Hurdles of Automatic Metric for Speech Translation Evaluation

Victor Zarzu
ETH Zurich
vzarzu@ethz.ch

Vilém Zouhar
ETH Zurich
vzouhar@ethz.ch

Abstract

Automatic evaluation of speech translation has so far relied on text-only automated metrics that ignore speech phenomena. One would expect that incorporating the source audio modality would improve the performance of automatic metrics. We implement two standard metric paradigms: a COMET-audio regression model using audio and text encoders, and one based on prompting a speech large language model. Surprisingly, both audio-infused models fail to reliably surpass text-only baselines. We attribute this failure to the noise pollution and audio-transcript mismatches present in the audio signal, which makes the modality unreliable from the metric's perspective. Furthermore, we argue that current human-annotated evaluation datasets for automated metrics predominantly feature technical content or short texts where paralinguistic features like prosody lack importance, rendering the extra audio information unhelpful for quality estimation (QE).

1 Introduction

Speech-to-text translation, due to its complexity, has often stayed behind text-to-text translation. In order to be able to continue improving speech translation models, we need to be able to evaluate them. The goal of automatic metrics (both reference-based and reference-free) is to predict an assessment of a translation, such as on a scale from 0% to 100% (Lavie et al., 2025). Evaluation of speech-to-text translations usually falls back to textual automated metrics, which miss the mark by omitting speech-only phenomena, such as prosody or hesitations (Abdulmumin et al., 2025).

Surprisingly, we find that common approaches to automatic metrics that infuse audio inputs (LLM-based and COMET+audio, illustrated in Figure 1) do not surpass text-only automated metrics. We attribute this to the lack of importance

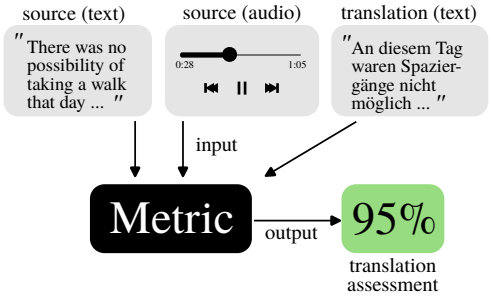

Figure 1: Example of a reference-free automated metric. The input is the translation text and either the source text, source audio, or both.

of the extra information that resides in the audio (e.g. intonation, prosody, emphasis, tempo, emotion etc.) in the case of the evaluation dataset, the samples mostly containing technical information or short sentences. Moreover, the noise pollution and unrelated words in the audio along with mismatches between the audio and transcripts contribute to underperformance of adding the audio modality.

We focus on reference-free automated metrics, also known as quality estimation, since the primary concern is with the input audio modality.

This work is split between two parts: Section 3 outlines our two approaches for automated metrics and evaluates them on IWSLT 2026 metrics shared task (Adelani et al., 2026) development set, for which we describe our submission. Section 4 presents evidence suggesting why extra audio input does not improve performance of automated metrics.

2 Related Work

Learned metrics. COMET (Rei et al., 2020) is a neural framework for machine translation evaluation that trains regression models on top of cross-lingual pretrained language models. Given the source text, the machine translation output, and optionally a human reference, COMET pro-

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 305–315

duces a scalar quality score trained to correlate with human judgments. Its reference-free variant, CometKiwi (Rei et al., 2022), operates in a quality estimation setting by removing the dependency on a human reference, using InfoXLM (Chi et al., 2021) as its backbone encoder. Our approach "COMET+audio" draws direct inspiration from this architecture, extending it to incorporate the audio modality.

Another approach is to make use of instruction-following abilities of LLMs. GEMBA (Kocmi and Federmann, 2023) is simply a zeroshot prompted LLM used directly for translation quality assessment, which achieves state-of-the-art performance as an automated metric (Lavie et al., 2025). A key finding of the GEMBA's authors was that the approach only works reliably with GPT-3.5 and larger models, with smaller models failing to produce valid scores. Our second approach "Speech LLM" follows a similar philosophy but adapts it to a multimodal speech-instruction-tuned LLM and uses few-shot examples instead of zero-shot prompting.

