Enhancing End-to-end Multilingual Medical Speech Translation via Terminology Injection Mechanism

Shuanghong Huang[1], Chong Feng[1,2†], Xia Liu[3†], Jinlei Xu[1], Xuan Zhao[1,2], Ge Shi[1], Yuhang Guo[1], Yulong Gao[1]

[1]*School of Computer Science and Technology, Beijing Institute of Technology*

[2]*Southeast Academy of Information Technology, Beijing Institute of Technology*

[3]*Department of Rheumatology, Key Laboratory of Myositis, China-Japan Friendship Hospital*

Abstract

Speech-to-text translation (S2TT) in the medical domain presents significant challenges due to the complexity of medical terminology and the scarcity of high-quality multilingual data. To address these issues, we propose MMST (**M**ultilingual **M**edical **S**peech-to-text **T**ranslation), a novel framework that systematically integrates domain knowledge into S2TT. Initially, we develop a multilingual medical terminology dictionary utilizing a large language model to extract terms from multilingual medical corpora. Following this, we create MMST, which consists of two main components: a two-stage training approach that integrates general pretraining with fine-tuning specific to the medical domain, and a terminology injection mechanism that embeds target terms into translation prompts, directing the generation process during training. Experiments on a many-to-many multilingual medical dataset show that MMST consistently outperforms robust baselines, achieving an average BLEU improvement of +6.97 and higher BERTScores, especially on terminology-rich and low-resource language pairs.

Keywords: Speech-to-text translation; End-to-end; Terminology injection; Multilingual medical terminology dictionary; Large language Models

1. Introduction

Speech-to-text translation (S2TT) has made impressive progress in recent years [1, 2, 3, 4], enabling the conversion of spoken utterances into written text. This technology has unlocked a range of real-world use cases, including live video subtitling, cross-border communication, international conferencing, and language education in underserved regions. While general-purpose S2TT systems perform well on open-domain content, they often fall short when applied to domain-specific scenarios.

In global healthcare delivery, there is an increasing demand for artificial intelligence (AI) systems that accurately interpret medical conversations across languages. Such capabilities are critical for cross-lingual diagnosis, telemedicine consultations, and the multilingual dissemination of medical knowledge. However, conventional S2TT models trained predominantly on open-domain

[†]Corresponding authors: Chong Feng and Xia Liu (Email: fengchong@bit.edu.cn; ORCID:0000-0002-1691-1584)

data [5, 6] often struggle to maintain domain-specific fidelity and consistently use specialized medical terminology [7].

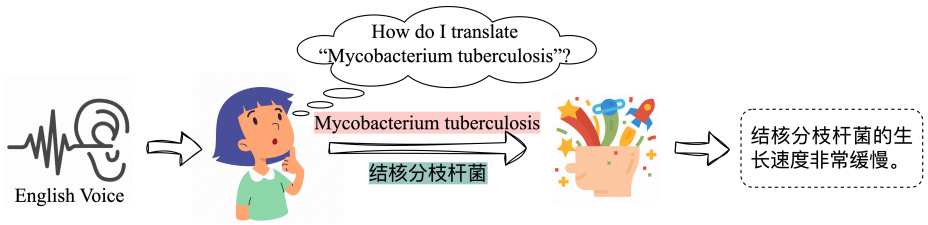

Figure 1: When faced with unfamiliar domain-specific terms in conversation, the model might have difficulty translating them accurately, similar to a human listener. By incorporating a medical terminology dictionary, our approach provides clear term mappings (e.g., “Mycobacterium tuberculosis”→“结核分枝杆菌”), enabling more accurate and reliable translation in medical scenarios.

Medical language is often rich in specialized terminology, acronym usage, and context-sensitive expressions. As shown in Figure 1, terms like “*Mycobacterium tuberculosis*” are crucial to the accuracy of translations. Existing S2TT systems, even those enhanced with large-scale pre-training (e.g., Whisper [8], SeamlessM4T [9]), typically underperform in such cases due to their reliance on general training corpora. Moreover, large language model (LLM)-based generation methods (e.g., LauraGPT [10], Qwen-audio [11]) generally lack explicit domain or terminology learning mechanisms. This often leads to suboptimal handling of medical terms, undermining their accuracy and limiting their reliability in speech translation tasks within healthcare settings.

To address these challenges, we propose **MMST**, a novel end-to-end **M**ultilingual **M**edical **S**peech-to-text **T**ranslation framework that improves translation accuracy through the terminology injection mechanism (TIM). Our core idea is explicitly incorporating domain-specific medical terms into the translation process by aligning recognized speech content with predefined terminology in the target language. This provides an external source of domain knowledge to guide the model toward accurate, consistent, and medically valid outputs.

To implement this proposal, we first construct a multilingual medical terminology dictionary by automatically extracting term pairs from high-quality medical corpora using LLMs. These term pairs are then integrated into the model through a specialized prompting strategy, enabling dynamic terminology guidance during translation.

