NYA's Offline Speech Translation System for IWSLT 2025

Wenxuan Wang, Yingxin Zhang, Yifan Jin, Binbin Du, Yuke Li
NetEase YiDun AI Lab, Hangzhou, China
{wangwenxuan,zhangyingxin03,jinyifan01,dubinbin,liyuke}@corp.netease.com

Abstract

This paper reports NYA's submissions to the IWSLT 2025 Offline Speech Translation (ST) task. The task includes three translation directions: English to Chinese, German, and Arabic. In detail, we adopt a cascaded speech translation architecture comprising automatic speech recognition (ASR) and machine translation (MT) components to participate in the unconstrained training track. For the ASR model, we use the Whisper medium model. For the neural machine translation (NMT) model, the wider and deeper Transformer is adopted as the backbone model. Building upon last year's work, we implement multiple techniques and strategies such as data augmentation, domain adaptation, and model ensemble to improve the translation quality of the NMT model. In addition, we adopt X-ALMA as the foundational LLM-based MT model, with domain-specific supervised fine-tuning applied to train and optimize our LLM-based MT model. Finally, by employing COMET-based Minimum Bayes Risk decoding to integrate and select translation candidates from both NMT and LLM-based MT systems, the translation quality of our ST system is significantly improved, and competitive results are obtained on the evaluation set.

1 Introduction

The Offline Speech Translation (ST) Task converts source audio into target text. Currently, two primary approaches dominate the ST field: the cascaded system and the end-to-end (E2E) system. The traditional cascade system (Matusov et al., 2005a) decouples the ST task into an automatic speech recognition (ASR) and a machine translation (MT) task. The source speech is first transcribed into text in the source language, which is then translated into text in the target language using a neural machine translation (NMT) model. However, it often leads to higher architectural complexity and error propagation (Duong et al., 2016), affecting subsequent MT tasks. To alleviate this problem, the end-to-end ST architecture (Bérard et al., 2016) is proposed. The E2E ST system employs a single neural network to directly map source-language audio to target-language text, bypassing intermediate symbolic representations. For end-to-end ST architectures, a key limitation is the scarcity of parallel speech-text data. In contrast, the widespread availability of large-scale ASR and MT datasets facilitates the development of high-precision ASR and MT systems through comprehensive training. Therefore, the cascaded ST system typically outperforms the E2E ST system (Anastasopoulos et al., 2022; Agarwal et al., 2023; Ahmad et al., 2024; Abdulmumin et al., 2025). Thus, we choose the cascaded ST scheme consisting of ASR and MT systems for the task.

The main architecture of the traditional NMT model is the encoder-decoder. Recently, large language models (LLMs) based on decoder-only architectures have demonstrated remarkable performance across various natural language processing (NLP) tasks. In the MT task, only the most advanced LLMs like GPT-4 (Achiam et al., 2023) can match the performance of supervised learning-based encoder-decoder state-of-the-art (SoTA) models such as NLLB (Costa-Jussà et al., 2022), yet their effectiveness still falls short of expectations in low-resource languages and specialized domains. Therefore, many studies (Xu et al., 2023, 2024b,a; Aryabumi et al., 2024) are focused on applying LLMs to smaller-scale models, broader language coverage, and more diverse application scenarios in machine translation, demonstrating significant advancements in the field. For example, X-ALMA (Xu et al., 2024a) is one of the top-performing translation models built on LLMs, capable of matching or even surpassing WMT winners and GPT-4 in some language pairs and scenarios. Therefore, unlike in previous work, we implement both NMT and LLM-based MT approaches

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 206–211
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

and investigate their combination to achieve improved translation performance.

