CUNI-NL@IWSLT 2025: End-to-end Offline Speech Translation and Instruction Following with LLMs

Nam Luu and **Ondřej Bojar**
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
namhoang.luu700@student.cuni.cz, bojar@ufal.mff.cuni.cz

Abstract

This paper describes the CUNI-NL team's submission to the IWSLT 2025 Offline Speech Translation and Instruction Following tasks, focusing on transcribing the English audio and translating the English audio to German text. Our systems follow the end-to-end approach, where each system consists of a pretrained, frozen speech encoder, along with a medium-sized large language model fine-tuned with LoRA on three tasks: 1) transcribing the English audio; 2) directly translating the English audio to German text; and 3) a combination of the above two tasks, i.e., simultaneously transcribing the English audio and translating the English audio to German text.

1 Introduction

End-to-end speech translation (ST) is a growing research direction that aims to ignore the intermediate speech recognition (ASR) step to directly translate the audio input into the corresponding text in another language. This approach simplifies the overall architecture and has been shown to match the performance of the cascaded counterpart (Bérard et al., 2018; Liu et al., 2019; Gaido et al., 2020).

Large language models (LLMs) have demonstrated their good performance in a large number of complex natural language tasks, including machine translation (Minaee et al., 2024; Zhang et al., 2024; Zhao et al., 2023; Naveed et al., 2024). With the ever-improving potential of LLMs, researchers have been trying to integrate different components used for other modalities, in order to extend their abilities to go beyond text-only tasks (Li et al., 2023a; Gao et al., 2023; Liu et al., 2023; Li et al., 2023c; Zhang et al., 2023).

Motivated by recent contributions in speech representation learning and LLMs, to participate in the IWSLT 2025 Offline Speech Translation and Instruction Following tasks, we aim to investigate an end-to-end architecture that can perform both ASR and ST. This architecture combines the high-quality audio representation from the pre-trained acoustic models with the excellent performance of LLMs to serve as an end-to-end speech translation system, while still having the ability to transcribe from the audio signal. Our systems, after being fine-tuned with the Low-Rank Adaptation (LoRA; Hu et al., 2021) technique, achieve a solid performance in both speech recognition and translation.

The paper is structured as follows:

- Section 2 describes the details of our chosen network architecture, along with the dataset used for its training and evaluation.
- Section 3 provides the ASR and ST evaluation results of the model in different public test sets.
- Section 4 proposes possible directions to improve the architecture.

2 Methods and Dataset

2.1 Architecture

The overall architecture is illustrated in Figure 1. For each training sample, given the speech signal, its corresponding transcript, and the translated text, the speech hidden features are obtained using a speech encoder. In this step, we experimented with SeamlessM4T (Barrault et al., 2025) and Whisper encoder (Radford et al., 2022).

Next, the speech features represented as a time sequence of vectors, at the "frame rate" of 20ms, are fed to a projection layer, in order to convert the feature dimension to match the LLM's embedding dimension. For a better match between the speech encoder and the LLM, we use a length adapter which effectively reduces the "frame rate" of the sequence. The resulting speech embeddings are subsequently given to the LLM as the prompt for it to generate the corresponding transcription and the translated text simultaneously. The LLM is fine-tuned in the next-token-prediction fashion to

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 282–288

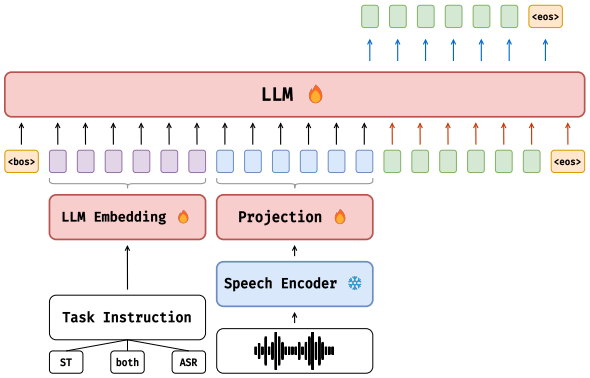

Figure 1: The overall architecture includes a frozen speech encoder component, a modal projection layer, and a fine-tuned LLM. **Red** arrows denote the usage of tokens during training, and **blue** arrows indicate tokens generated during inference; while **black** arrows represent the prompt fed to the LLM. The original modality of the model is indicated by the token color, from left to right: violet for text instruction, blue for source-language speech tokens, green for target-language text tokens. Length adapter is a part of the Projection step.

complete the sequence with the translation into the target language.