Meanwhile, we adopt a two-stage training strategy. In the first stage, the model is pretrained on large-scale automatic speech recognition (ASR) data to learn robust acoustic and linguistic representations. In the second stage, we fine-tune the model on multilingual medical S2TT corpora, where terminology injection is actively employed to enhance domain fidelity. This terminology-aware training pipeline enables the model to generalize effectively while accurately translating domain-specific expressions.

MMST is evaluated on a multilingual medical S2TT dataset involving five languages under many-to-many translation settings. Experimental results demonstrate that MMST consistently outperforms strong baselines, achieving an average improvement of +6.97 BLEU, with notable gains on terminology-rich segments.

Our main contributions are summarized as follows:

- We introduce the TIM that explicitly incorporates domain-specific medical terms into the end-to-end S2TT model, enhancing translation accuracy and consistency in the medical domain.

- We develop a terminology-aware framework that integrates LLM-based architecture and domain-specific training strategies for multilingual medical S2TT.
- We conduct comprehensive evaluations on a multilingual medical S2TT dataset, showing substantial improvements over strong baselines across multiple language pairs.

2. Related Work

2.1. Speech-to-Text Translation

S2TT relies on cascaded systems that combine ASR and machine translation (MT) modules [1, 12, 13]. However, these pipelines can suffer from error propagation, degrading overall translation quality. End-to-end S2TT models provide a compelling alternative, translating speech inputs to text outputs in one neural architecture [14, 15, 16, 17]. Recent studies [8, 9] indicate these models can match or surpass cascaded systems in performance while being more efficient and simpler.

Building on this advancement, researchers have shifted their focus to LLMs to improve speech translation further. Utilizing the robust language modeling and instruction-following features of LLMs, hybrid architectures have been created that integrate a speech encoder with either a frozen or fine-tuned LLM. Noteworthy examples are LauraGPT [10], Qwen-audio [11], and SALMONN [18], all of which cater to various speech tasks and deliver results that are competitive with those of task-specific models.

However, these systems are usually trained on general-domain data, which hinders their adaptation to specialized fields like medicine. Furthermore, they seldom incorporate external knowledge (e.g., terminology) during training and inference, which restricts accuracy.

2.2. Terminology Control and Prompting Engineering

Managing terminology in domain-specific MT, particularly in healthcare, law, and finance, remains challenging. Conventional neural MT systems [19] typically rely on external lexicons or limited decoding methods for accurate terminology. However, these approaches are often inflexible and difficult to integrate into modern end-to-end architectures.

Recent advances in LLMs have introduced more flexible methods for terminology control through prompt engineering. In the MT domain, techniques such as in-context learning and template prompting have shown promise in guiding LLMs to adhere to predefined glossaries without retraining [20, 21, 22]. However, these methods typically assume clean, well-aligned text input and output, which limits their applicability in speech-based scenarios.

Conversely, S2TT remains relatively neglected in terms of terminology control. End-to-end S2TT models infrequently integrate domain-specific prompts, particularly in multilingual contexts where both speech and text must be concurrently processed. To address this gap, we propose that the TIM integrate medical knowledge into the translation process during training to improve the model's translation performance within domain-specific and multilingual contexts.

3. Methodology

This section presents our method in detail. We begin by introducing the task definition of S2TT in Section 3.1 Then, we describe the architecture of the proposed model in Section 3.2, followed by the training and inference procedures in Section 3.3.