In contrast, all prior metrics, COMET and GEMBA, operate exclusively on textual inputs and translations. Closest to our work, Han et al. (2024) formulate quality estimation for speech translation and propose an end-to-end system that injects audio into a translation-tuned text LLM via a learned modality adapter. They report that this end-to-end approach outperforms cascaded baselines built on state-of-the-art ASR, arguing that SpeechQE should be studied separately from text-QE.

Methodologically, our approaches do not rely on training a dedicated speech-LLM adapter. COMET+audio extends the COMET framework with a lightweight cross-modal fusion over pre-trained encoders, while Speech LLM uses few-shot prompting of an open-source multimodal LLM without any task-specific tuning. Moreover, we find that incorporating audio yields only marginal gains over text-only baselines.

Speech translation. Speech translation systems have traditionally operated as cascaded pipelines, where an automatic speech recognition module first transcribes the source audio, followed by a machine translation model processing the text (Ney, 1999; Sperber and Paulik, 2020; Etchegoyhen et al., 2022). However, these methods are increasingly shifting toward end-to-end architectures (Seamless et al., 2023) that directly map audio to the target language. These end-to-end approaches circumvent errors of automatic speech recognition and can, in theory, leverage acoustic cues, such as prosody and intonation, which are not contained in the text. However, the field of speech translation still lags behind text-to-text translation in both data availability and benchmark maturity, and the evaluations remaining overwhelmingly text-centric (Abdulmumin et al., 2025; Radford et al., 2023; Seamless et al., 2023).

While incorporating source audio into automated metrics is expected to improve the quality estimation due to the extra paralinguistic features, the reality is different. Extracting a reliable signal from raw audio introduces multiple hurdles and we show that background noise, acoustic artifacts, and mismatches between audio and transcripts often counteract the theoretical benefits of the audio modality. However, these observations might be strongly influenced by the used dataset, while the general translation quality estimation task could be having slightly different properties.

3 Methods & Results

To determine the usefulness of the audio input for the speech translation quality estimation task, we experiment with both speech instruction-tuned LLM prompting and a COMET+audio approach.

3.1 Method: Speech LLM

Throughout our evaluations, we use the open-source *Phi-4-multimodal-instruct* (Abouelenin et al., 2025) model, instructed to output a single score for a given sample. The prompt (see Appendix A.1) includes the task description, guidelines, and few-shot examples sampled from the training data split. To simulate Chain-of-Thought reasoning (Wei et al., 2022) and improve the model's reasoning capabilities, it is also asked to provide an explanation for its assigned score beforehand.

To align the model with the dataset's grading scale, we include few-shot examples in the prompt (Brown et al., 2020), each containing the source and target texts along with the corresponding quality score. They are organized into five bins according to their ground-truth scores, spanning from the lowest range (15%–25%) to the highest (95%–100%). At inference time, few-shot examples are randomly drawn from a per-bin pool of ten examples. The final quality score for each translation is obtained by prompting the model five times and averaging the resulting scores.

