CUNI at WMT25 General Translation Task

Miroslav Hrabal, Josef Jon, Martin Popel, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics
{hrabal,jon,popel,bojar}@ufal.mff.cuni.cz

Abstract

This paper describes the CUNI submissions to the WMT25 General Translation task, namely for the English to Czech, English to Serbian, Czech to German and Czech to Ukrainian language pairs. We worked in multiple teams, each with a different approach, spanning from traditional, smaller Transformer NMT models trained on both sentence and document level, to fine-tuning LLMs using LoRA and CPO. We show that these methods are effective in improving automatic MT evaluation scores compared to the base pretrained models.

1 Introduction

We have entered the shared task as a number of small teams with different approaches, each with its own submission. We will describe the datasets, methods and evaluation results of each submission in the following sections. Here we will present just a brief overview of all the systems we submitted.

CUNI-MH-v2 is a constrained system trained on partially synthetic data sampled from the CzEng 2.0 (Kocmi et al., 2020) dataset using LoRA (Hu et al., 2021) and Contrastive Preference Optimization (Xu et al., 2024). We will release both the model weights and the filtered training data. The model itself is fine-tuned from the EuroLLM-9B-Instruct model. We currently only support two language directions, (en→cs) and (cs→de), and offer separate LoRA adapters for each. The translations were done on the paragraph level.

CUNI-EdUKate-v1 is an unconstrained system trained on educational domain data using LoRA, SFT, and Contrastive Preference Optimization. It is also fine-tuned from the EuroLLM-9B-Instruct model. It only supports cs2uk language direction and, unlike CUNI-MH-v2, both training and inference were done on sentence level.

CUNI-SFT models were created by a simple supervised finetuning using LoRA on a small amount of publicly available training data.

CUNI-Transformer and **CUNI-DocTransformer** are resubmissions of systems from previous years.

2 Methods

This section describes the approaches used for training our submissions.

2.1 CUNI-SFT (en2cs, en2sr, cs2uk)

We have finetuned multiple pretrained models for document-level and sentence-level translation using LoRA. We have used learning rate $lr = 2e-4$, LoRA ranks $r = 8$ and $r = 16$, LoRA $\alpha = 2 * r$ and batch size of 2 with 16 gradient accumulation steps, resulting in effective batch size of 32. We trained the models for 10k updates. We compared sentence-level translation without context, sentence-level with context shown to the LLM and pure document-level prompt. The prompts are shown in Section A.

2.2 CUNI-MH-v2 (en2cs, cs2de)

Considering that EuroLLM-9B-Instruct is already reasonably good at English to Czech translation, we chose to skip the supervised fine-tuning stage, thereby departing from year's CUNI-MH (Hrabal et al., 2024), and fine-tuned the model solely using Contrastive Preference Optimization (CPO) (Xu et al., 2024).

For the two language directions, (en→cs) and (cs→de), we trained separate LoRA adapters with rank $r = 32$, LoRA $\alpha = 64$, LoRA dropout of 0.05 and effective batch size of 8. We used cosine learning rate scheduler and trained for $10\,\text{k}$ steps.

2.3 CUNI-EdUKate-v1 (cs2uk)

CUNI-EdUKate-v1 was trained from EuroLLM-9B-Instruct model using LoRA in two stages. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 233–240
November 8–9, 2025.

the first stage, we train it on internal sentence-level educational domain parallel data. In the second stage, we train it on partially synthetic internal preference sentence-level educational domain data.

2.4 CUNI-(Doc)Transformer (cs2uk, en2cs)

CUNI-Transformer (cs→uk) and CUNI-DocTransformer (en→cs) are the same systems as submitted in previous years (Jon et al., 2023), relying on standard NMT training with Block backtranslation (Popel, 2018; Popel et al., 2020) and (in the case of CUNI-DocTransformer) document-level training.

3 Data

3.1 CUNI-SFT

We downloaded corpora for Czech to English, Croatian, Serbian[1], Bosnian, German and Ukrainian and English to Croatian, Serbian, German and Ukrainian from OPUS, keeping the document boundaries where possible. The datasets we used are: `DGT`, `DocHPLT`, `ELITR-ECA`, `EMEA`, `GlobalVoices`, `JRC-Acquis`, `News-Commentary`, `SETIMES`, `StanfordNLP-NMT`, `Tatoeba`, `TED2020`, `tico-19`, `TildeMODEL` and `WMT-News`. We scored these datasets with wmt22-cometkiwi-da QE model using Marian. We have selected the top 5% scoring documents (scores are computed on sentence-level and averaged) from each dataset for each direction, with at most 200 documents per dataset and direction. Documents longer than 60 sentences are split into 60-sentence chunks for scoring and training.