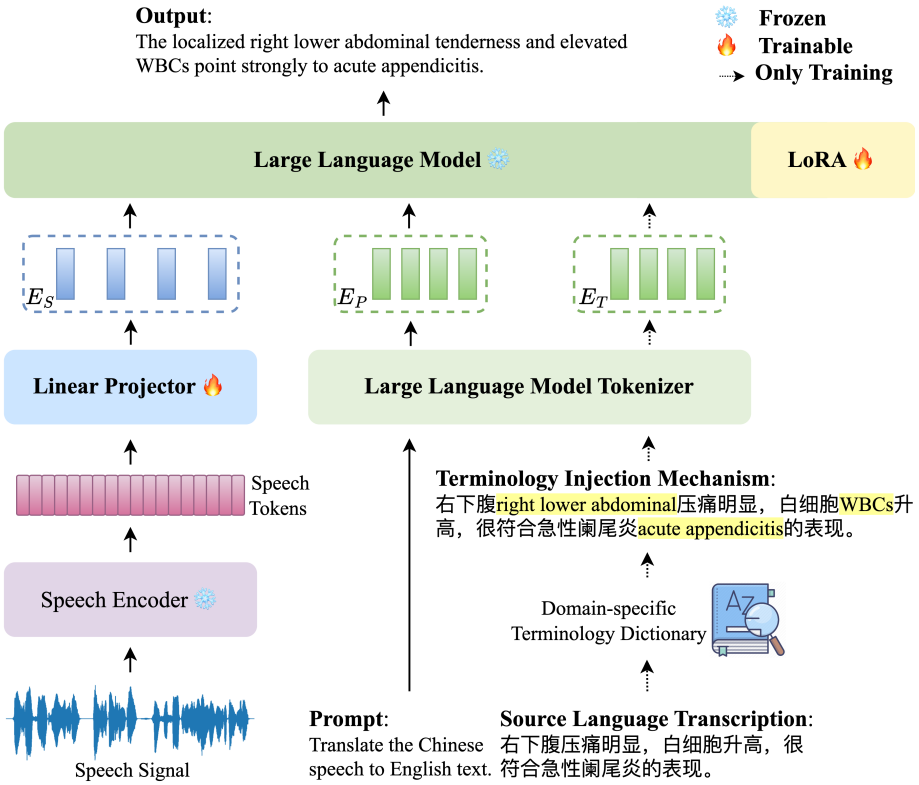

Figure 2: **Overview of the MMST model architecture.** The model architecture consists of a frozen speech encoder for extracting fine-grained acoustic features, a trainable linear projector for mapping these features to the LLM input space, and a decoder-only LLM fine-tuned via LoRA. During training, the TIM guides the model toward accurate and domain-consistent outputs.

3.1. Task Definition

Given a S2TT dataset $\mathcal{D} = (\mathbf{S}, \mathbf{X}_{src}, \mathbf{Y}_{tgt})$, where $\mathbf{S}$ denotes the input speech signal, $\mathbf{X}_{src}$ denotes the source language transcription, and $\mathbf{Y}_{tgt}$ corresponds to the target language text. To process the input speech $\mathbf{S}$, we initially extract acoustic features (like mel-spectrograms), denoted as $\mathbf{X}_S$. The feature extraction process is defined as:

$$\mathbf{X}_S = \mathcal{F}_{extract}(\mathbf{S}), \mathbf{X}_S = \{x_1, x_2, \cdots, x_T\} \tag{1}$$

where $\mathcal{F}_{extract}$ denotes the acoustic feature extraction function, and T is the total number of timesteps in the extracted feature sequence.

The goal of S2TT is to generate the prediction text of target language $\hat{\mathbf{Y}}_{tgt}$ based on the source speech $\mathbf{S}$. This can be formulated as:

$$\hat{\mathbf{Y}}_{tgt} = \mathcal{F}(\mathbf{S}) \tag{2}$$

where $\mathcal{F}$ denotes the end-to-end S2TT model, which can be adapted to handle multilingual input. The performance of the predicted transcription $\hat{\mathbf{Y}}_{tgt}$ is evaluated against the ground truth $\mathbf{Y}_{tgt}$ using standard metrics such as BLEU [23] and BERTScore [24].

3.2. Model Architecture

Our objective is to create the MMST model featuring a lightweight yet efficient architecture, as shown in Figure 2. The model comprises three key components: a frozen speech encoder that captures detailed representations from input speech signals, a trainable linear projector that transforms speech features into a format suitable for LLMs, and a decoder-only LLM that has been efficiently fine-tuned to generate accurate target text

To improve adaptability to different domains and enhance translation precision, we implement a **T**erminology **I**njection **M**echanism (TIM) during the training phase. This process involves inserting domain-specific terms into the LLM input, directing the model to produce outputs that are more accurate and consistent with the terminology. As a result, this mechanism enables the model to manage domain-specific terms more effectively in the S2TT task.

Speech Encoder. The acoustic features $\mathbf{X}_S$ encapsulate rich information, including semantic content, speaker identity, emotional state, paralinguistic cues, and background noise. The core function of the speech encoder is to disentangle these heterogeneous factors and extract robust linguistic representations, denoted as $\mathbf{H}_S$. This transformation is formally defined as:

$$\mathbf{H}_S = \mathcal{F}_{se}(\mathbf{X}_S) \tag{3}$$

where $\mathcal{F}_{se}$ denotes the speech encoder function.

Linear Projector. The linear projector serves as a crucial bridge between the speech encoder and the LLM, featuring a lightweight set of trainable parameters. This module aligns the extracted speech representations with the LLM's embedding space by learning an effective transformation. Specifically, it projects the speech encoder output $\mathbf{H}_S$ into the LLM's input embedding space, producing the transformed representation $\mathbf{E}_S$:

$$\mathbf{E}_S = \mathcal{F}_{proj}(\mathbf{H}_S) \tag{4}$$

where $\mathcal{F}_{proj}$ denotes the linear projector function.

This transformation allows for the smooth integration of features derived from speech into the LLM's text-based representation space. In our approach, we utilize a two-layer multilayer perceptron (MLP) as the projector to effectively connect the modality gap between speech and text.

Terminology Injection Mechanism. To improve domain awareness and ensure the correct use of terminology in MMST, we propose the TIM, which relies on a predefined multilingual medical terminology dictionary $\mathcal{T} = (s_i, t_i)$, where s_i and t_i represent the terms in the source and target languages, respectively. Appendix B provides a detailed description of the dictionary.

During preprocessing, we construct the terminology-enhanced source transcription $\mathbf{T}_{term}$ by identifying source terms in the original transcription $\mathbf{X}_{src}$ and appending their corresponding target terms t_i from the dictionary $\mathcal{T}$ immediately after each match. This results in enriched input containing both the source term and its translation, which helps guide the model towards more accurate and terminology-consistent generation. Formally:

$$\mathbf{T}_{term} = \mathcal{F}_{TIM}(\mathbf{X}_{src},\ \{(s_i, t_i) \mid (s_i, t_i) \in \mathcal{T},\ s_i \in \mathbf{X}_{src}\}) \tag{5}$$

where $\mathcal{F}_{TIM}$ denotes the TIM function.