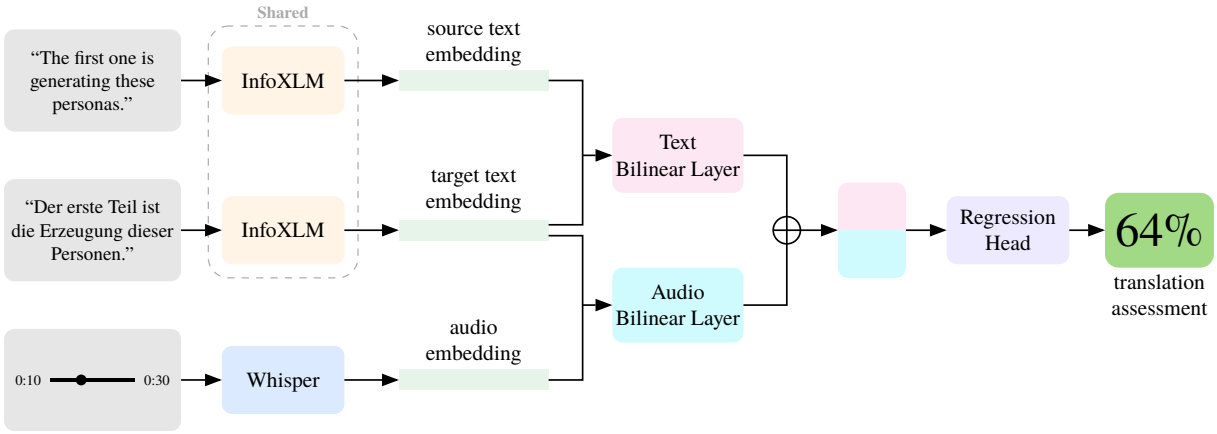

Figure 2: Overview of the COMET+audio architecture. Either the source audio or text can are optional for the input, experiments for which we adapt and train separate architectures.

The few-shot approach proved essential for calibrating the model's output. Without it, the LLM exhibited a strong bias toward extreme scores and failed to predict middle-range values (e.g. 43).

3.2 Method: COMET+audio

We adopt an architecture inspired by COMET and extend it to accommodate the audio modality alongside source and target text. An overview of the architecture is presented in Figure 2. We employ the base variant of InfoXLM as a shared text encoder for both the source transcript and the target translation, and the medium variant of Whisper (Radford et al., 2023) as the audio encoder. To efficiently adapt both pretrained encoders to the quality estimation task, we apply LoRA (Hu et al., 2022) adapters of rank 8 with an alpha scaling factor of 32 to the query and value projection matrices across all attention layers. Additionally, to provide lightweight language conditioning, each transcript is prepended with a language tag (e.g. `[en]` for English, `[de]` for German).

Audio and text representations. For the text modality, we obtain a fixed-length representation by extracting the hidden state at the `[CLS]` token position from the InfoXLM encoder. For the audio modality, since Whisper's encoder produces a variable-length sequence of hidden states, we apply an attention pooling mechanism to derive a fixed-length vector. This mechanism consists of a two-layer network (with a Tanh activation) that computes a scalar attention score for each time step, followed by a softmax normalization. The resulting attention weights are used to compute a weighted sum over the encoder outputs, yielding a single audio representation.

Cross-modal fusion. Each modality-specific representation is then projected into a shared 256-dimensional space through separate linear projections. To capture cross-modal interactions, we apply two bilinear layers: one modeling the interaction between the source audio and target representations, and another between the source transcript and target representations. Each bilinear layer produces a 128-dimensional output, and the two interaction vectors are further concatenated into a single fused vector.

Regression head. The fused representation is then passed through a regression head consisting of layer normalization, a linear layer with Tanh activation and ends with the final linear layer that outputs the predicted quality score. In between these layers dropout is applied to avoid overfitting.

Other approaches and negative results. Beyond the strategies in Table 1, we explored several other approaches to improve the "Speech LLM" method.

The simplest variant replaces dynamically sampled few-shots with a fixed set of five examples shared across all samples. Another simple alternative uses ten more fine-grained bins to sample few-shot examples from, which also doubles their number. Furthermore, after observing that the model struggled to produce mid-range scores (e.g., 64), we reduced the scoring range and prompted the model to give "star"-based scores, for both the 5 and 10 star ranges. We further explored language-specific few-shots in place of general

	Speech LLM			COMET+audio		
	Segment %	System %	MAE	Segment %	System %	MAE
Text and audio	18.4	88.0	21.29	18.3	83.6	18.20 ⋆
Audio-only	17.3	85.5	23.81	16.4	68.7	18.49
Text-only	27.4 ⋆	82.0	21.07	18.2	80.3	18.21
Text and wrong audio	17.7	90.1 ⋆	22.11	16.4	81.0	18.53
Translation-only	14.8	81.2	23.72	13.9	51.9	18.66

