SkipCLM: Enhancing Crosslingual Alignment of Decoder Transformer Models via Contrastive Learning and Skip Connection

Nikita Sushko[1,2] Alexander Panchenko[2,1] Elena Tutubalina[1,3,4]
[1]AIRI [2]Skoltech [3]Sber AI [4]Kazan Federal University

Correspondence: sushko@airi.net

Abstract

This paper proposes **SkipCLM**, a novel method for improving multilingual machine translation in Decoder Transformers. We augment contrastive learning for cross-lingual alignment with a trainable skip connection to preserve information crucial for accurate target language generation. Experiments with XGLM-564M on the Flores-101 benchmark demonstrate improved performance, particularly for en-de and en-zh direction translations, compared to direct sequence-to-sequence training and existing contrastive learning methods. Code is available at: https://github.com/s-nlp/skipclm.

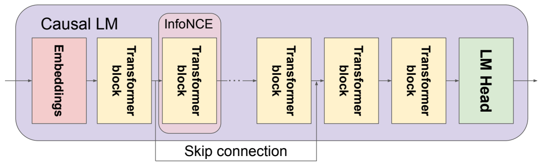

Figure 1: In **SkipCLM** we've added an InfoNCE to the final loss function to facilitate better cross-lingual alignment and a skip connection, to pass through information, which is potentially lost after training with InfoNCE.

1 Introduction

Recently, multilingual Decoder Transformer models (Vaswani et al., 2023), such as XGLM (Lin et al., 2022), Gemini (Georgiev et al., 2024), Unbabel Tower (Rei et al., 2024), Claude 3 Sonnet (Anthropic, 2024) became highly performant in the machine translation tasks (Kocmi et al., 2024). To better understand the mechanisms behind the emergence of this strong performance, researchers began to explore the inner workings of these models, which revealed a multi-stage evolution of internal representations within these Decoder Transformer models (Wendler et al., 2024; Li et al., 2024; Zhao et al., 2024). Initially, transformer (Vaswani et al., 2023) blocks project input token embeddings into a shared subspace. Subsequently, layers enrich the residual stream with different features, corresponding to token prediction, contextual information, and tasks represented in the prompts of the model (Ilharco et al., 2023). Finally, these enriched representations are mapped to output tokens (Wendler et al., 2024). Additionally, logit lens analysis indicates that tokens generated from layer activations in this second stage show a strong alignment with the dominant language in the model's training data (Wendler et al., 2024; nostalgebraist, 2020). However, this alignment is much less effective for underrepresented languages, negatively impacting prompt comprehension and task performance.

Existing techniques such as AFP (Li et al., 2024) and Lens (Zhao et al., 2024) address multilingual misalignment for low-resource languages by incorporating an auxiliary contrastive loss to improve the alignment of initial layer representations with the pivot language. While improving performance on tasks like translation, adding contrastive loss alone suffers from a potential loss of information within the residual stream, which hurts the model's performance in such aspects as original language preservation, context understanding, and instruction following. The authors of AFP added a separate instruction tuning stage to mitigate this information loss, but this greatly limited the applications of such models due to them being instruction tuned instead of utilized in a zero-shot manner.

This paper proposes **SkipCLM**, a novel method of enhancing cross-lingual alignment of multilingual embeddings in Decoder Transformer models. We introduce a linear skip connection to transfer hidden representations from the initial stages directly to the final transformer blocks. This, in conjunction with contrastive learning, facilitates both

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 517–528
April 30 - May 1, 2025

improved alignment of input embeddings with the pivot language and subsequent effective remapping to the original language, mitigating the information loss associated with only relying on contrastive learning.

2 Background and Related Work

In Sec. 2.1, we discuss the "Do Llamas Work in English" paper, which presented the interpretational framework, on which stems the idea of multilingual alignment. In Sec. 2.2, we discuss InfoNCE loss, which is essential for aligning the representations of parallel texts in several languages. In Sec. 2.3 and Sec. 2.4, we discuss pioneer works, which explored cross-lingual alignment using contrastive learning approaches.