The terminology-enhanced sequence $\mathbf{T}_{term}$ is subsequently tokenized using the LLM tokenizer to produce the term-augmented representation $\mathbf{E}_T$, which is utilized during training to guide the model towards consistent and accurate terminology usage in domain-specific generation.

Large Language Model. To enable the LLM to generate multilingual outputs from audio inputs, we construct an input sequence consisting of three parts: the projected speech representation $\mathbf{E}_S$ that is compatible with the LLM, a textual prompt representation $\mathbf{E}_P$ derived from tokenizing and embedding the prompt that satisfies the desired translation behavior, and the terminology-enhanced representation $\mathbf{E}_T$ obtained through TIM. These elements are concatenated and fed into the LLM to guide the multilingual generation process:

$$\hat{\mathbf{Y}}_{tgt} = \mathcal{F}_{llm}([\mathbf{E}_S, \mathbf{E}_P, \mathbf{E}_T]) \tag{6}$$

where [,] denotes the concatenation operation. For simplicity, we omit the bos and eos tokens.

3.3. Training and Inference

Training. To improve training efficiency during fine-tuning, we adopt Low-Rank Adaptation (LoRA) [25] for efficient model updates. LoRA adds low-rank decomposition matrices into each Transformer layer, enabling effective adaptation to the medical domain with far fewer trainable parameters, while keeping most pre-trained weights frozen. This allows the model to retain general linguistic knowledge from the ASR pretraining stage while incorporating domain-specific terminology through targeted updates.

We perform instruction tuning based on the LLM's inherent autoregressive objective. For a target sequence $\mathbf{Y}_{tgt}$ of length N, the goal of training is to optimize the conditional likelihood:

$$P(\mathbf{Y}_{tgt}|\mathbf{X}) = \prod_{i=1}^{N} P_\theta(y_i|\mathbf{X}, \mathbf{Y}_{tgt,<i}) \tag{7}$$

where $\mathbf{X} = \{\mathbf{X}_S, \mathbf{T}_{Prompt}, \mathbf{T}_{Term}\}$ represents the concatenated input.

This approach reduces computational overhead and alleviates overfitting, making it ideal for low-resource settings and enabling rapid adaptation to domain-specific tasks.

Inference. For the multilingual generation, we adopt beam search decoding with a beam size of 4, offering a favorable balance between generation quality and computational efficiency.

4. Experiments Settings

4.1. Datasets

The MMST model undergoes training and evaluation using the MultiMed-ST [7] dataset. The statistics for this dataset are presented in Table A.2 in Appendix A. MultiMed-ST facilitates many-to-many multilingual S2TT across five languages: Vietnamese (vi), English (en), German (de), French (fr), and Chinese (zh), encompassing around 48K samples that span all translation directions.

The dataset consists of training, development, and test sets, with **en** as the main source language due to its prevalent use in real-world medical communication. This varied and fairly balanced dataset enables thorough evaluation of multilingual S2TT in specialized contexts. All audio samples are downsampled to 16kHz to maintain consistency throughout the experiments.

4.2. Configurations and hyperparameters

The MMST model employs Whisper-large-v2 [8] as the speech encoder, contributing approximately 1 billion parameters. A lightweight two-layer MLP projects the 1280-dimensional speech features into the input space of the downstream LLM. Consequently, the overall parameter count is primarily determined by the LLM, for which we utilize Llama3-8B [26] as the backbone decoder.

All models are optimized using the AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate schedule with warm-up followed by linear decay is employed, featuring a peak learning rate of 2e-4. Unless otherwise specified, the LoRA rank and scaling factor α are set to 8 and 32, respectively. Training is conducted with a batch size of 1 per GPU and gradient accumulation steps of 8. All experiments are carried out on four NVIDIA RTX 3090 GPUs (24GB each).

4.3. Baselines

To assess the effectiveness of our proposed method, we compare it against various strong baselines, categorized into cascaded and end-to-end models based on their architectural design.

Cascaded Models. Cascaded models separate the S2TT into two stages: ASR followed by MT. We use Whisper-large-v2 [8] as the ASR component and pair it with the subsequent MT models:
• **mBART-large-50** [27]: is a multilingual sequence-to-sequence model aimed at facilitating multilingual MT by jointly fine-tuning across various language directions. This version enhances the original mBART [28] by incorporating 25 new languages, bringing the total to 50 languages for multilingual MT.
• **M2M100-418M** [29]: is a multilingual encoder-decoder model designed for many-to-many multilingual MT. It supports direct translation across 9,900 directions involving 100 languages.

