

GMU WMT 2026 CreoleMT System Description

Chutong Meng
George Mason University
cmeng2@gmu.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Abstract

We present our submissions to the machine translation track of the WMT26 Creole Language Translation Shared Task (Robinson et al., 2026). Our systems cover all English-Creole language pairs supported by the KREYÒL-MT dataset (Robinson et al., 2024), comprising 80 translation directions. We mainly explore the use of large language models (LLMs). In particular, we fine-tune LLMs and, given the limited amount of parallel data available for many language pairs, also evaluate few-shot prompting as an alternative to supervised fine-tuning. We also explore data cleaning, temperature-based multilingual data sampling, and extracting additional parallel data from archival grammar books. Our experiments show that LLM-based systems outperform NLLB-based systems in most cases. For the lowest-resource language pairs, few-shot prompting proves substantially more effective than supervised fine-tuning.¹

1 Introduction

Recent advances in machine translation (MT) have been driven by large language models (LLMs) trained on massive multilingual corpora (Xu et al., 2024; Alves et al., 2024; Xu et al., 2025; Zheng et al., 2025; Finkelstein et al., 2026). These models have demonstrated superior performance over traditional encoder-decoder MT models such as NLLB (NLLB Team et al., 2022).

The Creole-MT shared task provides the publicly available KREYÒL-MT dataset (Robinson et al., 2024), which supports many-to-many translation across 40 Creole languages. In this work, we focus on English↔Creole language translation. Among the 40 Creole languages, only Haitian (hat) has substantial training data, with more than 600K parallel sentence pairs, whereas most of the re-

maining languages contain fewer than 1,000 parallel sentence pairs. This extreme data imbalance presents two key challenges: (1) effectively leveraging LLMs for extremely low-resource languages; and (2) acquiring additional parallel data.

To address the first challenge, we investigate both supervised fine-tuning and few-shot prompting with LLMs. Our experiments show that few-shot prompting substantially outperforms supervised fine-tuning for the lowest-resource language pairs. To address the second challenge, we use a vision language model (VLM) to extract incidental parallel data from archival grammar books (Virk et al., 2020). Although the extracted bitext is noisy, incorporating it into training often improves translation quality.

2 Data

We mainly rely on the parallel data between English and 40 Creole languages from the official KREYÒL-MT dataset (Robinson et al., 2024). As shown in Figure 1, the dataset is extremely imbalanced. hat contains more than 600k sentence pairs, whereas most languages contain fewer than 1,000. We first apply sentence cleaning and filtering to the training data (Section 2.1). We then augment the training data with additional parallel sentences extracted from archival grammar books (Section 2.2).

2.1 Data Cleaning and Filtering

We first preprocess sentences by removing parentheses and text inside them, as such content is often noisy and rarely contributes to the sentence meaning (e.g., "[. . .]").

Motivated by Koehn et al. (2018, 2019), we further filter out noisy sentence pairs that satisfy any of the following conditions:

1. Fewer than 50% of the characters are alphabetic;
2. Either sentence contains fewer than 3 words or more than 60 words;

¹We release the additional parallel sentences at <https://github.com/mct10/CreoleMT-2026> to ensure the reproducibility of our work.