2.2 Speech Encoder

We adopted SeamlessM4T (Barrault et al., 2025) and Whisper (Radford et al., 2022) as the speech encoders, utilizing their capability of extracting high-quality representation from audio data. From both architectures, we only employed the encoder part of the pre-trained `seamless-m4t-v2-large`[1] and `whisper-large-v3`,[2] respectively, in order to extract the audio hidden features.

2.3 Length Adapter

The length of the speech feature sequence can be longer than the supported length of the LLM, as a result, it is more favorable to shorten it beforehand.

Because the speech encoder part of the SeamlessM4T architecture already contains a length adapter layer (Figure 2), we decided not to add any adapter layer, but used it as-is for SeamlessM4T-based models.

For Whisper-based models, a convolution-based downsampling layer with a kernel size of 25x5 and a stride of 25 is used to reduce the length of the

[1] https://huggingface.co/facebook/seamless-m4t-v2-large

[2] https://huggingface.co/openai/whisper-large-v3

[3] https://github.com/facebookresearch/seamless_communication/blob/main/docs/m4t/README.md

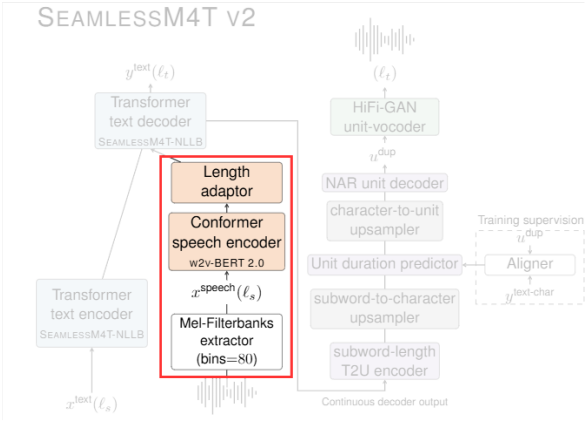

Figure 2: Details of the SeamlessM4T-v2 architecture,[3] where the red region is the speech encoder part

speech feature sequence from 1500 tokens (each corresponds to 20ms) to 60 tokens (each encodes 500ms). The detail of the convolution adapter layer is illustrated in Figure 3.

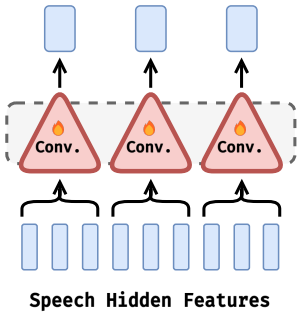

Figure 3: Details of the convolution adapter. Note that the convolution windows do not overlap.

2.4 Projection Layer

For the projection layer, we used only one simple three-layer ReLU-activated feed-forward network, with the hidden size of 4096, to map from the encoder's hidden size to the corresponding LLM's hidden size. This layer ensures the resulting speech representation is well integrated into the LLM's embedding space, giving it enough information for the downstream task.

2.5 LLMs

We experimented with three different pre-trained LLMs available on HuggingFace, namely `Llama-3.1-8B-Instruct`,[4]

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

EuroLLM-9B-Instruct,[5] and gemma-3-12b-it.[6] We summarize the examined combinations of components in Table 1.

Speech Enc.	LLM	Adapter
seamless-m4t -v2-large (S)	Llama-3.1-8B-Instruct (L) EuroLLM-9B-Instruct (E) gemma-3-12b-it (G)	N/A
whisper -large-v3 (W)	Llama-3.1-8B-Instruct (L) EuroLLM-9B-Instruct (E) gemma-3-12b-it (G)	25x5 Convolution

Table 1: Details of our six examined combinations of components, testing each of the speech encoders ((S), (W)) with each of the LLMs ((L), (E), (G)).