End-to-end Models. End-to-end models translate speech directly into the target language, bypassing intermediate transcriptions and reducing error propagation.
• **Whisper-large** [8]: is a multitask speech model trained with weak supervision on large-scale multilingual data. It supports ASR and S2TT.
• **SeamlessM4T-large-v2** [9]: is an integrated model designed for speech and text translation, accommodating almost 100 languages. It performs numerous functions such as speech-to-speech (S2ST), S2TT, text-to-speech (T2ST), text-to-text (T2TT), and ASR, making it ideal for diverse multilingual translation.
• **QwenAudio-2-7B-Instruct** [11]: is a large-scale audio-language model developed to process various audio inputs. It supports tasks such as audio comprehension and following speech instructions, generating text-based responses directly from audio signals.

5. Experiments Results

5.1. Main Result

Table 1 reports BLEU and BERTScore results across 20 language directions, comparing cascaded systems, end-to-end baselines, and our proposed model. In terms of BLEU, our approach consistently outperforms existing end-to-end systems in the majority of language pairs. For instance, on the **vi-en** task, our model achieves 20.62 BLEU, substantially higher than *SeamlessM4T-large-v2* (14.40), *QwenAudio-2-7B-Instruct* (1.66), and *Whisper-large* (8.18). A similar advantage is observed in the **zh-en** direction (23.36 vs. 14.22 and 19.63), as well as in **vi-fr**, **de-en**, and **de-fr**, with BLEU scores of 12.41, 31.72, and 21.13, respectively.

Table 1: BLEU and BERTScore results for 20 language directions. The *Whisper-large* model only supports translation from X to English. Only the best BLEU scores from end-to-end models are highlighted in **Bold**.

Model	Metrics	en-vi	en-fr	en-zh	en-de	vi-en	vi-fr	vi-zh	vi-de	fr-en	fr-vi	fr-zh	fr-de	de-en	de-vi	de-fr	de-zh	zh-en	zh-vi	zh-fr	zh-de	Avg.
Cascaded																						
Whipser-large + mBart-large-50	BLEU	53.43	47.69	40.82	39.19	6.71	8.67	11.30	4.19	29.47	28.01	20.63	21.39	35.29	35.96	34.56	28.81	7.41	12.39	9.51	5.90	24.07
	BERTScore	0.89	0.88	0.84	0.86	0.86	0.72	0.65	0.68	0.92	0.82	0.74	0.79	0.93	0.85	0.85	0.79	0.81	0.61	0.74	0.71	0.80
Whipser-large + M2M100-418M	BLEU	53.42	47.96	42.05	40.52	10.85	9.68	11.45	7.76	32.19	29.84	25.52	25.25	37.90	38.51	37.72	28.69	18.71	24.20	16.83	13.66	27.64
	BERTScore	0.96	0.93	0.92	0.93	0.73	0.72	0.70	0.71	0.84	0.79	0.79	0.79	0.81	0.83	0.86	0.74	0.78	0.85	0.78	0.78	0.81
End-to-end																						
Whisper-large	BLEU	/	/	/	/	8.18	/	/	/	26.06	/	/	/	**37.32**	/	/	/	16.54	/	/	/	/
	BERTScore	/	/	/	/	0.75	/	/	/	0.81	/	/	/	0.85	/	/	/	0.79	/	/	/	/
SeamlessM4T-large-v2	BLEU	24.59	25.68	20.43	20.19	14.4	10.19	11.49	7.4	29.23	17.49	11.37	15.94	25.09	15.07	12.88	11.45	14.22	11.39	6.83	4.16	15.47
	BERTScore	0.81	0.82	0.76	0.8	0.77	0.75	0.74	0.72	0.82	0.78	0.72	0.77	0.82	0.77	0.75	0.73	0.79	0.74	0.73	0.70	0.76
QwenAudio-2-7B-Instruct	BLEU	24.46	30.16	23.3	22.69	1.66	1.17	2.36	1.13	23.63	11.49	**15.37**	14.51	23.29	11.07	14.88	16.04	19.63	15.72	13.52	10.37	14.82
	BERTScore	0.80	0.82	0.76	0.79	0.66	0.65	0.65	0.66	0.79	0.74	0.71	0.74	0.80	0.73	0.76	0.72	0.80	0.78	0.77	0.77	0.75
Our Model																						
MMST	BLEU	**28.30**	**33.33**	**26.80**	**25.41**	**20.62**	**12.41**	**17.58**	**14.75**	**32.73**	**23.37**	10.17	**20.90**	31.72	**20.43**	**21.13**	**18.39**	**23.36**	**20.21**	**18.07**	**16.11**	**21.79**
	BERTScore	0.85	0.86	0.80	0.85	0.83	0.81	0.82	0.80	0.84	0.80	0.71	0.78	0.84	0.79	0.80	0.75	0.82	0.84	0.78	0.81	0.81
MMST w/o TIM	BLEU	20.59	31.41	19.15	22.59	16.25	9.28	15.18	10.12	17.29	19.28	9.68	18.98	24.63	18.35	21.25	11.48	21.59	18.07	16.44	12.69	17.72
	BERTScore	0.76	0.84	0.74	0.82	0.78	0.74	0.76	0.79	0.83	0.79	0.70	0.77	0.80	0.78	0.81	0.73	0.80	0.81	0.77	0.79	0.78

While cascaded systems, particularly *Whisper + M2M100*, still yield the highest scores in some directions (e.g., 42.05 BLEU for **en-zh**, 37.72 for **de-fr**), they come at the expense of higher inference latency and increased susceptibility to error propagation. In contrast, our model strikes a more favorable balance between performance and efficiency. For example, in the **fr-en** direction, our system achieves 32.73 BLEU, closely matching *Whisper + M2M100* (32.19) while providing a more streamlined architecture.