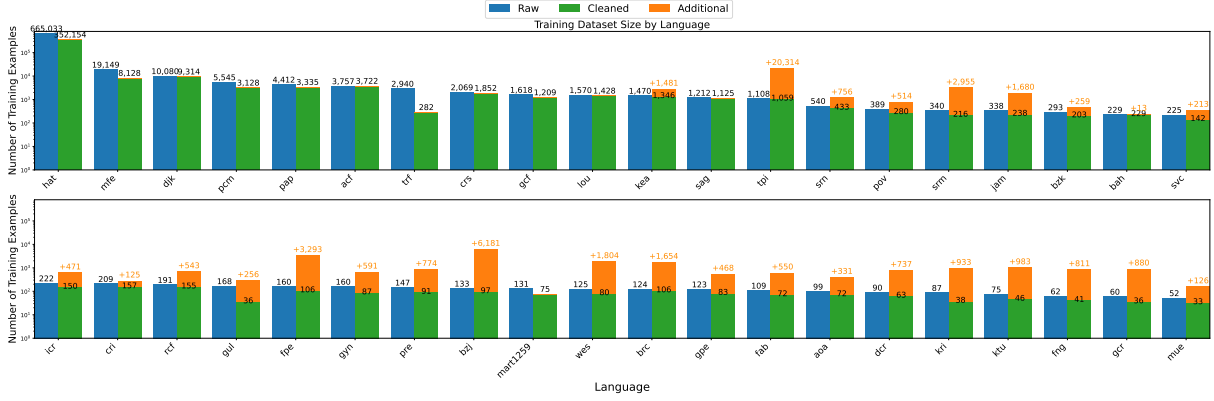


Figure 1: Number of sentence pairs of the raw training data, the cleaned training data, and the additional data we extracted from Virk et al. (2020). The y-axis is in log-scale.

3. The source-to-target or target-to-source character length ratio exceeds 2.2;
4. The Levenshtein ratio exceeds 0.8;
5. Either side is identified as the wrong language by the language identification model released with NLLB (NLLB Team et al., 2022). Specifically, the English sentence must be identified as English, while the Creole sentence must not be identified as English.

2.2 Additional Bitext from Archival Data

Because most languages contain fewer than 1,000 parallel sentence pairs, we extract additional bitext to help with model training. We use the DReaM corpus (Virk et al., 2020) as our source, as it contains linguistic resources, including grammar books and dictionaries, for more than 4,000 languages.

Since the documents are available only in PDF format, we first perform optical character recognition (OCR) using dots.ocr (Li et al., 2025), which provides strong multilingual OCR performance.

Next, we use the VLM Qwen/Qwen3.5-72B-FP8² (Qwen Team, 2026a) to extract incidental bitext from each page independently. Our key observation is that parallel text often exhibits distinctive visual layouts, such as interlinear glosses or adjacent bilingual columns, making it well suited for VLM-based extraction.

To reduce computational cost, we apply this pipeline only to the 28 lowest-resource languages. As shown in Figure 1, the extracted bitext substantially increases the amount of training data for these languages.

We acknowledge that the extracted bitext is noisy. Some sentence pairs may use non-standard or

Translate this from SRC_LANG to TGT_LANG:
SRC_LANG: SRC_SENTENCE
TGT_LANG: TGT_SENTENCE

Figure 2: Instruction template used for supervised fine-tuning. We only compute loss for TGT_SENTENCE.

thographies, while others may suffer from OCR or VLM errors. Nevertheless, we consider this level of noise acceptable, as large-scale sentence mining methods such as global mining (Schwenk et al., 2021; NLLB Team et al., 2022) also produce imperfect alignments.

We removed sentence pairs for which either side has a Levenshtein similarity ratio greater than 0.9 with a sentence in the dev sets. However, since the test sets are released during the evaluation period, after the systems have been developed, we have not applied the same filtering to remove potential duplicates with the test sets.

3 Experiments

We investigate two approaches based on large language models (LLMs): supervised fine-tuning and few-shot prompting. We also fine-tune a NLLB model (NLLB Team et al., 2022) as baseline.

3.1 Model Fine-tuning

Since hat contains substantially more training data than other languages, we train one model exclusively for hat and a separate multilingual model for the remaining languages.

For multilingual training, we employ temperature-based sampling to upsample

²<https://huggingface.co/Qwen/Qwen3.5-72B-FP8>

System: You are a professional translator. Do not say you do not speak the language. Do not say you do not have enough information. You must provide a translation. Do not include explanations, comments, or extra text. The entire response must be a single line containing only the translation. When provided with example translations, use them as reference to help ensure correct translation.

User: Translate the following sentence from SRC_LANG to TGT_LANG: SRC_SENTENCE
Below are some example translations that might be useful:
SRC_LANG_1: SRC_SENTENCE_1
TGT_LANG_1: TGT_SENTENCE_1
SRC_LANG_2: SRC_SENTENCE_2
TGT_LANG_2: TGT_SENTENCE_2
...
Now, please provide the translation.
SRC_LANG: SRC_SENTENCE
TGT_LANG:

Figure 3: Instruction template used for few-shot prompting.

low-resource languages:

$$p_i = \frac{|D_i|^{1/T}}{\sum_j |D_j|^{1/T}}$$

where D_i denotes the number of training sentence pairs for language i , T is the temperature, and p_i is the probability of sampling a sentence pair from language i at each training step. We mainly consider $T = 5$, following the setting used by the baseline model.³

We fine-tune the pre-trained Gemma-3 model, gemma-3-4b-pt (Gemma Team et al., 2025) (hereafter gemma-3), using LoRA adapters (Hu et al., 2022), due to its strong multilingual capabilities. The prompt template is shown in Figure 2. We compute the training loss only over the target translation.