2.1 Do Llamas Work in English

Wendler et al. (2024) investigate the latent representations within Decoder Transformer large language models (LLMs), focusing on the role of a potential internal "pivot" language. Their analysis reveals a three-stage process within the Decoder Transformer models. The early layers focus on processing the input information, and if we apply the logit lens nostalgebraist (2020) technique, we can see that hidden representations do not have any prevalence for a specific output language. In the middle layers, English emerges as the dominant language according to the language probability metric. This means that the model employs an internal latent representation closely aligned with the pivot language, which, in the case of the Llama-2 model, was English, being the most prevalent language in the training dataset. In the final layers, the most prevalent language becomes the target language.

The reliance on a pivot language during the intermediate stage can lead to information loss and suboptimal alignment for languages distant from the pivot. This misalignment reduces the model's ability to accurately capture nuances and context specific to the source language, impacting the translation quality.

2.2 InfoNCE

Van den Oord et al. (2019) introduced InfoNCE, a type of contrastive loss function used for self-supervised learning. It is used to train models to learn representations that are useful for predicting future samples in unsupervised learning tasks. Given a set of N random samples containing one positive sample from $p(x_{t+k}|c_t)$ and $N-1$ negative samples from a proposal distribution $p(x_{t+k})$, the InfoNCE loss is defined as:

$$\mathcal{L}_N = -\frac{E}{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right],$$

Where $f_k(x_{t+k}, c_t)$ is a function that estimates the density ratio between the conditional distribution and the proposal distribution. Optimizing this loss results in $f_k(x_{t+k}, c_t)$ estimating the density ratio $\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$. Minimizing the InfoNCE loss maximizes a lower bound on the mutual information between the context representation c_t and the future input x_{t+k}.

We utilize InfoNCE loss for aligning the embeddings between the translated versions of the input texts in the middle layers of our decoder LLM.

2.3 Lens

Zhao et al. (2024) propose Lens, a method for enhancing the multilingual capabilities of LLMs. Their approach leverages a decomposition of the multilingual latent subspace into language-agnostic and language-specific components via singular value decomposition. By identifying the components associated with each role, they employ contrastive learning to align the language-agnostic components across all languages. Simultaneously, they guide the language-specific components toward their respective language directions, increasing multilingual alignment. Finally, an $L2$ penalty is applied to maintain the integrity of the representations for a designated central language.

Experiments were conducted on English-centric decoder-only transformer models, such as Llama-3-8b (Grattafiori et al., 2024) and Phi-3.5-mini (Abdin et al., 2024), focusing on improving Chinese language performance. The authors did not provide evaluations for machine translation task, thus, we could not directly compare to their approach.

2.4 Align After Pre-Train

Li et al. (2024) introduce Align After Pre-training (AFP), a two-loss approach for cross-lingual adaptation of transformer models. The method leverages contrastive learning to spatially align the embeddings of translations of input examples for Decoder Transformer LLMs via InfoNCE loss. Additionally, authors incorporate cross-lingual instruction tuning, which explicitly instruct the models to generate responses in the target language. The final loss function for the models is a weighted combination of the contrastive loss and a cross-entropy loss. The models in the experiments are trained on a curated subset of the Bactrian-X dataset (Li et al., 2023), with machine translation performance assessed using BLEU score (Papineni et al., 2002) on the Flores-101 dev set (Goyal et al., 2021).

Since application of the contrastive loss to a certain layer of the model leads to some loss of information, which is represented in the hidden activations of the models, this approach is suboptimal. Our approach addresses this by adding a skip connection to preserve critical information from layers, that are earlier than the layer with contrastive loss, ensuring it is available for final token generation. In our paper, we directly compare our approach to AFP, using the same training and development data, the same metrics and the same model.

3 Methodology

3.1 Proposed Approach

This work proposes two key modifications to the Decoder Transformer architecture and training procedure:

1. **Incorporating InfoNCE Loss:** Following the approach of AFP (Lin et al., 2022), we integrate an InfoNCE loss function to enhance cross-lingual alignment between the pivot language (English) and other selected languages. This aims to improve the quality of multilingual representations and increase the translation abilities of the final model.