Our model showcases remarkable semantic performance, as highlighted by BERTScore results. Notably, it achieves scores of 0.81 on **vi-fr** and 0.84 on **zh-vi**, substantially outperforming *QwenAudio-2-7B-Instruct* and *SeamlessM4T-large-v2*.

Our approach provides robust and scalable performance across the multilingual medical S2TT task. It achieves results that are competitive or superior to both cascaded and end-to-end baselines, while maintaining low inference complexity and demonstrating strong fidelity in translating domain-specific content. Appendix C and Appendix D, respectively, provide additional results, including ASR performance and qualitative case studies.

5.2. Analysis

To evaluate the effects of the proposed TIM, we conduct an ablation study comparing the full model with a version that removes terminology injection, called *MMST w/o TIM*. As shown in Table 1, omitting TIM consistently decreases performance across all language directions, as indicated by both BLEU and BERTScore metrics.

Notably, the performance gains brought by TIM vary across language pairs. In particular, low-resource directions, such as *vi-en*, benefit significantly more from TIM compared to high-resource directions, like *en-fr*. This pattern indicates that in low-resource settings, TIM plays a vital role in enhancing term alignment and guiding the model to produce accurate translations. In contrast, high-resource language pairs likely benefit less because the LLMs have already gained sufficient alignment and terminology knowledge through pretraining and fine-tuning on large amounts of data. Therefore, TIM acts as a stronger correction tool when LLMs' capacity alone is insufficient, thereby increasing its effect in low-resource scenarios.

We further observed that the effectiveness of TIM is closely linked to terminology density, with translations containing more matched terms generally showing greater improvements. Preliminary experiments using partial injection strategies, such as injecting only high-confidence terms, yielded limited gains compared to complete injection, indicating that ensuring comprehensive terminology coverage is essential in clinical settings. Moreover, incorrect term injection,

including numerical mismatches or loosely related paraphrases, was found to negatively affect translation quality during our experiments, especially in medically sensitive contexts.

6. Conclusion

In this work, we propose MMST, a multilingual S2TT framework tailored for the medical domain. MMST integrates a terminology dictionary constructed via LLM-based extraction, incorporates a TIM for terminology-aware translation, and employs a two-stage training strategy combining general ASR pretraining with domain-specific fine-tuning. Experiments on the MultiMed-ST dataset show that MMST achieves an average improvement of +6.97 BLEU over strong baselines, particularly on terminology-rich segments. To our knowledge, this is the first work to incorporate TIM into multilingual S2TT systematically. Future work includes extending MMST to support more low-resource languages and specialized domains such as law and finance.

Acknowledgements

We would like to thank all the anonymous reviewers for their insightful and valuable comments. This work was supported by the Beijing Natural Science Foundation-Haidian Original Innovation Joint Foundation (Grant L232119) and the National Natural Science Foundation of China (Grant No. 81302586).

References

[1] H. Ney, Speech translation: coupling of recognition and translation, in: Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, Arizona, USA, March 15-19, 1999, IEEE Computer Society, 1999, pp. 517–520. doi:10.1109/ICASSP.1999.758176.
URL https://doi.org/10.1109/ICASSP.1999.758176

[2] C. Xu, B. Hu, Y. Li, Y. Zhang, S. Huang, Q. Ju, T. Xiao, J. Zhu, Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, Association for Computational Linguistics, 2021, pp. 2619–2630. doi:10.18653/V1/2021.ACL-LONG.204.
URL https://doi.org/10.18653/v1/2021.acl-long.204

[3] S. Indurthi, S. Chollampatt, R. Agrawal, M. Turchi, CLAD-ST: contrastive learning with adversarial data for robust speech translation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 9049–9056. doi:10.18653/V1/2023.EMNLP-MAIN.560.
URL https://doi.org/10.18653/v1/2023.emnlp-main.560

[4] X. Chen, S. Zhang, Q. Bai, K. Chen, S. Nakamura, Llast: Improved end-to-end speech translation system leveraged by large language models, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 6976–6987. doi:10.18653/V1/2024.FINDINGS-ACL.416.
URL https://doi.org/10.18653/v1/2024.findings-acl.416

[5] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 4218–4222.
URL https://aclanthology.org/2020.lrec-1.520/

[6] M. A. D. Gangi, R. Cattoni, L. Bentivogli, M. Negri, M. Turchi, Must-c: a multilingual speech translation corpus, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 2012–2017. doi:10.18653/V1/N19-1202.
URL https://doi.org/10.18653/v1/n19-1202

[7] K. Le-Duc, T. Tran, B. P. Tat, N. K. H. Bui, Q. Dang, H.-P. Tran, T.-T. Nguyen, L. Nguyen, T.-M. Phan, T. T. P. Tran, et al., Multimed-st: Large-scale many-to-many multilingual medical speech translation, arXiv preprint arXiv:2504.03546 (2025).