As a baseline, we also fine-tune nllb-200-distilled-1.3B (NLLB Team et al., 2022) (hereafter nllb) using LoRA adapters. For Creole languages not supported by NLLB, we introduce new language code tokens.

We choose relatively small models (4B for Gemma-3 and 1.3B for NLLB) mainly due to computational constraints and small training sets.

3.2 Few-shot Prompting

Even after augmenting the training data with incidental bitext, a few thousand sentence pairs are often insufficient for effective fine-tuning. We therefore investigate few-shot prompting as an alterna-

tive. We use Qwen/Qwen3.6-27B⁴ (Qwen Team, 2026b) as the backbone model. We choose a much bigger model than those used for fine-tuning because few-shot prompting requires much less computation than fine-tuning. To enable inference on a single 40GB GPU, we apply 4-bit quantization.

The prompt template is in Figure 3. For each input sentence, we retrieve similar examples from the training set of the corresponding language using the chrF2++ score (Popović, 2017). The 512 highest-scored sentence pairs are then included as in-context examples. If the training set contains fewer than 512 sentence pairs, we use all available pairs.

3.3 Evaluation Metrics

We evaluate our systems on the KREYÒL-MT dev set. We leave the dev set untouched and do not apply any cleaning or filtering strategies.

We report spBLEU⁵ with the flores200 tokenizer (NLLB Team et al., 2022) and chrF2++⁶ scores, computed by SacreBLEU (Post, 2018).

4 Results and Analysis

We report the results of fine-tuning nllb-200-distilled-1.3B and gemma-3-4b-pt in Section 4.1. Results for few-shot experiments are presented in Section 4.2.

4.1 Model Fine-tuning

Results for the eng-hat system and the multilingual systems are presented in Table 1 and Table 2. Given the large number of language directions, we report the average performance of the multilingual systems.

For both nllb and gemma-3, we progressively introduce modifications to improve performance. The baseline models are fine-tuned on the raw KREYÒL-MT dataset.

Applying data cleaning yields significant gains for eng→hat with gemma-3. In contrast, it degrades the performance of the multilingual systems. We hypothesize that data cleaning reduces the training data to a level that is insufficient for these low-resource language pairs.

Temperature-based data sampling does not improve the performance of multilingual nllb. In

⁴<https://huggingface.co/Qwen/Qwen3.6-27B>

⁵nrefs:1 + case:mixed + eff:no + tok:flores200 + smooth:exp + version:2.6.0

⁶nrefs:1 + case:mixed + eff:yes + nc:6 + nw:2 + space:no + version:2.6.0

³<https://huggingface.co/pranavgupta/gemma4-creole-mt-2026-baseline-adapter>

Method	hat		Multilingual w/o hat	
	spBLEU	chrF2++	spBLEU	chrF2++
nllb-out-of-box	38.67	55.73	–	–
nllb-ft	45.91	63.31	16.07	32.24
+Data clean	45.09	63.21	13.32	28.93
+ $T = 5$	–	–	14.08	28.12
+Extra data	–	–	16.77	33.41
gemma-3-ft	48.06	65.90	19.85	38.16
+Data clean	52.13	67.18	17.97	35.78
+ $T = 5$	–	–	<u>23.54</u>	<u>41.60</u>
+Extra data	–	–	24.45	43.11

Table 1: Dev set performance from English *into Creole languages*. nllb-ft and gemma-3-ft denote fine-tuned versions of nllb-200-distilled-1.3B and gemma-3-4b-pt, respectively. **Bold** indicates the best result, and underlining indicates the second-best result.