2.6 Dataset

All models were trained using the CoVoST2 dataset (Wang et al., 2020), a large multilingual corpus built from the Common Voice corpora (Ardila et al., 2020), which contains the audio data, the English transcription of such audio and the translation of the transcription into multiple languages. Specifically, we used the English-to-German subset of the dataset, with approximately 184 hours of audio data.

For evaluation, we used the test sets from the Offline Speech Translation track of IWSLT 2021[7] and 2022,[8] because they are the two latest development sets whose golden labels are available. These datasets are from the TED domain, in which the audios contain clean speech from the speaker mixed with some occasional noise from the audience; thus, we believe these are suitable for development. As all models can perform both ASR and ST, evaluation results for both tasks are described in Sections 3.2 and 3.3, respectively.

2.7 Multi-task Training

To obtain a system that can perform both ASR and ST tasks, we decided to train the model on the following three tasks:

1. Transcribing the English audio to English text;
2. Directly translating the English audio to German text; and
3. Simultaneously transcribing the English audio to English text, and translating such audio to German text.

With tasks (1) and (3), the LLM is given corresponding instructions depending on each task. For task (2), we used two different prompts in either English or German, to prepare the model for processing both English and German instructions. With this setup, we randomly divided the training dataset into four subsets using a uniform distribution, where each part was associated with an instruction according to the task and the relevant language. We decided to split the dataset and train the model on only one epoch, instead of duplicating the dataset four times and training for four epochs, due to limited time and resources. Details about each task and the corresponding instructions are described in Table 2.

Task	Instruction	# examples
transcribe (1)	Transcribe the English audio	72,067
translate_en (2)	Translate the English audio to German	72,600
translate_de (2)	Übersetzen Sie den englischen Ton ins Deutsche	72,455
both (3)	Transcribe then translate the following English audio to German	72,291

Table 2: Details of our four tasks, each demonstrated by roughly a quarter of the fine-tuning items

2.8 Training and Inference Details

All systems were fine-tuned using 16-bit LoRA (Hu et al., 2021) adapters in bfloat16 precision, with the following LoRA parameters: rank of $r = 256$, alpha of $\alpha = 256$. The effective batch size was set to 8. Other training hyperparameters included the learning rate of $1e-5$ with 100 warmup steps, and an AdamW optimizer (Loshchilov and Hutter, 2019) with a cosine scheduler (Loshchilov and Hutter, 2017). All systems were trained for 1 epoch.

For each example, the training data is formatted as follows: "<bos> {user_header} {instruction} {audio_features} {assistant_header} {output} <eos>". The cross-entropy loss was computed only for the tokens following "{assistant_header}". Each system's training loss details are illustrated in Figure 4.

During inference, for each audio data, the LLMs were prompted using the following format: "<bos> {user_header} {instruction} {audio_features} {assistant_header}", then generated the output, subject to the task, in an autoregressive manner. We performed inference using the beam search algorithm, with a beam size of 2

[5] https://huggingface.co/utter-project/EuroLLM-9B-Instruct
[6] https://huggingface.co/google/gemma-3-12b-it
[7] https://iwslt.org/2021/offline
[8] https://iwslt.org/2022/offline

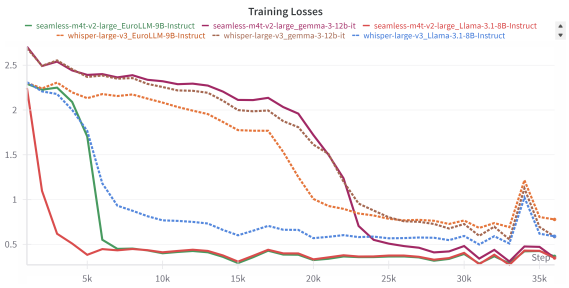

Figure 4: Training loss of systems

for all systems. All evaluation results are described in Sections 3.2 and 3.3.

3 Evaluation

3.1 Metrics and Tools

For the Offline Speech Translation task, we evaluated all models using standard metrics, namely BLEU (Papineni et al., 2002), $\text{COMET}_{22}^{\text{DA}}$ (Rei et al., 2022a),[9] and $\text{COMET}_{22}^{\text{KIWI-DA}}$ (Rei et al., 2022b).[10] For the Automatic Speech Recognition task, we used WER, the standard metric for speech recognition.

