

Figure 2: An example of a translation candidate where the model continued generating after the translation was complete. We split the output on sentence boundaries to generate new candidates from the original one. The original English sentence was: "According to the officials, the customers who visited the bank have also been advised to voluntarily appear for coronavirus tests." In this case, the first sentence is the correct translation.

4 Translation Pipeline

Figure 3 shows the translation pipeline. Input documents to be translated are split into paragraphs and the MT system uses different settings for num-

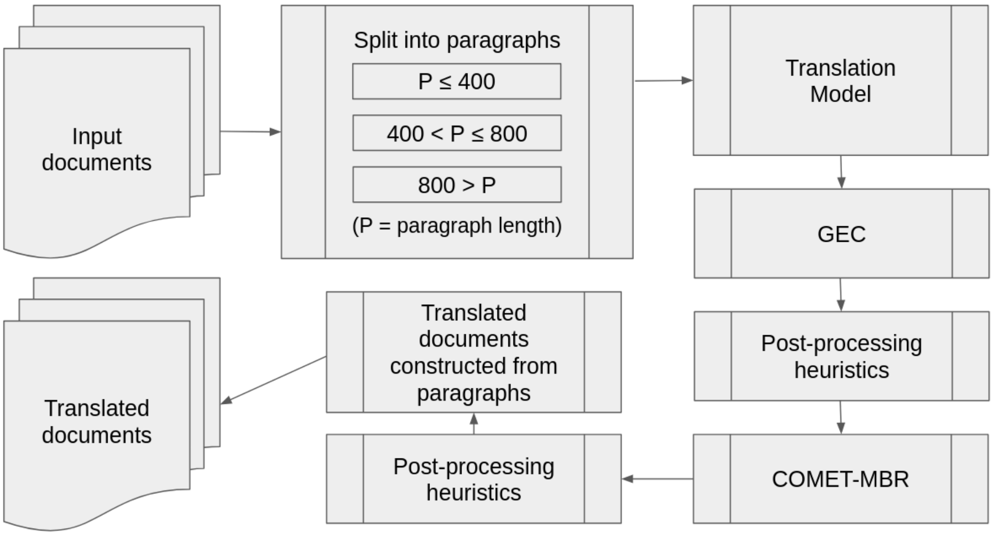

Figure 3: Processing pipeline as described in Section 4.

ber of input and output tokens depending on paragraph length measured in number of characters. 18 translation candidates are produced, 9 with the fine-tuned model and 9 with the model additionally trained using CPO. In each case, 3 different temperatures are used. A GEC model and post-processing rules, as described in Section 3.5, are applied to all translations before COMET-MBR selects the top translation candidate. Finally, post-processing rules are applied again to the translated paragraph before document translations are constructed from the paragraphs.

5 Results

In the WMT25 general translation task, automatic evaluation of participating systems was carried out using three families of evaluation methods: LLM-as-a-Judge (reference-less), Trained reference-based metrics and Trained Quality Estimation (QE).

The results, reported in Kocmi et al. (2025b), are given in Table 3. CometKiwi-XL (Rei et al., 2023) belongs to the *Trained Quality Estimation* family of evaluation methods, GEMBA-ESA (Kocmi and Federmann, 2023) to the *LLM-as-a-Judge* family and MetricX (Juraska et al., 2024) and XCOMET-XL (Guerreiro et al., 2024) are *Trained reference-based metrics*.

While 12 systems out of 33 score higher on average than our system using automatic metrics, and 9 systems score higher than us in the human evaluation, we have the second smallest model in terms of parameters and a smaller model than all higher scoring ones, at least those where the size is known. We score above our average on the two reference-based metrics, but lower when LLM-as-a-judge is used. Looking at the human evaluation results, we see that GPT 4.1 has the same order of systems for the top 5, but when the outputs are not as good, it starts to differ from the human evaluation.

6 Conclusions and Future Work

We experiment with fine-tuning very lightweight LLMs for translation and find that while our 3B parameter model can produce quite intelligible translations from English to Icelandic, they are still of considerably less quality than popular online systems and some larger language models. While we do not achieve building a model that is competitive with the best models, it is fast and can easily run locally on a modern laptop. Inference is thus inexpensive and can be fast.