2. **Trainable Skip Connection:** We introduce a trainable skip connection, implemented as a linear layer within the Decoder Transformer. This connection is designed to selectively filter language-specific information using a linear layer with a ReLU activation function, preserving only the information relevant for subsequent translation to the target language. Applying the linear transformation with the activation function effectively creates a learnable non-linear filter, which removes unwanted noise from the residual connection from the start to the end of the model. This mitigates information loss during processing, improving the model's ability to reconstruct vital information otherwise lost in the standard architecture when contrastive loss is applied. The skip connection is placed immediately before the layer to which the contrastive loss is applied, ensuring critical information is preserved before potential loss within the contrastive layer. The architecture of the final model is shown in Fig. 1.

The skip connection is integrated back into the residual stream of the Decoder Transformer by multiplying the transformed skip connection output by a fraction of $\frac{1}{3}$ and adding the result to the model's hidden states. Specifically, the hidden state after layer α, denoted as R_α, is updated as follows:

$$R_\alpha = H_\alpha + \frac{\lambda}{3} \cdot \text{Skip}(H_\beta)$$

Where H_α is the layer, after which the skip connection is integrated into the residual stream, H_β represents the hidden state at the source layer of the skip connection β, $\text{Skip}(\cdot)$ denotes the linear transformation applied by the skip connection, and λ is a scaling coefficient.

During training, λ is gradually increased from 0 to 1 using a warm-up schedule; during inference, λ is set to 1. The choice of layers α and β is explored in Sec. 4.3. The selection of the normalizing constant $\frac{1}{3}$ was done empirically, with higher coefficients leading to model breakage.

3.2 Model Selection

For our experiments, we have used XGLM-564M (Lin et al., 2022) multilingual autoregressive LM. It was pretrained on a diverse corpus encompassing 30 languages, ranging from high-resource languages such as English, German, French, Chinese, and Russian to low-resource languages including Turkish, Vietnamese, Arabic, and Swahili.

3.3 Data

3.3.1 Training Data

Our models were trained on the Bactrian-X dataset (Li et al., 2023), a multilingual corpus comprising 3.4 million instruction-response pairs across 52 languages. This dataset leverages and expands upon the alpaca-52k (Taori et al., 2023) and Dolly-15k (Conover et al., 2023) datasets, with translation to all 52 languages performed using the Google Translate API. Responses in each language were generated using the GPT-3.5 model (Ouyang et al., 2022). To ensure comparability with prior work, data preparation followed the procedures outlined in the AFP repository[1]. Separate models were trained for Chinese, German, and Turkish, utilizing only the translated instruction-response pairs; no instruction tuning was performed on synthetic response data.

3.3.2 Test Data

Model evaluation was conducted using the development set of the Flores-101 benchmark (Goyal et al., 2021). We focused on the English-to-Chinese (`en-zh`), English-to-German (`en-de`), and English-to-Turkish (`en-tr`) translation directions. This selection reflects the language distribution within the training data of the XLMR-567M model, with German representing a high-resource European language, Chinese representing a high-resource non-European language, and Turkish representing a low-resource non-European language.

4 Experiments

4.1 Metrics

To evaluate our approach, we've used six different metrics: BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), chrF (Popović, 2015), BERTScore (Zhang et al., 2020), TER (Snover et al., 2006) and COMET (Rei et al., 2020). The primary metric for our evaluation we are using COMET, as it showed the best agreement with human labeling. More information on the metrics can be found in Appendix A.

4.2 Baselines

This work evaluates two baseline approaches: XGLM-564M trained directly on the parallel translation corpus (denoted as **Seq2Seq Training** in the Tab. 1); and a reproduction of the AFP method where skip connections were frozen and the hyperparameter λ, controlling the summation of hidden representations, was set to zero (denoted as **Align After Pretraining** in the Table 1). Additionally, we have included non-comprehensive evaluation from (Lin et al., 2022) to illustrate comparison between our and their approaches.