Table 1: Performance evaluation of the explored strategies. The LLM prompting results are obtained with constant sampling with dynamic few-shots as this combination led to highest values. For COMET+audio, we adapt and train separate architectures for the *"Audio-only"*, *"Text-only"* and *"Translation"* strategies as the input differs.

ones, hypothesizing that this can help the model accommodate scale differences in ground-truth scores across languages. Additionally, to help the model assess each translation relative to its alternatives, we included all "sibling" translations of the same source sentence in the context, without their ground-truth scores.

Moreover, the high standard deviation of predicted scores for a single translation suggested limited model confidence, motivating a confidence-sampling approach: we resample scores until either the standard deviation falls below a threshold or a maximum number of generations is reached.

Hypothesizing that low performance stemmed from limited context, we prompted an LLM to generate plausible continuations of each sample, which we then appended to the prompt alongside the source, target, and few-shots. Finally, we apply UniPrompt (Juneja et al., 2025) to automatically optimize the prompt to fit the data.

However, none of them showed better results on average than the strategy using constant sampling with dynamic few-shots. More details on these explored strategies and their results can be found in Appendix A.2.

3.3 Experimental Setup

For all our experiments we used the IWSLT 2026 shared task dataset that contains, for each sample, the source language, transcript and audio, the target text translation and the ground truth score. For our COMET-audio training we used the train split, while for all our evaluations for the results within this paper we took the development split into consideration. However, while the train data contains multiple *source* → *target* language pairs, the development one contains only the *English* → *German* and *English* → *Chinese* pairs.

3.4 Experimental Results

To evaluate performance, we employed IWSLT's segment- and system-level meta-metrics alongside Mean Absolute Error (MAE). The results for both strategies are presented in Table 1.

For the *"Text and wrong audio"* strategy, we paired each sample's correct source transcript with an audio of a different and randomly drawn example from the evaluation set. In the *"Translation-only"* condition, the model received only the target translation, with no access to the original source data. Table 1 indicates that:

1. The audio modality is not as effective as the text one and does not add a lot of value when combined with the transcript.
2. Both methods (especially Speech LLM) obtain unreasonably high results even when the source is not present in the input, hinting at a slight preference for the fluency of the translation over its adequacy in the evaluation data.

4 Audio does not add a lot of value

Surprised by the low performance of audio-based automated metrics, we investigate why the addition of recordings into the input does not increase the performance. This section covers three potential hypotheses and evidence, based on manual annotations of data:

- low additional information in audio,
- mismatched audio samples, and
- model preference for text

Moreover, as previously stated, these might only be particular properties of the dataset we've been given rather than a characteristic of the task itself.

4.1 Low extra information in the audio

We manually inspect 100 audios from the IWSLT 2026 metrics shared task dataset and classify each

based on its topic such as "technical". Table 2 contains the distribution of these topics, the technical one dominating with 76% and mainly focusing on machine learning content such as [technical audio]. Combined with "information sharing", these two categories make 96% of data. At the same time, these kinds of audios do not contain rich additional signals, such as intonation or prosody. Furthermore, the informative share contains a consistent portion of very short (up to 4 words) like a [hello message] and plain texts such as an [enumeration of names] that further reduces the benefits of the audio compared to text-only approaches. The absence of additional signal could explain why adding audio does not improve the performance of the automated metrics.

Audio cuts off mid-phrase	27%
Audio-text information mismatch	34%
Transcript punctuation issue	3%
Typos in transcript	1%
Audio aligned with transcript	61.0

Table 3: Distribution of content annotation for 100 random samples from the development data split.

4.2 Mismatched text + audio

The audio samples are not always perfectly aligned with the transcripts and, by extension, with the translations. We annotate these audios across multiple dimensions based on the occurred problem, the results being showcased in Table 3. Out of these data issues, we distinguish two of them that directly impact the effectiveness of the audio: (1) unintentional audio cuts, and (2) mismatches with the textual content, a superset of the former.