We participate in the unconstrained training track of the offline speech translation task. And the Whisper (Radford et al., 2023) medium model is directly employed for the ASR system in the source language. We also explore audio segmentation methods, such as Supervised Hybrid Audio Segmentation (SHAS) (Tsiamas et al., 2022), to segment the source audio for better ST results. In the MT task, we widely collect a large amount of parallel data and monolingual data from various data sources. For the NMT system, we use the Transformer architecture (Vaswani et al., 2017) as the backbone model and implement multiple optimization techniques and strategies such as Back Translation (BT) (Sennrich et al., 2016), Forward Translation (FT), Domain Adaptation (DA), and Ensemble (Ganaie et al., 2022) to improve the translation quality of the NMT model. For the LLM-based MT system, we use X-ALMA as the foundational model and adopt supervised fine-tuning (SFT) to train and optimize the LLM-based MT model. Subsequently, we adopt Minimum Bayes Risk (MBR) (Kumar and Byrne, 2004) decoding to select the translation candidates from both NMT and LLM-based MT systems and obtain significant improvements in translation quality.

2 Dataset

2.1 Text Data

The training set is divided into two parts: general data and domain data. For general data, we retain the same data configuration of En2Zh and En2De as last year (Zhang et al., 2024). For En2Zh and En2De, we make full use of a large amount of monolingual data through BT and FT. For En2Ar, in addition to utilizing the data provided by IWSLT 2025, we incorporate several large-scale open-source text datasets such as NLLB (Costa-Jussà et al., 2022), CCAligned (El-Kishky et al., 2019), HPLT (Aulamo et al., 2023) and etc. For domain-specific data, we crawl a substantial amount of domain-specific videos from websites and use the bilingual subtitles provided by these sites to create domain-specific training sets.

We employ sBERT (Reimers and Gurevych, 2019, 2020) to calculate semantic similarity for all parallel text data and filter out text pairs with similarity scores lower than 0.7. Table 1 presents the size of our MT corpus after filtering.

Corpus	En2Zh	En2De	En2Ar
General data	27M	20M	126M
Domain data	4M	4M	236K

Table 1: Data statistics of MT corpus.

2.2 Data Pre-processing

For semantically filtered data, we perform text pre-processing according to last year's rules and procedure (Zhang et al., 2024) to enhance data quality.

After text pre-processing, these sentences are tokenized by a SentencePiece (SPM) model (Kudo and Richardson, 2018). The SPM model is trained separately on sampled data, with vocabulary sizes set as follows: 40k in English, 37k in Chinese, 37k in German, and 40k in Arabic. Both the source and target sides share the same dictionary.

3 Speech Translation System

3.1 ASR System

We utilize the Whisper [1] (Radford et al., 2023) model in conjunction with the SHAS [2] (Tsiamas et al., 2022) method to implement our ASR system within a cascaded framework.

SHAS functions as a Voice Activity Detection (VAD) mechanism within the ASR system, enabling the segmentation of lengthy audio files into shorter segments. We experiment with various parameters and ultimately settle on the parameter set of (5, 30, 0.5), which we apply across all scenarios except for the accent challenge data.

Whisper is an advanced multilingual ASR system, providing robust performance across various audio conditions, including accented speech and noisy environments. The open-source models range from tiny to large, addressing different computational needs. We choose the medium-sized Whisper model for its suitability as the ASR model in our speech translation system.

3.2 MT System

Due to differences in training paradigms and learning objectives, traditional NMT tends to produce more literal translations while LLM-based MT generates more paraphrased outputs. The LLM approach shows better fluency and greater robustness to ASR errors, though it may occasionally overlook details or produce redundant hallucinations. These

[1] https://github.com/openai/whisper
[2] https://github.com/mt-upc/SHAS

two approaches exhibit complementary strengths in machine translation. Therefore, both NMT and LLM-based MT approaches are developed and integrated for our machine translation system.

3.2.1 NMT Model

Our NMT model in the speech translation system is built using the Transformer architecture implemented with the Fairseq toolkit (Ott et al., 2019). This model is designed with a wider and deeper structure, including an 18-layer encoder, 6-layer decoder, and 16 self-attention heads. This architecture enables the model to capture complex patterns and dependencies in the data effectively. Our NMT model is trained on parallel data from three language directions (English to Chinese, German, and Arabic) to form a one-to-many translation model.