For the evaluation purpose, we used the `SLTev` (Ansari et al., 2021) library,[11] because it supports both MT and ASR evaluation in one package, using `sacreBLEU` (Post, 2018) to calculate BLEU score. However, since `SLTev` does not report any COMET-family metrics, we had to change the structure of the sentence with `mwerSegmenter`,[12] to automatically resegment the models' output according to the reference, before evaluating with the `unbabel-comet`[13] package. The evaluation was done using `python-3.11.5`, `SLTev-1.2.3`, and `unbabel-comet-2.2.2`.

3.2 ASR Results

Table 3 details the ASR evaluation results against the IWSLT 2022 test set (tst2022). We reported the WER score after applying the "LPW" preprocessing strategy available in `SLTev`, which first lowercased every character, removed all punctuation, then used the built-in `mwerSegmenter` tool to resegment the output transcripts. Due to some bugs when processing the IWSLT 2021 test set (tst2021), `mwerSegmenter` failed to run during evaluation, hence we could not obtain the results. It can be seen that the model with `seamless-m4t-v2-large` as the speech encoder and `EuroLLM-9B-Instruct` as the decoder has the best result among all systems.

3.3 Offline ST Results

Tables 4 and 5 report the BLEU and COMET-family scores, respectively, on the two test sets, with two corresponding instructions. For evaluating with BLEU, we included both `docAsWhole` score, which concatenated all reference segments and candidate complete segments as two documents, and `mwerSegmenter` score, which resegments complete candidate segments according to reference segments to minimize WER. Similar to Section 3.2, `mwerSegmenter` scores for IWSLT 2021 test set could not be obtained, hence we did not include them.

We observe that the system with `seamless-m4t-v2-large` as the encoder and `EuroLLM-9B-Instruct` as the language model achieves the best scores in all evaluation metrics, compared to the other systems. With the instruction associated with the task "`both`" (Table 2), the system excels in translation results, suggesting that the inclusion of English transcript provided useful assistance in translation.

Comparing between the two prompt variations "`translate_en`" and "`translate_de`" for this task, the latter one leads to more solid overall results. For example, consider the Ⓢ+Ⓔ system: for tst2022, while "`translate_en`" instruction might outperform that of "`translate_de`", but the difference is small; while results for tst2021 shows a contrastive situation, where "`translate_de`" surpasses "`translate_en`' by a considerable amount. This behavior also appears in other systems, leading us to believe that the system can perform better when the instruction provided is in the relevant target language. As a result, we chose "`translate_de`" prompt with Ⓢ and Ⓔ as our submission to the Offline Speech Translation and Instruction Following task, under the "*constrained+LLM*" evaluation condition.

4 Future Work

For the IWSLT 2025 Offline Speech Translation and Instruction Following tasks, we have only conducted experiments for the English-to-German di-

[9] https://huggingface.co/Unbabel/wmt22-comet-da

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

[11] https://github.com/ELITR/SLTev

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

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

Model		transcribe	both
Enc.	**LLM**	**tst2022**	**tst2022**
Ⓢ	Ⓛ	14.1%	17.3%
	Ⓔ	**13.4%**	16.7%
	Ⓖ	20.0%	24.2%
Ⓦ	Ⓛ	24.3%	26.6%
	Ⓔ	47.9%	47.5%
	Ⓖ	38.6%	38.5%

Table 3: ASR evaluation results (WER↓)

Model		translate_en		translate_de		both	
Enc.	**LLM**	**tst2021**	**tst2022**	**tst2021**	**tst2022**	**tst2021**	**tst2022**
Ⓢ	Ⓛ	39.53 / -	37.55 / 26.58	39.50 / -	37.58 / 26.66	42.01 / -	38.15 / 29.78
	Ⓔ	41.50 / -	**38.47 / 30.65**	**41.94** / -	37.73 / 29.83	44.28 / -	40.82 / 32.33
	Ⓖ	37.37 / -	33.72 / 24.93	36.70 / -	33.66 / 25.25	42.17 / -	37.70 / 29.73
Ⓦ	Ⓛ	33.02 / -	31.54 / 19.47	33.64 / -	30.02 / 19.62	39.48 / -	37.76 / 26.64
	Ⓔ	22.43 / -	22.02 / 9.24	22.21 / -	23.32 / 9.83	32.91 / -	31.50 / 19.56
	Ⓖ	27.23 / -	27.34 / 14.67	27.34 / -	27.81 / 14.67	35.37 / -	34.72 / 22.95