3.2 CUNI-MH-v2

In order to create the preference dataset necessary for the CPO method, we first sampled paragraphs from the CzEng 2.0 dataset and translated them using different models. For en→cs dataset, we used EuroLLM-9B Instruct and CUNI-MH from last year. We also used the reference translations as one of the possible candidate translations. For cs→de dataset, we used EuroLLM-9B Instruct, Qwen 2 and Qwen 3.

We then scored the translations (all synthetic candidates and the reference for the en→cs direction) using MetricX24 (used as a reference-free metric). From this, we created (source, preferred, dis-preferred) triplets by taking the highest-scoring translation as preferred and worse scored translations as possible dis-preferred translations.

Unlike the dataset used in previous year, where we gradually built paragraphs sentence by sentence (Hrabal et al., 2024), this year we chose to select the preference on the level of whole documents.

We further filtered these triplets using a version of our work-in-progress experimental metric based on Gemma 3 27b-it model, which we refer to as r1.1. We assigned the MetricX24 and r1.1 scores to each translation candidate. Afterwards, we considered the best candidate with the best MetricX24 score as preferred and all other candidates as dis-preferred. Out of those pairs, we kept only those that met the following criteria:[2]

1. The chosen and rejected translations differ.
2. $\mathrm{MetricX24(chosen)}$ is better than $\mathrm{MetricX24(rejected)}$ by at least 1.0 points.
3. $\mathrm{MetricX24(chosen)} < 10.0$.
4. $r1.1(\mathrm{chosen}) - r1.1(\mathrm{rejected}) \geq 1.0$.

The resulting en→cs dataset consists of 25530 preference triplets, and the cs→de dataset consists of 14797 preference triplets. All datasets and models will be available on Hugging Face:

- en→cs preference dataset: https://huggingface.co/hrabalm/CUNI-MH-v2-encs-data
- cs→de preference dataset: https://huggingface.co/hrabalm/CUNI-MH-v2-csde-data
- en→cs trained model: https://huggingface.co/hrabalm/CUNI-MH-v2-encs
- cs→de trained model: https://huggingface.co/hrabalm/CUNI-MH-v2-csde

3.3 CUNI-EdUKate-v1

For the CUNI-EdUKate-v1 model, we used our internal sentence-level Czech-Ukrainian parallel dataset covering the educational domain. This

[1]We transliterated all Serbian texts written in Cyrillic into the Latin script.

[2]Note that here we work with the raw MetricX24 outputs, which are greater than or equal to 0, and where lower is better.

Table 1: CUNI-MH-v2 en→cs performance on the development set. MetricX24 is google/metricx-24-hybrid-xl-v2p6-bfloat16. CometKiwi22 is Unbabel/wmt22-cometkiwi-da. r1.1 is our internal metric based on Gemma 3 27b-it assigning DA scores.

	wmt23				wmt23-para			
Model	BLEU	MetricX24	CometKiwi22	r1.1	BLEU	MetricX24	CometKiwi22	r1.1
CUNI-MH	36.52	–	83.16	–	35.42	–	**74.82**	–
EuroLLM-9B-Instruct	36.14	−3.74	82.90	89.66	36.69	−7.68	72.67	88.98
CUNI-MH-v2	**37.33**	−3.69	**83.38**	90.36	**37.81**	−7.53	73.75	91.81

Table 2: CUNI-MH-v2 en→cs performance compared with selected WMT24 models on the WMT24 test set.

	wmt24			
Model	BLEU	MetricX24	CometKiwi22	r1.1
Unbabel-Tower70B	24.72	**−3.70**	**83.04**	88.54
Claude-3.5	**32.04**	−4.62	80.79	**90.56**
CUNI-MH	27.62	−4.53	81.10	88.21
EuroLLM-9B-Instruct	26.04	−4.77	80.51	87.19
CUNI-MH-v2	27.89	−4.62	80.99	87.85

dataset is the only reason why our submission is unconstrained.

The creation of the preference dataset for the CPO stage was done in a similar way to the CUNI-MH-v2 model but using different selection of models to generate translation candidates and to score and filter them.