Unintentional audio cuts refers to audio not fully delivering the entire message and stops before finishing a sentence. In this [example], the audio cuts off with three words before finishing the sentence that is captured in the transcript.

Other mismatches not caused by audio cuts include samples where the audio has more information than the transcript. Here, we distinguish between the recording containing extra information either before or after the text content. As an example, for [this data point], the audio starts with an additional explanation and only afterwards continues with the content of the text. The audio in this example also happens to miss the last four words.

Technical	76%
Information sharing	20%
Advertising	3%
Citation	1%

Table 2: Topic distribution of inspected samples from the development split.

When audio and transcript disagree, the metric receives conflicting signals about the source: a translation faithful to the transcript may diverge from the audio, so the assigned score is conditioned on an inconsistent reference. Rather than providing complementary evidence, the audio modality adds noise, encouraging the model to rely on text alone and reducing the expected gains of multimodal systems.

4.3 Metrics rely solely on text

By closely looking at the results presented in Table 1 for the *"Text and wrong audio"* strategy and comparing them to the other strategies that include the text modality, the results are comparable or even higher, surprisingly.

Inspired by Hua et al., 2024, to rule out the possibility of modality collapse caused by training of the COMET+audio approach, we consider replacing the joint optimization with Alternating Gradient Descent (Akbari et al., 2023) (see Appendix B.2).

Furthermore, the *"Audio-only"* strategy yields one of the lowest scores across all strategies that provide the source of the translation. We consider this as a strong evidence that both the Speech LLM and our trained COMET+audio metric simply drop the audio modality when combined with text for the source input. This behavior can be also attributed to the misalignments with the transcripts and low-variance prosody of the audio samples.

5 Conclusion

We explored the effects of the source audio in the context of speech translation quality estimation. The study implied the use of two methodologies (LLM prompting and regression model training) for assessing the impact of this modality on the performance for this problem. Surprisingly, we found that incorporating audio yields only marginal gains over text-only baselines. We attribute this to audio-transcript misalignments and lack of rich acoustic signals (such as prosody and intonation) due to the technical nature of the content within the data. However, these results might only

show a particular peculiarity of the dataset we've been given rather than an innate property of the task.

Acknowledgements

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship. He is also the co-organizer of the IWSLT Metrics Shared Task and declares no privileges were used in the making of this paper.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Fortuné Kponou, Mateusz Krubiński, 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, Ashwin Sankar, 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)*, pages 412–481, Vienna, Austria (in-person and online).

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *CoRR* abs/2503.01743. arXiv: 2503.01743.

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, Danni Liu, Nam Luu, Min Ma, Dominik Macháček, Marie Maltais, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Chutong Meng, Mohammadamini Mohammad, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John Ortega, Siqi Ouyang, Sara Papi, Peter Polák, Fabian Retkowski, Beatrice Savoldi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marie Tahon, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. 2026. Speech Translation and Metrics in 2026: Findings of the IWSLT Campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US.

Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam. 2023. Alternating Gradient Descent and Mixture-of-Experts for Integrated Multimodal Perception. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3576–3588.

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Alvarez, Iván G. Torre, Juan Manuel Martín-Doñas, Ander González-Docasal, and Edson Benites Fernandez. 2022. Cascade or Direct Speech Translation? A Case Study. *Applied Sciences* 12(3).

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the Quality of Direct Speech Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21852–21867, Miami, Florida, USA.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. 2024. ReconBoost: Boosting Can Achieve Modality Reconcilement. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, pages 19573–19597.

Gurusha Juneja, Gautam Jajoo, Hua Li, Jian Jiao, Nagarajan Natarajan, and Amit Sharma. 2025. Task Facet Learning: A Structured Approach To Prompt Optimization. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23473–23496.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

H. Ney. 1999. Speech translation: coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pages 517-520 vol.1.

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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 28492–28518.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. Souza, Taisiya Glushkova, Duarte M. Alves, Luísa 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 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645.