Method	hat		Multilingual w/o hat	
	spBLEU	chrF2++	spBLEU	chrF2++
nllb-out-of-box	41.72	59.49	–	–
nllb-ft	55.72	71.77	30.72	47.58
+Data clean	56.04	71.90	29.24	46.42
+ $T = 5$	–	–	26.68	42.52
+Extra data	–	–	36.41	53.49
gemma-3-ft	57.89	73.03	33.29	49.69
+Data clean	58.37	<u>72.91</u>	32.79	49.11
+ $T = 5$	–	–	32.92	49.02
+Extra data	–	–	<u>35.83</u>	<u>52.64</u>

Table 2: Dev set performance from Creole languages *into English*. nllb-ft and gemma-3-ft denote fine-tuned versions of nllb-200-distilled-1.3B and gemma-3-4b-pt, respectively. **Bold** indicates the best result, and underlining indicates the second-best result.

contrast, it substantially improves multilingual gemma-3 for translation into Creole languages, increasing the average BLEU score from 17.97 to 23.54.

Finally, incorporating additional training data consistently improves both base models. For translation into Creole languages, the additional data improves multilingual nllb by more than 2 BLEU and multilingual gemma-3 by nearly 1 BLEU. For translation into English, the gains are even larger. There is nearly 10 BLEU improvement for multilingual nllb and almost 3 BLEU improvement for multilingual gemma-3. These results suggest that incorporating additional training data is beneficial, even when the data may contain some noise.

4.2 Few-shot Prompting

Due to time and computational constraints, we conduct few-shot experiments on only 20 languages (40 translation directions). The results are presented in Table 3.

We are interested in comparing few-shot prompting and fine-tuning for LLMs under these extremely

low-resource settings. We use our best multilingual fine-tuned gemma-3-4b-pt model as the baseline, incorporating all methods described in the previous section. For few-shot prompting, we evaluate two settings: with and without the additional training data.

Among the 40 translation directions, the fine-tuned model achieves the best performance on only 7 directions, 5 of which are translation into English. This result suggests that, under such limited-data conditions, few-shot prompting is generally more effective than fine-tuning, especially for into-low-resource-language translation.

Consistent with our observations from the fine-tuning experiments, incorporating the additional training data generally improves few-shot performance. Few-shot prompting *with* the additional data achieves the best results on 22 translation directions, compared with 11 directions for few-shot prompting *without* the additional data.

Model	aoa		bah		brc		bzk		cri	
	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng
Best gemma-3-ft	15.02	20.20	18.32	20.47	6.26	18.45	26.60	46.11	21.37	28.57
Few-shot w/o extra data	23.37	16.26	16.43	21.38	12.57	10.58	34.01	50.85	27.92	22.28
Few-shot w/ extra data	27.75	20.28	11.38	20.81	15.08	20.37	35.24	53.57	31.57	23.99
Zero-shot	0.28	1.97	3.06	20.70	0.46	2.91	0.94	35.32	2.12	3.10

Model	dcr		fab		fng		gcr		gpe	
	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng
Best gemma-3-ft	5.05	33.23	2.84	12.63	19.82	12.48	18.49	24.88	15.49	28.03
Few-shot w/o extra data	19.53	27.67	5.22	10.37	29.62	21.38	22.06	34.75	28.59	29.59
Few-shot w/ extra data	19.42	38.40	6.92	16.28	24.57	21.02	26.86	33.78	10.86	30.94
Zero-shot	0.77	18.68	0.56	0.76	0.71	7.51	2.33	26.81	0.98	14.27

Model	gul		gyn		icr		jam		kri	
	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng
Best gemma-3-ft	8.09	30.64	27.70	32.47	55.19	62.85	35.28	46.67	18.80	33.18
Few-shot w/o extra data	22.00	39.32	25.42	33.88	53.21	57.14	52.47	78.03	28.85	41.91
Few-shot w/ extra data	22.15	45.30	29.54	34.20	55.75	58.32	43.25	79.21	23.02	38.12
Zero-shot	15.87	34.01	1.46	19.51	0.85	19.02	6.60	70.20	8.64	21.21

Model	pov		pre		rcf		svc		tpi	
	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng	eng→	→eng
Best gemma-3-ft	39.58	51.07	36.92	52.91	18.58	26.81	42.21	45.77	26.72	53.93
Few-shot w/o extra data	40.62	39.61	35.11	30.14	26.58	32.88	46.26	44.96	26.57	34.46
Few-shot w/ extra data	35.14	51.28	41.31	40.80	24.31	36.30	49.61	44.38	24.38	34.22
Zero-shot	7.86	23.63	1.19	4.46	3.53	24.29	10.15	30.26	9.14	28.22

Model	# Wins	
	eng→	→eng
Best gemma-3-ft	2	5
Few-shot w/o extra data	7	4
Few-shot w/ extra data	11	11

Table 3: Dev set spBLEU scores of the best fine-tuned model and few-shot experiments for 20 selected languages. Few-shot and zero-shot experiments are performed using Qwen3.6-27B.