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, Vol. 202 of Proceedings of Machine Learning Research, PMLR, 2023, pp. 28492–28518.
URL https://proceedings.mlr.press/v202/radford23a.html

[9] S. Communication, L. Barrault, Y. Chung, M. C. Meglioli, D. Dale, N. Dong, P. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, C. Klaiber, P. Li, D. Licht, J. Maillard, A. Rakotoarison, K. R. Sadagopan, G. Wenzek, E. Ye, B. Akula, P. Chen, N. E. Hachem, B. Ellis, G. M. Gonzalez, J. Haaheim, P. Hansanti, R. Howes, B. Huang, M. Hwang, H. Inaguma, S. Jain, E. Kalbassi, A. Kallet, I. Kulikov, J. Lam, D. Li, X. Ma, R. Mavlyutov, B. N. Peloquin, M. Ramadan, A. Ramakrishnan, A. Y. Sun, K. Tran, T. Tran, I. Tufanov, V. Vogeti, C. Wood, Y. Yang, B. Yu, P. Andrews, C. Balioglu, M. R. Costa-jussà, O. Celebi, M. Elbayad, C. Gao, F. Guzmán, J. Kao, A. Lee, A. Mourachko, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, P. Tomasello, C. Wang, J. Wang, S. Wang, Seamlessm4t-massively multilingual & multimodal machine translation, CoRR abs/2308.11596 (2023). arXiv:2308.11596, doi:10.48550/ARXIV.2308.11596.
URL https://doi.org/10.48550/arXiv.2308.11596

[10] J. Wang, Z. Du, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma, W. Wang, S. Zheng, C. Zhou, Z. Yan, S. Zhang, Lauragpt: Listen, attend, understand, and regenerate audio with GPT, CoRR abs/2310.04673 (2023). arXiv:2310.04673, doi:10.48550/ARXIV.2310.04673.
URL https://doi.org/10.48550/arXiv.2310.04673

[11] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, J. Zhou, Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, CoRR abs/2311.07919 (2023). arXiv:2311.07919, doi:10.48550/ARXIV.2311.07919.
URL https://doi.org/10.48550/arXiv.2311.07919

[12] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, S. Yamamoto, The ATR multilingual speech-to-speech translation system, IEEE Trans. Speech Audio Process. 14 (2) (2006) 365–376. doi:10.1109/TSA.2005.860774.
URL https://doi.org/10.1109/TSA.2005.860774

[13] T. K. Lam, S. Schamoni, S. Riezler, Cascaded models with cyclic feedback for direct speech translation, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021, IEEE, 2021, pp. 7508–7512. doi:10.1109/ICASSP39728.2021.9413719.
URL https://doi.org/10.1109/ICASSP39728.2021.9413719

[14] M. Sperber, G. Neubig, J. Niehues, A. Waibel, Attention-passing models for robust and data-efficient end-to-end speech translation, Trans. Assoc. Comput. Linguistics 7 (2019) 313–325. doi:10.1162/TACL_A_00270.
URL https://doi.org/10.1162/tacl_a_00270

[15] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, J. M. Pino, Fairseq S2T: fast speech-to-text modeling with fairseq, in: D. Wong, D. Kiela (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020, Association for Computational Linguistics, 2020, pp. 33–39.
URL https://aclanthology.org/2020.aacl-demo.6/

[16] R. Ye, M. Wang, L. Li, End-to-end speech translation via cross-modal progressive training, in: H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, P. Motlícek (Eds.), 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021, ISCA, 2021, pp. 2267–2271. doi:10.21437/INTERSPEECH.2021-1065.
URL https://doi.org/10.21437/Interspeech.2021-1065

[17] R. Ye, M. Wang, L. Li, Cross-modal contrastive learning for speech translation, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 5099–5113. doi:10.18653/V1/2022.NAACL-MAIN.376.