Communication Seamless, 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 N. Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Y. Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, 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, and Skyler Wang. 2023. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *CoRR* abs/2308.11596. arXiv: 2308.11596.

Matthias Sperber and Matthias Paulik. 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*

35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

A Speech LLM

This section covers implementation details of Speech LLM along with the other explored strategies for this approach.

A.1 Prompt

For the *"Text and audio"* prompting strategy, we use the prompt shown in Prompt 1, which combines a system-level task description with a user prompt containing few-shot examples and the translation to be evaluated. Since the model tends to reproduce the dummy explanations from the few-shot examples, we include an explicit instruction discouraging this behavior to encourage independent reasoning. The prompts for the other strategies are very similar, differing only in the inputs provided for each sample (e.g. removing the audio) and the corresponding introductory line that references them.

A.2 Other explored strategies

This section includes technical details for various strategies that we explored for the *"Text and audio"* strategy.

Confidence sampling Motivated by the high standard deviation of the sampled scores, we implement confidence sampling as follows: sample 3 scores and keep sampling until a number of 15 is reached or the standard deviation falls under a value of 5. Figure 3 reports the average number of samples drawn across bins defined by the ground-truth scores, showing that the model is more confident on good translations than on bad or mediocre ones.

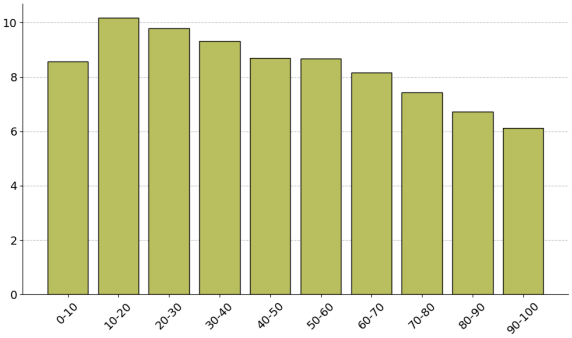

Figure 3: Average number of samples drawn per instance across ground-truth score bins.

Generating continuations for increased context For creating extra context to be used by the model in translation quality assessment, we prompt the same *Phi-4-multimodal-instruct* using Prompt 2. At inference, we inject this continuations into the context and extend the instructions list from Prompt 1 with an extra rule, steering the model into using this extra context *"to better understand the nuances, intent, and direction of the source"*.

UniPrompt To automatically optimize the prompt for our task, we apply the UniPrompt algorithm and adapt the official implementation. We use *Phi-4-multimodal-instruct* in both the *Solver* and *Expert* roles, the audio being passed just to the *Solver*, the *Expert* operating fully on text. Optimization is performed on 1,500 samples drawn from the train split, with a 90/10 train/validation partition. To ensure balanced coverage across the full 0–100 quality range, we apply stratified sam-

```
System Instruction:
You are a translation quality evaluator. Your task is to estimate the quality of the
translation from en to de based on the source audio and transcript and the translated text.
Instructions:
1. Analyze meaning preservation and grammar using the provided transcript as ground truth.
2. Provide a detailed explanation inside <explanation> tags. You must justify your score by
   referencing words or phrases from the source and translation and explain the mistakes or
   good practices. Do not just copy the generic explanations from the examples.
3. Provide a final score from 0 (terrible) to 100 (perfect) inside <score> tags.
---
Evaluation Content:
Here are some examples for scoring translations using just the transcripts.

FEW_SHOTS

Now evaluate this: <|audio_1|>
Transcript of the audio in en language: "The first one is generating these personas."
Translation in de language: "Der erste Teil ist die Erzeugung dieser Personen."
```

Prompt 1: Prompt used for the *"Text and audio"* strategy

```
System Instruction:
You are a helpful monolingual assistant fluent in en. Your task is to read the provided
transcript and listen to the audio, then generate a natural, plausible continuation of the
text in en. Do not provide a translation. Do not provide commentary. Just write the next few
sentences that logically follow the transcript.
```

```
User Message:
<|audio_1|>
Transcript (en): "In watermark injection, we first define a target embedding."

Continuation:
```

Prompt 2: Prompt used for generating plausible continuations as additional context.

pling over 10 score bins of width 10, and restrict the data to *English* → *German* and *English* → *Chinese* pairs.