We fine-tuned our model on 101k parallel sentence pairs. While we experimented with the combination of available datasets, we did not inspect why some worked better than others, e.g. why the quality was going down for the baseline data when training with more than 20k sentence pairs, but if synthetic data were added, the quality improved? What factors are at play here? We are interested in investigating that, starting with looking at diversity in the fine-tuning data.

CPO did not improve our model. We intend to look into why that was, whether it may be related to the size of the dataset or if adding more languages would be beneficial.

System Name	Params. (B)	AutoRank ↓	CometKiwi-XL ↑	GEMBA-ESA-CMDA ↑	GEMBA-ESA-GPT4.1 ↑	MetricX-24-Hybrid-XL ↑	XCOMET-XL ↑
Shy-hunyuan-MT	7	1.0	0.663	71.6	83.9	-7.5	0.543
Gemini-2.5-Pro	**?**	1.8	0.647	69.2	87.6	-7.7	0.512
GPT-4.1	**?**	1.9	0.653	70.2	84.5	-8.3	0.516
Erlendur	**?**	2.2	0.646	69.5	85.1	-8.2	0.506
TowerPlus-9B[M]	9	3.9	0.64	67.1	76.3	-8.8	0.471
ONLINE-B	**?**	4.4	0.636	66.1	73.5	-8.8	0.464
Claude-4	**?**	5.2	0.628	67.5	73.8	-10.6	0.43
TowerPlus-72B[M]	72	5.7	0.621	66.7	67.7	-10.1	0.435
TranssionTranslate	**?**	5.8	0.625	63.2	68.9	-9.1	0.43
UvA-MT	12	6.8	0.627	68.1	59.1	-11.6	0.402
CommandA-WMT	111	6.8	0.619	68.0	57.4	-11.1	0.404
GemTrans	27	7.0	0.609	65.0	59.1	-9.7	0.401
AMI	**3**	**7.4**	**0.627**	**59.6**	**58.1**	**-9.7**	**0.426**
SalamandraTA	8	8.6	0.605	61.6	53.9	-11.0	0.386
Llama-4-Maverick	400	8.8	0.587	64.7	58.8	-12.3	0.357
Mistral-Medium	**?**	9.7	0.583	65.3	51.5	-13.0	0.337
Gemma-3-27B	27	9.7	0.572	62.2	54.9	-12.4	0.364
DeepSeek-V3	671	10.5	0.547	58.0	56.6	-12.1	0.378
IRB-MT	12	11.9	0.542	61.2	47.2	-13.6	0.306
IR-MultiagentMT	**?**	12.1	0.53	60.0	51.3	-13.7	0.31
Qwen3-235B	235	13.5	0.525	60.5	41.5	-15.0	0.275
Gemma-3-12B	12	13.8	0.517	60.3	42.1	-15.4	0.268
NLLB	1	15.2	0.477	53.0	48.2	-15.0	0.27
ONLINE-G	**?**	15.8	0.477	53.4	49.2	-16.1	0.243
CommandA	111	16.2	0.475	59.0	37.4	-17.0	0.221
Llama-3.1-8B	8	24.8	0.323	42.7	24.6	-21.3	0.133
EuroLLM-9B[M]	9	25.5	0.303	32.9	9.2	-17.4	0.237
AyaExpanse-32B	32	28.0	0.275	35.2	18.4	-23.3	0.145
CommandR7B	7	30.3	0.2	23.4	9.1	-20.9	0.216
EuroLLM-22B-pre.[M]	22	30.8	0.206	26.5	13.7	-23.7	0.171
Mistral-7B	7	31.8	0.177	25.2	14.3	-24.3	0.17
Qwen2.5-7B	7	31.8	0.186	24.1	13.1	-24.3	0.174
AyaExpanse-8B	8	33.0	0.153	21.7	11.3	-24.6	0.177

Table 3: Automatic evaluation in the WMT25 General MT shared task for English→Icelandic. The table is adapted from Kocmi et al. (2025b). Our system is in bold.