Table 4: Offline ST en2de BLEU results, with both `docAsWhole`↑ and `mwerSegmenter`↑ scores, respectively

Model		translate_en		translate_de		both	
Enc.	**LLM**	**tst2021**	**tst2022**	**tst2021**	**tst2022**	**tst2021**	**tst2022**
Ⓢ	Ⓛ	61.11 / 53.85	68.05 / 62.38	68.69 / 62.63	67.88 / 62.02	69.49 / 64.66	69.48 / 64.80
	Ⓔ	62.57 / 56.13	**71.06 / 66.04**	**70.38 / 65.11**	70.62 / 65.46	71.59 / 66.93	71.05 / 66.53
	Ⓖ	59.09 / 51.87	66.01 / 60.33	66.33 / 60.48	66.08 / 60.10	67.90 / 62.75	68.20 / 63.28
Ⓦ	Ⓛ	52.80 / 43.13	61.99 / 53.47	62.73 / 54.55	62.02 / 53.73	66.66 / 60.71	66.38 / 60.00
	Ⓔ	42.73 / 30.45	48.33 / 35.93	48.22 / 35.80	49.48 / 36.95	58.35 / 47.78	56.85 / 46.70
	Ⓖ	48.10 / 36.40	55.36 / 44.11	55.75 / 44.10	54.10 / 42.70	61.49 / 52.65	61.54 / 52.86

Table 5: Offline ST en2de $\text{COMET}_{22}^{\text{DA}}$↑ and $\text{COMET}_{22}^{\text{KIWI-DA}}$↑ results, respectively

rection; hence, in the future, we will expand our experiments to more language pairs and directions. In addition, we have some ideas to improve the pipeline:

- Try other modal adapter methods, like Q-Former (Li et al., 2023b).
- Experiment with smaller variants of the LLMs for faster training and inference, while retaining the quality in translation, by distilling knowledge from fine-tuned systems.
- Build a Direct Preference Optimization (DPO; Rafailov et al., 2024) or Contrastive Preference Optimization (CPO; Xu et al., 2024) dataset to apply into the training pipeline. Xu et al. (2024) showed that their CPO approach improved the performance of medium-sized LLMs, so we will try following the same idea.

5 Conclusion

In this paper, we leveraged pre-trained speech encoders and LLMs and connected them into an end-to-end architecture to participate in the IWSLT 2025 Offline Speech Translation and Instruction Following tasks. Our primary goal was to develop a system that could perform both ASR for English audio, and ST from English audio to German text. In our experiments, the model with `seamless-m4t-v2-large` as the speech encoder and `EuroLLM-9B-Instruct` as the LLM yielded the best results in evaluation of both ASR and ST tasks, suggesting that this pair could be a promising combination for end-to-end models.

6 Acknowledgment

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Nam Luu has been supported by the Erasmus Mundus program in Language and Communication Technologies (LCT).

Ondřej Bojar has received funding from the Project OP JAK Mezisektorová spolupráce Nr. CZ.02.01.01/00/23_020/0008518 named "Jazykověda, umělá inteligence a jazykové a řečové

technologie: od výzkumu k aplikacím."

References

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTEV: Comprehensive Evaluation of Spoken Language Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Çelebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-End Automatic Speech Translation of Audiobooks.

Marco Gaido, Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-End Speech-Translation with Knowledge Distillation: FBK@IWSLT2020.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023c. Prompting Large Language Models for Zero-Shot Domain Adaptation in Speech Recognition.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. ArXiv:1711.05101 [cs, math].

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A Comprehensive Overview of Large Language Models.

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.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. 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, Abu Dhabi, United Arab Emirates (Hybrid). 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 M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task.

Changhan Wang, Anne Wu, and Juan Pino. 2020. CoVoST 2 and Massively Multilingual Speech-to-Text Translation.

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.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction Tuning for Large Language Models: A Survey.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models.