4.3 Hyperparameter Selection

Optimal values for the hyperparameters α and β were determined via grid search, with $\alpha \in [1, 3]$ and $\beta \in [15, 22]$. These ranges were selected based on the AFP paper's finding that the first layers are optimal for applying the contrastive loss. During

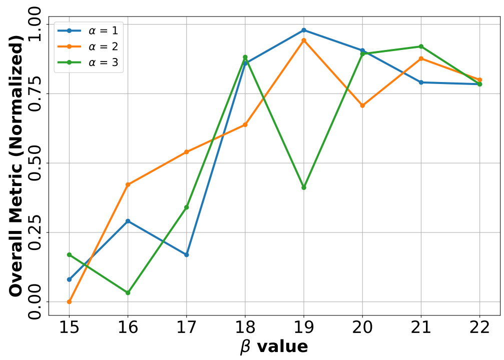

Figure 2: Grid search results for α and β hyperparameters for German language.

grid search, the models were trained on a subset of the German training data, comprising 3000 examples, and tested on a separate smaller development set, consisting of 100 examples from Flores-101. BLEU, METEOR, chrF, TER, F1 from BERTScore and COMET metrics were collected, normalized and averaged, to get one overall metric, which represents the final performance of the models. Since a lower score in the TER metric signifies better performance, we've inverted the values of this metric to maintain consistency with other metrics. The results of this grid search are presented in Fig. 2. The configuration $\alpha = 1, \beta = 19$ yielded the highest overall score and was thus selected for the final training phase. Additionally, it is shown that the $\beta = 19$ is a stable peak of the performance for all three evaluated α values, making this the optimal hyperparameter for training final models.

The λ hyperparameter for combining the output of skip connection with embeddings is initialized as 0 and then warmed up for 300 steps towards 1. This gradual warm-up prevents the model from being overwhelmed by a sudden influx of new information. A coefficient of 1e-2 was used to combine the loss functions, as it was empirically found to be the most stable across our experiments.

Model training was conducted on a single NVIDIA Tesla A100 80GB GPU. The models were trained for 1 epoch using a batch size of 16, a weight decay of 0.1, a cosine learning rate scheduler, and a learning rate of 5e-5. For consistency, the baseline models employed identical hyperparameter settings, with the contrastive loss applied to layer 1 for the AFP baseline.

[1] https://github.com/chongli17/cross-lingualalignment

Model	BLEU ↑	METEOR ↑	chrF ↑	BERTScore F1 ↑	TER ↓	COMET ↑
En-De						
SkipCLM (Ours)	**15.12**	**0.41**	**45.12**	**0.81**	**87.41**	**0.65**
Align After Pretraining	8.67	0.34	37.96	0.78	137.44	0.63
Seq2Seq Training	13.36	0.39	43.19	0.80	98.58	0.64
En-Tr						
SkipCLM (Ours)	8.61	0.30	37.29	0.78	98.00	0.66
Align After Pretraining	8.70	0.30	38.51	0.78	100.37	0.67
Seq2Seq Training	**9.78**	**0.31**	**38.82**	**0.79**	**90.65**	**0.68**
En-Zh						
AFP (Lin et al., 2022)	-	-	-	-	-	0.53
SkipCLM (Ours)	5.80	0.13	7.86	0.77	258.56	**0.57**
Align After Pretraining	6.00	0.13	8.05	0.77	291.58	0.54
Seq2Seq Training	**6.29**	**0.14**	**8.24**	**0.78**	**227.10**	0.56

Table 1: Evaluation results on the FLORES-101 dataset.

5 Results and Discussion

We have trained three models for each language: a model with applied skip connection and with contrastive loss (our approach), a model with only contrastive loss (AFP-like training) and a sequence-to-sequence trained model. Tab. 1 shows our results.

For English-German translation direction, our approach performs the strongest, achieving the highest scores in all metrics, including a notably lower TER compared to AFP baseline. Seq2Seq Training trails closely behind in this language pair. However, for English-Turkish, Seq2Seq Training shows best results, outperforming both our approach and AFP in every metric, including a higher BLEU score and lower TER. Our approach is slightly behind AFP in chrF, though COMET scores for all models are tightly grouped, suggesting similar perceived translation quality.