One notable difference was that we also trained EuroLLM-9B-Instruct to predict Direct Assesment scores and used the result as one of the models used to filter the preference triplets.

As a development set, we used 3770 segments split from the training data.

4 Evaluation

4.1 CUNI-SFT

We compared translation quality after finetuning across four pretrained models: `EuroLLM 9B`, `Aya Expanse 8B`, `Mistral Instruct v0.3 7B` and `Granite 3.3 8B`. We measured BLEU (Papineni et al., 2002) and chrF (Popović, 2015) on `newstest2019` (Barrault et al., 2019) in the English to Czech direction, NTREX (Federmann et al., 2022) for English to Serbian and `wmttest24` (Kocmi et al., 2024) for Czech to Ukrainian. The result for simple sentence-level and context-aware sentence-level prompts are shown in Table 3. We do not present results for the doc-level prompt, since we were not able to retrieve sentence-level alignment for source and translated sentences. Overall, we see that our approach to finetuning is effective for languages that are not well covered by the base model. For high resource combinations (e.g. eng-ces in EuroLLM), the finetuning does either not change the evaluation scores, or decreses them.

4.2 CUNI-MH-v2

During inference, we use vLLM and greedy decoding.

In Table 1, we show the performance of the en→cs CUNI-MH-v2 model on the development set. In Table 2, we compare its performance with best performing WMT23 models on WMT23 test set.

Interestingly, we can see that CUNI-MH-v2 improves in BLEU score compared to the base EuroLLM-9B-Instruct model, while we saw the opposite happen in the previous year (Hrabal et al., 2024), where the BLEU/chrF metrics got worse while the COMET22 and CometKiwi22 metrics improved. On the other hand, CUNI-MH-v2 gets higher CometKiwi22 score on sentence-level wmt23 dataset but lower score on the document-level version. Overall, we were able to achieve modest improvements in all metrics compared to the base model on both the development and test set.

For translation of the final WMT25 test set, we use the official script provided by WMT organizers

Context	Language	Model	Base		Finetuned	
			BLEU	ChrF	BLEU	ChrF
Yes	eng-ces	aya-expanse-8b	25.9	57.8	23.3	51.8
		EuroLLM-9B-Instruct	29.9	56.7	28.5	56.2
		granite-3.3-8b-instruct	22.1	51.5	18.5	47.3
		Mistral-7B-Instruct-v0.3	16.9	48.3	15.8	44.3
	eng-srb	aya-expanse-8b	3.3	20.9	7.3	35.2
		EuroLLM-9B-Instruct	15.4	46.6	15.6	46.6
		granite-3.3-8b-instruct	3.1	17.2	4.2	29.8
		Mistral-7B-Instruct-v0.3	2.3	14.8	11.2	40.4
	ces-ukr	aya-expanse-8b	27.3	56.2	25.5	52.0
		EuroLLM-9B-Instruct	28.7	56.4	26.8	54.7
		granite-3.3-8b-instruct	7.0	31.7	6.9	27.6
		Mistral-7B-Instruct-v0.3	15.7	47.7	13.3	39.0
		GPT-4.1-mini	33.7	61.7	-	-
No	eng-ces	aya-expanse-8b	25.4	51.8	26.4	54.9
		EuroLLM-9B-Instruct	31.7	59.0	31.1	59.1
		granite-3.3-8b-instruct	21.8	51.2	22.1	51.5
		Mistral-7B-Instruct-v0.3	13.0	43.4	20.2	49.7
		GPT-4.1-mini	32.5	59.2	-	-
	eng-srb	aya-expanse-8b	8.8	38.0	17.1	47.9
		EuroLLM-9B-Instruct	16.9	48.3	22.6	52.4
		granite-3.3-8b-instruct	6.7	34.8	15.2	45.6
		Mistral-7B-Instruct-v0.3	9.1	41.2	17.4	47.9
		GPT-4.1-mini	29.3	57.9	-	-
	ces-ukr	aya-expanse-8b	24.3	55.1	24.4	51.9
		EuroLLM-9B-Instruct	31.0	59.0	28.2	55.8
		granite-3.3-8b-instruct	6.6	44.4	10.5	35.3
		Mistral-7B-Instruct-v0.3	13.4	39.8	19.2	46.0
		GPT-4.1-mini	33.5	61.6	-	-

Table 3: BLEU and ChrF scores of base and finetuned CUNI-SFT models on devsets (newstest2019 for eng-ces and eng-srb, wmttest2024 for ces-ukr.

to extract paragraph-level segments. During the inference, we further split the paragraphs to chunks of at most 256 tokens by using the sentence-splitter Python library.