Results Table 4 presents the results of the explored alternatives to the Speech LLM *"Text and audio"* strategy.

	Segment %	System %	MAE
Static few-shots	24.7	79.2	21.07
10 few-shots	15.0	77.3	22.91
5 star rating	21.2	81.4	22.60
10 star rating	19.2	82.6	22.23
Siblings	16.5	84.0	20.31
Confidence sampling	21.3	77.5	21.36
Continuations	18.4	81.2	21.16
UniPrompt	18.0	77.7	21.71

Table 4: Results of the alternative explored strategies for the Speech LLM method using both the source text and audio.

Audio impact across languages. While audio brings no overall gain, Table 6 reveals an asymmetric effect across language pairs. Compared to the *Text-only* baseline, adding audio to text improves both methods for *English* → *Chinese*, but degrades their performance for the *English* → *German* translation pair.

B COMET+audio

This section covers implementation details and analysis of the COMET+audio method.

B.1 Optimization details

We optimized the model using the AdamW algorithm (Loshchilov and Hutter, 2019) with a Mean Squared Error (MSE) loss objective. The training configuration included a learning rate of 5×10^{-5} and a weight decay coefficient of 0.03. To stabilize training, we employed a warmup period for the initial 10% of steps, succeeded by an inverse square root decay schedule and gradient clipping at a maximum norm of 1.0.

B.2 Alternating Gradient Descent

To verify whether modality collapse occurs during the joint optimization of COMET+audio, we re-train the metric using Alternating Gradient Descent in place of standard joint optimization. Concretely, we partition the parameters into three disjoint groups (audio-specific, text-specific, and the shared regression head) and alternate updates between the audio and text groups at each batch, while the shared head is updated at every step. By decoupling the parameter updates across modalities, each modality is guaranteed an unbiased optimization step, mitigating the risk that one modality dominates the gradient signal during training.

The results, reported in Table 5, show that substituting the correct audio with a mismatched sample leads to only a marginal drop in performance. This finding reinforces our earlier observation that the metric largely disregards the audio modality, even under an optimization strategy explicitly designed to prevent such collapse.

	Segment %	System %	MAE
Text and audio	17.3	82.2	18.93
Text and wrong audio	14.6	79.9	19.71

Table 5: Results for the COMET+audio method when optimized using Alternating Gradient Descent.

English → German	Speech LLM			COMET+audio		
	Segment %	System %	MAE	Segment %	System %	MAE
Text and audio	17.2	88.3	21.75	15.5	68.1	19.31
Audio-only	17.0	89.4	24.21	12.9	48.0	19.78
Text-only	23.8	96.4	20.99	16.0	72.4	19.29
Text and wrong audio	16.2	85.8	22.55	15.2	70.4	19.65
Translation-only	12.8	91.7	24.17	11.0	35.4	19.90

English → Chinese	Speech LLM			COMET+audio		
	Segment %	System %	MAE	Segment %	System %	MAE
Text and audio	19.7	87.7	20.43	21.1	99.0	16.11
Audio-only	17.6	81.6	23.03	19.9	89.4	16.05
Text-only	31.0	67.5	21.23	20.3	88.1	16.16
Text and wrong audio	19.3	94.4	21.29	17.7	91.6	16.45
Translation-only	16.8	70.6	22.88	16.7	68.4	16.32

Table 6: Detailed results broken down by language pair for each method and main strategy.

C Experimental setup

For running the experiments, we used GeForce RTX 4090 GPUs with 24GB of VRAM for both the Speech LLM inference and COMET+audio training.