English-Chinese results are mixed. Seq2Seq Training leads in most metrics like BLEU and TER, but our approach achieves the highest COMET score, surpassing both Seq2Seq Training and AFP baseline. AFP baseline consistently underperforms, confirming our concerns, that simply adding a contrastive loss, as shown in AFP paper, leads to performance degradation, compared to the standard seq2seq training across all languages, underscoring the limitations of that approach. Interestingly, our implementation of the contrastive baseline surpasses the results reported in the AFP paper, likely due to improved hyperparameter tuning. Examples of translation being done by each model are shown in Appx. B.

We hypothesize, that the performance discrepancy between German, Chinese and Turkish can be explained by optimizing α and β hyperparameters for the German language, which shows the best results. Additionally, we believe that the performance of our method can be increased when training is being carried out on a multidirectional translation dataset instead of a single direction translation.

6 Conclusion

We present a novel method for enhancing multilingual machine translation in Decoder Transformers by augmenting contrastive learning with a trainable skip connection. This approach aimed to mitigate the information loss often associated with contrastive learning methods while simultaneously improving cross-lingual alignment with a pivot language. Our experiments on the Flores-101 benchmark, using XGLM-564M, demonstrated the effectiveness of this strategy, showing consistently better performance for German translation across all evaluation metrics, while being competitive for Chinese and slightly worse for Turkish languages.

7 Limitations and Future Work

This work has investigated the translational performance of the proposed method. However, its efficacy on tasks beyond sequence-to-sequence translation, such as multilingual understanding and generation, remains an open question. Future research could explore the application of the proposed algorithm to language model training. Furthermore, the investigation of multilingual training paradigms, with a combination of different training directions and the potential for cross-lingual transfer learning represents a promising future work direction.

Additionally, our approach is underperforming in the Turkish language, making necessary additional ablations and hyperparameter tuning for this language.

Ethics Statement

This work focuses on improving machine translation performance for multilingual decoder models. We primarily use publicly available datasets (Bactrian-X derived data, Flores-101) and pre-trained models (XGLM-564M). We acknowledge that language models can perpetuate societal biases present in their training data. The Bactrian-X dataset uses machine translation and AI-generated responses, which may introduce artifacts or reflect biases from those systems. Our method shows varying performance across language pairs, highlighting the need for careful evaluation, particularly for lower-resource languages. We release our code to encourage further research.

Acknowledgments

The contribution of E.T. was supported by the Kazan Federal University Strategic Academic Leadership Program ("PRIORITY-2030"), Strategic Project No 5.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Anthropic. 2024. Claude 3.5 sonnet model card. Accessed from Anthropic's Claude 3 Model Family documentation.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and Anmol Gulati et. al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Preprint*, arXiv:2106.03193.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, and Alan Schelten et. al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. *Preprint*, arXiv:2212.04089.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, and Anton et. al. Dvorkovich. 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.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. *Preprint*, arXiv:2311.08089.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *Preprint*, arXiv:2305.15011.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. 2024. Lens: Rethinking multilingual enhancement for large language models. *Preprint*, arXiv:2410.04407.

A Metrics Description

In our work, we've evaluated our models using the following six metrics:

- **BLEU (Papineni et al., 2002):** Measures how many n-grams in the generated text match the reference text. It focuses on precision and is commonly used for machine translation. Higher scores indicate better overlap, but it may not account for fluency or meaning.
- **METEOR (Lavie and Agarwal, 2007):** Evaluates translations by considering precision, recall, and alignment of words, including synonyms and stemming. It is more sensitive to word choice and meaning than BLEU, making it a useful complement.
- **chrF (Popović, 2015):** Based on character-level n-grams, this metric calculates an F-score that balances precision and recall. It is particularly effective for languages with complex morphology or tokenization challenges. For Chinese language, we've utilized Chinese tokenizer, used in SacreBLEU library (Post, 2018).

- **BERTScore** (Zhang et al., 2020): Uses contextual embeddings from BERT to compare semantic similarity between generated and reference text.
- **TER** (Snover et al., 2006): Measures the number of edits (insertions, deletions, substitutions and shifts) needed to turn the generated text into the reference text, normalized by length. Lower scores indicate better translations, as fewer edits imply higher similarity.
- **COMET** (Rei et al., 2020): A neural-based metric that evaluates translations by comparing their semantic meaning to reference translations and source inputs. It leverages pretrained LMs for a more nuanced understanding of quality.