Data augmentation techniques like back translation (Sennrich et al., 2016) and forward translation are employed to enhance the quality and diversity of the training data. Back translation involves translating the target language back into the source language, while forward translation transforms the source language into the target language. These methods leverage additional monolingual resources to generate synthetic bilingual data. In total, we utilize approximately 23M sentences of BT and FT data, including 18M sentences of En2Zh data and 5M sentences of En2De data. When employing the data generated by BT or FT models, we adopt the tagged BT method (Caswell et al., 2019) by appending a distinctive <BT> token at the beginning of the source sentence. This approach enables the model to distinguish between supervised and semi-supervised data during the training process.

Domain adaptation is also performed to fine-tune the model for specific domains. In-domain data is selected and used to train monolingual language models, which then score all language pairs. Specific thresholds are set to filter parallel data that is closer to the target domain. This process ensures that the model performs well in domain-specific scenarios, enhancing its overall translation quality and adaptability to different contexts.

3.2.2 LLM-based MT

LLMs have demonstrated impressive performance across various NLP tasks. Since most LLMs are primarily pre-trained on English, they still face limitations in multilingual translation tasks. Consequently, the paradigm of applying LLMs to multilingual translation tasks has been extensively studied. Among these, X-ALMA currently represents the state-of-the-art in open-source multilingual machine translation models. It supports bidirectional translation between English and 49 languages, achieving SoTA performance on the COMET-22 metric across all 50 language directions.

In this task, we find that the release of the X-ALMA[3] open-source model already achieves competent translation quality. Building upon the baseline, we perform supervised fine-tuning to enhance its domain-specific capabilities. In order to ensure data quality, we filter in-domain parallel data based on the reference-free CometKiwi (Rei et al., 2022b) metric. Subsequently, we conduct parameter-efficient adaptation of the model through Low-Rank Adaptation (LoRA) (Hu et al., 2022) fine-tuning, which is applied to all modules of the feed-forward network.

3.2.3 Minimum Bayes Risk Decoding

Unlike Maximum-A-Posteriori (MAP) estimation, which selects the single most probable hypothesis, Minimum Bayes Risk (MBR) (Kumar and Byrne, 2004) considers the entire distribution of possible outcomes and chooses the decision that minimizes the average loss across them. For MT, MBR decoding employs evaluation metrics like COMET (Rei et al., 2022a) to choose the hypothesis with the highest average score against other candidates. A substantial body of research (Fernandes et al., 2022; Finkelstein et al., 2023) has demonstrated that MBR decoding can effectively enhance translation quality across both NMT and LLM-based MT models. The N-best candidates from the NMT model are produced via beam search, while those from the LLM-based MT model are generated through temperature scaling and nucleus sampling. We employ COMET-based MBR decoding to rerank all the translation candidates from both subsystems, ultimately selecting the final translation output.

4 Experiments and Results

All NMT models are implemented using the open-source Fairseq toolkit (Ott et al., 2019). For LLM fine-tuning, we utilize the open-source ALMA toolkit [4] (Xu et al., 2024a). We evaluate the performance of MT models using case-sensitive Sacre-

[3] https://huggingface.co/haoranxu/X-ALMA
[4] https://github.com/fe1ixxu/ALMA

Model	En2Zh		En2De	
	COMET	BLEU	COMET	BLEU
NMT baseline	0.7988	34.70	0.7081	25.23
+ BT & FT	0.7995	35.03	0.7248	**26.36**
+ DA	0.8181	35.21	0.7315	26.34
+ MBR	0.8342	35.53	0.7572	25.41
LLM baseline (X-ALMA)	0.8200	32.74	0.7437	25.44
+ SFT	0.8221	33.44	-	-
+ MBR	0.8337	33.58	0.7560	24.77
MBR Ensemble NMT&LLM	**0.8417**	**36.01**	**0.7708**	25.71

Table 2: COMET and BLEU scores of NMT and LLM-based MT systems on the IWSLT tst-2022 test set

BLEU[5] (Post, 2018) and COMET[6] (Rei et al., 2022a) metrics, based on the tst2022 and tst2010 test sets. Specifically, tst2022 is used to assess En2De and En2Zh, while tst2010 is applied for En2Ar. For audio segmentation, we adopt SHAS with parameters set to (5,30,0.5). Finally, we utilize mwerSegmenter [7] (Matusov et al., 2005b) toolkit for the resegmentation and alignment of translation results.