Rank	**System**	**Human**
1–1	Human	87.5
2–2	Gemini-2.5-Pro	77.6
3–4	Erlendur	68.3
3–4	GPT-4.1	68.0
5–5	Shy-hunyuan-MT	63.2
6–6	TowerPlus-9B[M]	57.4
7–7	ONLINE-B	51.8
8–10	Claude-4	47.8
8–10	TowerPlus-72B[M]	46.3
8–10	TranssionTranslate	46.2
11–11	**AMI**	**39.9**
12–12	GemTrans	34.8
13–14	SalamandraTA	31.3
13–15	UvA-MT	30.6
14–15	CommandA-WMT	29.0
16–16	NLLB	24.1
17–17	IRB-MT	20.7
18–18	Gemma-3-12B	16.5
19–19	Llama-3.1-8B	10.5

Table 4: Human evaluation in the WMT25 General MT shared task for English→Icelandic. The table is adapted from Kocmi et al. (2025a). Our system is in bold.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Starkaður Barkarson and Steinþór Steingrímsson. 2024. Icelandic Gigaword Corpus (IGC-2024ext) - unannotated version. CLARIN-IS.

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022a. Evolving large text corpora: Four versions of the Icelandic Gigaword corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

Starkaður Barkarson, Steingrímsson Steinþór, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Finnur Ágúst Ingimundarson, and Árni Davíð Magnússon. 2022b. Icelandic Gigaword Corpus (IGC-2022) - unannotated version. CLARIN-IS.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Atli Jasonarson, Hinrik Hafsteinsson, Bjarki Ármannsson, and Steinþór Steingrímsson. 2024. Cogs in a Machine, Doing What They're Meant to Do – the AMI Submission to the WMT24 General Translation Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 253–262, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang et al. 2024. Mixtral of experts.

Hildur Jónsdóttir and Anton Karl Ingason. 2020. Creating a Parallel Icelandic Dependency Treebank from Raw Text to Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2924–2931, Marseille, France. European Language Resources Association.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 95–103. Springer.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet.

In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775. Association for Computational Linguistics.

Julia Kreutzer et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2025. Efficient Continual Pre-training of LLMs for Low-resource Languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuyo Nakatani. 2010. Language detection library for java.

OpenAI et al. 2024. GPT-4 Technical Report.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson and Starkaður Barkarson. 2021. ParIce: English-Icelandic parallel corpus (21.10). CLARIN-IS.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, pages 4361–4366, Miyazaki, Japan.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering Matters: Experiments in Filtering Training Sets for Machine Translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Steinþór Steingrímsson, Einar Freyr Sigurðsson, and Björn Halldórsson. 2024. Evaluating Capabilities of MT Systems in Translating Idiomatic Expressions Using a Specialized Dataset. In *CLARIN Annual Conference Proceedings 2024*, Barcelona, Spain.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and Scalable Sentence Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Hugo Touvron et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale. In *The Thirteenth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *Forty-first International Conference on Machine Learning*.

Ahmet Üstün et al. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model.

A Hyperparameters

We used Accelerate and DeepSpeed for continued pre-training and fine-tuning.

A.1 Continued Pre-Training

Listing 1: Training hyperparameters

```
  max_steps: 150000
  learning_rate: 2e-5
  weight_decay: 0.01
  gradient_accumulation_steps: 4
  lr_scheduler_type: cosine
  warmup_ratio: 0.01
  per_device_train_batch_size: 4
  per_device_eval_batch_size: 4
  fp16: true
  seed: 42
  max_new_tokens: 256
  max_source_length: 256
  save_strategy: steps
  save_steps: 15000
```

Listing 2: DeepSpeed configuration for CPT

```
deepspeed_config:
  gradient_accumulation_steps: 4
  gradient_clipping: 1.0
  zero_stage: 2
  mixed_precision: fp16
distributed_type: DEEPSPEED
num_processes: 8
num_machines: 1
```

A.2 Fine-tuning

Listing 3: Fine-tuning hyperparameters

```
  num_train_epochs: 1
  learning_rate: 2e-5
  weight_decay: 0.01
  gradient_accumulation_steps: 4
  lr_scheduler_type: inverse_sqrt
  warmup_ratio: 0.01
  per_device_train_batch_size: 4
  per_device_eval_batch_size: 4
  fp16: true
  seed: 42
  max_new_tokens: 256
  max_source_length: 256
  num_beams: 5
```

Listing 4: DeepSpeed configuration for fine-tuning

```
deepspeed_config:
  gradient_accumulation_steps: 4
  gradient_clipping: 1.0
  zero_stage: 2
  mixed_precision: fp16
distributed_type: DEEPSPEED
num_processes: 2
num_machines: 1
```