B Translation Examples

Examples of the translations can be seen in the Table 2.

Model	Language Pair	English	Translation
SkipCLM (Ours)	de	English: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.	Übersetzen: "Wir haben jetzt 4-Monat-Lebewesen, die keine Diabetes-Pille sind, die früher diabetisch waren, und die jetzt nicht gesund sind."
		English: Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.	Übersetzen: Dr. Ehud Ur, Professor der Medizin an der Dalhousie University in Halifax, Nova Scotia und Vorsitzender der klinischen und wissenschaftlichen Division der amerikanischen Diabetes-Gemeinschaft warnte vor, dass die Forschung noch in ihrer frühen Phase ist.
		English: Danius said, "Right now we are doing nothing. I have called and sent emails to his closest collaborator and received very friendly replies. For now, that is certainly enough."	Übersetzen: Danius hat gesagt: "Hier ist gerade nichts. Ich habe mich an seinen nächstgelegenen Ansprechpartner gemeldet und sehr freundlichen Antworten erhalten. Bis jetzt ist das wirklich genug."
	zh	English: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.	翻译: “我们现在拥有4 个月的性小鼠,他们以前都是糖尿病患者,现在都不是,”他补充。
		English: USA Gymnastics and the USOC have the same goal — making the sport of gymnastics, and others, as safe as possible for athletes to follow their dreams in a safe, positive and empowered environment.	翻译: 美国柔术和美国奥运会都一样目标:让柔术运动成为运动,以及其他运动,安全、积极和充满力量。
		English: During the 1976 selections he advised Carter on foreign policy, then served as National Security Advisor (NSA) from 1977 to 1981, succeeding Henry Kissinger.	翻译: 1977 年选举期间,他建议卡特对外交政策,然后担任国家安全顾问(NSA) 直到1981 年。
	tr	Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.	Dr. Ehud Ur, doktor, Kanada'daki Dalhousie Üniversitesi'ndeki profesörlüğüyle ilgili ve Kanada Diyabet Derneği'nin klinik ve bilimsel bölümündeki çalışmaların son aşamalarında olduğunu kınandı.

Continued on next page

Model	Language Pair	English	Translation
		English: On Monday, Sara Danius, permanent secretary of the Nobel Committee for Literature at the Swedish Academy, publicly announced during a radio program on Sveriges Radio in Sweden the committee, unable to reach Bob Dylan directly about winning the 2016 Nobel Prize in Literature, had abandoned its efforts to reach him.	Swedish Akademi'de Nobel Edebiyat Ödülü'nü kazanan Sara Danius, Stockholm'deki Swedish Akademi'nin, Stockholm'de Bob Dylan'ın doğrudan ulaşamadığı 2016 Nobel Edebiyat Ödülü'nü kazanmak için çabaladığı için açıkladığı radyo programında, Danius'un bu konudaki çalışmaları sürdüremediği bildirildi.
		Danius said, "Right now we are doing nothing. I have called and sent emails to his closest collaborator and received very friendly replies. For now, that is certainly enough."	Danius, "Bugün hiçbir şey yapmadık. Arkadaşımla konuştuk ve çok dostça yanıt aldık. Bu, kesinlikle yeterli."
AFP	de	English: For example visiting castles in the Loire Valley, the Rhine valley or taking a cruise to interesting cites on the Danube or boating along the Erie Canal.	Übersetzen: In Bezug auf die Burgen im Loire-Tal, den Rhein-Tal oder die Fahrt mit dem Binnenschiff entlang des Danubiums.
		English: Many of the traditions surrounding the holiday have been adopted also by non-believers in Christian countries and non-Christians around the world.	English: Viele der Traditionen rund um das Fest der Osterfeiertage wurden auch von Nicht-Bürgern in christlichen Ländern und Nicht-Christen in den Weltgegenden adoptiert.
		English: Japanese work culture is more hierarchical and formal that what Westerners may be used to.	English: Japanese work culture is more hierarchical and formal than what Westerners might use to.
	zh	English: Due to the long distance from the continent mammals were unable to make the journey making the giant tortoise the primary grazing animal in the Galapagos.	翻译: 由于太平洋乳动物无法前往太平洋大陆,导致了乳动物的旅程难以进行。
		English: If you visit the Arctic or Antarctic areas in the winter you will experience the polar night, which means that the sun doesn't rise above the horizon.	翻译: 如果您在冬天访问北极或南极地区,您将感受到北极夜,意味着太阳没有从天空中升起。
		English: This offers a good opportunity to see the Aurora borealis, as the sky will be dark more or less around the clock.	翻译: 这将是一个很好的机会来观察日食星,当天天天空会更暗一些。