Table 2 presents the COMET and BLEU scores for various NMT and LLM systems on the tst2022 test set. For NMT models, the integration of BT&FT data and domain adaptation demonstrates a notable enhancement of nearly 2% COMET scores across both En2Zh and En2De. This highlights the importance of domain-specific data for model performance. For LLM-based MT models, we perform LoRA fine-tuning on the X-ALMA pre-trained model with in-domain parallel data filtered by COMET-Kiwi (threshold is 0.82) for En2Zh, which brings slight translation improvements. The COMET-based MBR decoding achieves significant improvements in COMET scores, whether applied to candidate selection for a single translation model or two different types of translation systems (NMT and LLM). It is noteworthy that the En2De results of the single system indicate an inverse relationship between the COMET and BLEU scores.

Table 3 presents the performance of our final submitted ST system in the unconstrained training track of the offline speech translation task. Based on the results of "MBR Ensemble NMT&LLM" in Table 2, we train multiple models using a similar approach and achieve further improvements in COMET scores by integrating them through MBR decoding. The COMET scores for En2Zh and En2De reach 84.54% and 77.36% on tst2022, respectively. Since the En2Ar track does not provide an in-domain development set, we present the performance of En2Ar on the out-of-domain set tst2010 for reference.

	Test set	COMET	BLEU
En2Zh	*tst-2022*	0.8454	35.89
En2De	*tst-2022*	0.7736	25.81
En2Ar	*tst-2010*	0.8689	23.85

Table 3: COMET and BLEU scores of the ST system on the IWSLT test sets

5 Conclusion

This paper presents our submission to the IWSLT 2025 offline speech translation task. For the unconstrained track, we adopt a cascaded speech translation architecture consisting of the ASR and MT systems. For the ASR system, we directly employ the open-source Whisper medium model, which has shown outstanding performance and strong robustness across various scenarios for English speech recognition tasks. For the MT system, we investigate both NMT-based and LLM-based approaches and explore optimization strategies including data augmentation, domain adaptation, MBR decoding, and model ensemble. Experimental results demonstrate that integrating NMT with LLM-based MT models while applying these techniques yields significant performance improvements. Our final system achieves COMET scores of 0.8454, 0.7736, and 0.8689 for EN→ZH, EN→DE on the IWSLT tst-2022 test set, and EN→AR on the tst-2010 test set, respectively.

[5] https://github.com/mjpost/sacrebleu

[6] https://github.com/Unbabel/COMET

[7] https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria (in-person and online). Association for Computational Linguistics. To appear.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, et al. 2023. Findings of the iwslt 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61.

Ibrahim Sa'id Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, et al. 2024. Findings of the iwslt 2024 evaluation campaign. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, et al. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer Van Der Linde, and Jaume Zaragoza. 2023. Hplt: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 517–518.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2019. Ccaligned: A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*.

Patrick Fernandes, António Farinhas, Ricardo Rei, José GC de Souza, Perez Ogayo, Graham Neubig, and André FT Martins. 2022. Quality-aware decoding for neural machine translation. *arXiv preprint arXiv:2205.00978*.

Mara Finkelstein, Subhajit Naskar, Mehdi Mirzazadeh, Apurva Shah, and Markus Freitag. 2023. Mbr and qe finetuning: Training-time distillation of the best and most expensive decoding methods. *arXiv preprint arXiv:2309.10966*.

Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.

E. Matusov, S. Kanthak, and Hermann Ney. 2005a. On the integration of speech recognition and statistical machine translation. In *Proc. Interspeech 2005*, pages 3177–3180.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005b. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *arXiv preprint arXiv:2209.06243*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, pages 106–110.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024a. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

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. *arXiv preprint arXiv:2401.08417*.

Yingxin Zhang, Guodong Ma, and Binbin Du. 2024. The nya's offline speech translation system for iwslt 2024. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 39–45.