4.3 CUNI-EdUKate-v1

We show the automatic metrics of the CUNI-EdUKate-v1 model in Table 4. The EuroLLM-9B-Instruct model, which is also the base model, is used as a baseline.

Table 4: CUNI-EdUKate-v1 automatic metric scores on internal educational domain sentence-level development set.

	dev set	
Model	BLEU	MetricX24
EuroLLM-9B-Instruct	37.4	−3.59
CUNI-EdUKate-v1	**39.1**	**−3.33**

5 Tools

To give a proper credit, we list the tools we used during the development and inference with our models:

CUNI-MH-v2

- transformers (Wolf et al., 2020), peft (Man-

grulkar et al., 2022) and trl (von Werra et al., 2020) libraries for training

- vLLM (Kwon et al., 2023) for inference
- MetricX24 XL[3] (Juraska et al., 2024) for scoring, data filtering, evaluation
- DSPy (Khattab et al., 2024, 2022) and Gemma-3-27b-it (Team et al., 2025) for data filtering

CUNI-EdUKate-v1

- transformers, peft and trl libraries for training
- vLLM for inference
- LINDAT Translation[4] for segmentation and to serve the translation API
- CometKiwi22 (Rei et al., 2022) for scoring, data filtering, evaluation
- MetricX24 XL for scoring, data filtering, evaluation
- Gemma-3-27b-it for data filtering

CUNI-SFT

- transformers, peft and trl libraries for training
- vLLM for inference
- CometKiwi22[5] used through Marian (Junczys-Dowmunt et al., 2018) for data filtering

CUNI-(Doc)Transformer

- Tensor2Tensor (Vaswani et al., 2018)

6 Future work

We have several ideas to improve the performance of the future iterations of our CUNI-MH-v2 model. In particular, we plan to scale up the size of the preference dataset by using a larger portion of CzEng2.0 and by sampling more translation candidates.

We also plan on experimenting with including synthetically translated documents with no reference translations, to augment our dataset with longer examples.

[3] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6-bfloat16

[4] https://github.com/ufal/lindat-translation/

[5] https://huggingface.co/Unbabel/wmt22-cometkiwi-da-marian

7 Conclusion

In this paper, we presented the CUNI submissions to the WMT25 General Translation Task, covering English→Czech, Czech→German, English→Serbian, and Czech→Ukrainian language pairs. Future work will focus on scaling preference datasets and leveraging longer-context translation scenarios.

8 Acknowledgment

This work was supported by the project TQ01000458 (EdUKate) financed by the Technology Agency of the Czech Republic (www.tacr.cz) within the Sigma 3 Programme.

It was also partially supported by the Charles University Grant Agency in Prague (GAUK 244523), by SVV project number 260 821, by Czech Ministry of Education, Youth and Sports (grant MŠMT OP JAK Mezisektorová spolupráce CZ.02.01.01/00/23_020/0008518) and by National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO.

It has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2023062).

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Miroslav Hrabal, Josef Jon, Martin Popel, Nam Luu, Danil Semin, and Ondřej Bojar. 2024. CUNI at WMT24 general translation task: LLMs, (Q)LoRA, CPO and model merging. In *Proceedings of the Ninth Conference on Machine Translation*, pages 232–246, Miami, Florida, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI at WMT23 general translation task: MT and a genetic algorithm. In *Proceedings of the Eighth Conference on Machine Translation*, pages 119–127, Singapore. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

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.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines. In *The Twelfth International Conference on Learning Representations*.

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, Martin Popel, and Ondřej Bojar. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *arXiv:2007.03006*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

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

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa

Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning. `https://github.com/huggingface/trl`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

A CUNI-SFT Model Prompt Template

We have compared three ways of formatting the input. We present the corresponding prompts here.

Sentence-level:

```
Translate this {source_lang} sentence to
{target_lang}: {line}
```

Sentence-level with document context:

```
We need to translate one line from a
{source_lang} conversation into {target_lang}.
Source document: {document_src}
Already translated: {previous_translations}
Translate literally (no explanations) this
line: {line}
```

Document-level:

```
Translate from {source_lang} to {target_lang}:
{document}"
```