Continued on next page

Model	Language Pair	English	Translation
	tr	English: Christmas is one of the most important holidays of Christianity, and is celebrated as the birthday of Jesus.	İngilizce: Noel, Kilise'nin en önemli dini bayramlarından biridir ve İsa'nın doğum günü olarak kutlanır.
		English: There's a tradition to pass the Easter night awake at some exposed point to see the sunrise.	İngilizce: Bazı açık havada gün batımını görmek için bazı yerlerde uyku uykusuna ulaşmak için bir gelenek var.
		English: There are of course Christian theological explanations for this tradition, but it may well be a pre-Christian Spring and Fertility ritual.	Bu gelenek için de, o da bir Pre-Cristian Bahar ve Gebelik ritüel olabilir.
Seq2seq	de	English: Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have Type 1 diabetes.	Übersetzen: Als anderer Experten ist er skeptisch darüber, ob Diabetes geheilt werden kann, und bezieht sich dabei auf die Tatsache, dass diese Ergebnisse keine Relevanz haben, die Menschen mit Typ-1-Diabetes bereits haben.
		English: On Monday, Sara Danius, permanent secretary of the Nobel Committee for Literature at the Swedish Academy, publicly announced during a radio program on Sveriges Radio in Sweden the committee, unable to reach Bob Dylan directly about winning the 2016 Nobel Prize in Literature, had abandoned its efforts to reach him.	English: On Monday, Sara Danius, Permanent Secretary of the Nobel Committee for Literature at the Swedish Academy, publicly announced during a radio program on Sveriges Radio in Sweden the committee, unable to reach Bob Dylan directly about winning the 2016 Nobel Prize in Literature, had abandoned its efforts to reach him.
		English: He built a WiFi door bell, he said.	Übersetzung: Er hat einen WiFi-Schalter gebaut, sagte er.
	zh	English: Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.	翻译: 在丹佛大学亚尔福大学的艾滋病学教授埃尔·伊·奥利、教授、医学分院的主管和科学分管的加拿大糖尿病协会提醒说,研究仍在早期阶段。
		English: Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have Type 1 diabetes.	翻译: 他对糖尿病是否可以治愈持怀疑态度,并指出这些发现没有相关性,这些发现没有与已有1 型糖尿病患者相关。

Continued on next page

Model	Language Pair	English	Translation
		English: On Monday, Sara Danius, permanent secretary of the Nobel Committee for Literature at the Swedish Academy, publicly announced during a radio program on Sveriges Radio in Sweden the committee, unable to reach Bob Dylan directly about winning the 2016 Nobel Prize in Literature, had abandoned its efforts to reach him.	翻译: 在伦敦周日下午,萨拉·迪亚斯、瑞典斯坦福大学教授的永久秘书,在瑞典电视台在瑞典电视台播出的新闻节目中公开宣布,她无法直接向杰克逊·赖特直接联系,因为她无法直接向杰克逊·赖特直接联系。
	tr	English: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.	"Diyetisyen tarafından hipertansiyonlu olan 4 aylık kedilerimiz artık diyabetli değiller," ekledi.
		English: Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have Type 1 diabetes.	Diğer uzmanlar gibi diyabetin nasıl tedavi edilebileceğine dair şüphelidir, bu bulguların insanlarda Type 1 diyabet olup olmadığının hiçbir ilgisi olmadığını belirterek.
		English: Previously, Ring's CEO, Jamie Siminoff, remarked the company started when his doorbell wasn't audible from his shop in his garage.	"Ring CEO'su Jamie Siminoff, mağazasının kapısının sessiz olduğu sırada, şirketin başladığını söyledi."

Table 2: Selected translation examples by all models.