4.3 Submissions

Our primary submission is the best performing gemma-3-4b-pt model, fine-tuned with data cleaning, temperature-based sampling, and additional training data. Since no additional data were extracted for eng↔hat, this language pair is excluded. The primary system covers 78 translation directions.

Our contrastive 1 system is the gemma-3-4b-pt model fine-tuned on the cleaned data with temperature-based sampling. For eng↔hat, where data sampling is unnecessary, we use the gemma-3-4b-pt model fine-tuned on the cleaned eng↔hat data. These two systems cover all 80 translation directions.

Our contrastive 2 system uses the 512-shot prompting setup with additional training data and the Qwen3.6-27B backbone model. This system covers 20 selected languages (40 translation direc-

tions).

5 Conclusion

In this paper, we describe GMU systems for the WMT 2026 CreoleMT shared task. We explore both fine-tuning and few-shot prompting for low-resource translation. Our experiments show that incorporating additional parallel data automatically extracted from archival resources often improves translation performance, even when the data contains noise. Moreover, few-shot prompting generally outperforms fine-tuning when only a very limited amount of parallel data is available, especially for translation into low-resource languages.

Limitations

We did not perform a thorough audit of the parallel sentences automatically extracted from archival resources. As a result, the extracted data likely suf-

fers from issues such as orthography inconsistencies and language mismatch. More careful cleaning and normalization could further improve the quality of the extracted corpus and, consequently, downstream translation performance.

While we removed sentence pairs that overlap with sentences in the dev sets from the extracted additional data, some overlap with the test data may still remain.

Acknowledgments

We are grateful to the anonymous reviewers and area chairs for WMT whose feedback improved the paper. This work was financially supported by the National Science Foundation under CAREER award 2439202. We acknowledge the support by resources provided by the Office of Research Computing at George Mason University (<https://orc.gmu.edu>) and funded in part by grants from the NSF (Award Number 2018631).

References

- Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). In *First Conference on Language Modeling*.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. [TranslateGemma technical report](#). *Preprint*, arXiv:2601.09012.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi  re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Philipp Koehn, Francisco Guzm  n, Vishrav Chaudhary, and Juan Pino. 2019. [Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.
- Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. [Findings of the WMT 2018 shared task on parallel corpus filtering](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.
- Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. 2025. [dots.ocr: Multilingual document layout parsing in a single vision-language model](#). *Preprint*, arXiv:2512.02498.
- NLLB Team, Marta R. Costa-juss  , James Cross, Onur   elebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Maja Popovi  . 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Qwen Team. 2026a. [Qwen3.5: Towards native multimodal agents](#).
- Qwen Team. 2026b. [Qwen3.6-27B: Flagship-level coding in a 27B dense model](#).
- Nathaniel Robinson, Ruth-Ann Armstrong, Claire Bizon Monroc, Rasul Dent, Pranav Gupta, Raj Dabre, Andre Coy, and Kenton Murray. 2026. Findings of the second shared task for creole language machine translation, at WMT 2026. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Lo  c Grobol, Hasan Muhammad, Ashi Garg, Naome A. Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Dean Stutzman, Bismarck Bamfo Odoom, Sanjeev Khudanpur, Stephen D. Richardson, and Kenton Murray. 2024. [Krey  l-MT: Building MT for Latin American, Caribbean and colonial African creole languages](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

Papers), pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. [CCMatrix: Mining billions of high-quality parallel sentences on the web](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020. [The DReaM corpus: A multilingual annotated corpus of grammars for the world’s languages](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 878–884, Marseille, France. European Language Resources Association.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. [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*.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. [Hunyuan-mt technical report](#). *Preprint*, arXiv:2509.05209.