URL https://doi.org/10.18653/v1/2022.naacl-main.376
[18] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, C. Zhang, SALMONN: towards generic hearing abilities for large language models, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net, 2024.
URL https://openreview.net/forum?id=14rn7HpKVk
[19] G. Dinu, P. Mathur, M. Federico, Y. Al-Onaizan, Training neural machine translation to apply terminology constraints, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 3063–3068. doi:10.18653/V1/P19-1294.
URL https://doi.org/10.18653/v1/p19-1294
[20] OpenAI, GPT-4 technical report, CoRR abs/2303.08774 (2023). arXiv:2303.08774, doi:10.48550/ARXIV.2303.08774.
URL https://doi.org/10.48550/arXiv.2303.08774
[21] Q. Ye, M. Ahmed, R. Pryzant, F. Khani, Prompt engineering a prompt engineer, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 355–385. doi:10.18653/V1/2024.FINDINGS-ACL.21.
URL https://doi.org/10.18653/v1/2024.findings-acl.21
[22] Y. Wu, G. Hu, Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings, in: P. Koehn, B. Haddon, T. Kocmi, C. Monz (Eds.), Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023, Association for Computational Linguistics, 2023, pp. 166–169. doi:10.18653/V1/2023.WMT-1.15.
URL https://doi.org/10.18653/v1/2023.wmt-1.15
[23] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA, ACL, 2002, pp. 311–318. doi:10.3115/1073083.1073135.
URL https://aclanthology.org/P02-1040/
[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.
URL https://openreview.net/forum?id=SkeHuCVFDr
[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022.
URL https://openreview.net/forum?id=nZeVKeeFYf9
[26] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, et al., The llama 3 herd of models, CoRR abs/2407.21783 (2024). arXiv:2407.21783, doi:10.48550/ARXIV.2407.21783.
URL https://doi.org/10.48550/arXiv.2407.21783
[27] Y. Tang, C. Tran, X. Li, P. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, CoRR abs/2008.00401 (2020). arXiv:2008.00401.
URL https://arxiv.org/abs/2008.00401
[28] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Trans. Assoc. Comput. Linguistics 8 (2020) 726–742. doi:10.1162/TACL_A_00343.
URL https://doi.org/10.1162/tacl_a_00343
[29] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, M. Auli, A. Joulin, Beyond english-centric multilingual machine translation, J. Mach. Learn. Res. 22 (2021) 107:1–107:48.
[30] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu,

R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, CoRR abs/2412.15115 (2024). arXiv:2412.15115, doi:10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115

Appendix A. Dataset Statistics

Table A.2 summarizes the datasets used in our experiments along with their corresponding languages.

Table A.2: Statistics of the MultiMed-ST dataset categorized by language and data division. The notation "→X" indicates that the language on the left is the source language, while "X" represents all other target languages. The 'K' denotes thousands.

Language		vi→X	en→X	de→X	fr→X	zh→X
#Samples	Training	4K5	25K5	1K4	1K4	1K2
	Development	1K1	2K8	300	40	90
	Test	3K4	4K8	1K1	300	200
	All	9K1	33K1	2K8	1K8	1K6

Appendix B. Multilingual Medical Terminology Dictionary

To construct the dictionary, we create a four-step extraction pipeline specifically designed for multilingual medical texts. First, we align sentence- or clause-level segments across five languages using structural cues such as punctuation, numerals, named entities, and length similarity. Based on the aligned segments, we identify candidate terms by selecting medical noun phrases and removing expressions unlikely to be terminological (e.g., verbs, quantities, dates, or stopwords).

We then perform cross-lingual semantic verification to ensure conceptual consistency across languages, accommodating surface variations in expression, such as different word orders or mismatches in singular and plural forms. Finally, we normalize term formats with consistent styling, including capitalization, spacing, and abbreviation use, prioritizing full-form expressions when possible. The Qwen2.5-32B [30] model powers this process, producing a high-quality multilingual terminology dictionary with approximately 30,000 validated term pairs. Table B.3 shows an excerpt from the multilingual medical terminology dictionary created using the pipeline mentioned above.

Table B.3: Excerpt from the multilingual medical terminology dictionary.

vi	en	zh	de	fr
hen suyễn	asthma	哮喘	Asthma	asthme
bệnh tiểu đường	diabetes	糖尿病	Diabetes	diabète
đột quỵ	stroke	中风	schlaganfall	AVC
đau tim	heart attack	心脏病发作	Herzinfarkt	crise cardiaque
tăng huyết áp	hypertension	高血压	hypertonie	hypertension

Appendix C. Automatic Speech Recognition Results

To further assess the transcription quality of our model, we report ASR performance across five languages in Table C.4. The Character Error Rate (CER%) is used for **zh**, while the Word Error Rate (WER%) is employed for all other languages, with lower values indicating better performance.

Table C.4: Evaluation is conducted using CER% for zh and WER% for all other languages, with lower values indicating better performance.

ASR	vi	en	zh	de	fr
Whisper-small	33.40	40.90	89.80	19.60	55.30
Whisper-large	62.60	25.50	37.30	24.20	41.70
QwenAudio-2-7B-Instruct	54.51	29.61	37.23	23.39	65.62
MMST	**30.03**	**17.14**	**35.22**	**16.48**	**33.97**

Our approach consistently surpasses *Whisper-small*, *Whisper-large*, and *QwenAudio-2-7B-Instruct* in all languages evaluated. Specifically, it registers the lowest error rates in **en** (17.14%), **de** (16.48%), and **fr** (33.97%), showcasing robust generalization across diverse linguistic categories.

Furthermore, our model reduces the WER for **vi** by over 50% compared to *Whisper-large* (from 62.60% to 30.03%), highlighting the impact of domain adaptation, even in low-resource contexts. These results suggest that the improvements in S2TT's performance are not only due to the translation aspect but also arise from enhanced transcription quality.

Appendix D. Case Study

In addition to improvements in medical terminology translation, Table D.5 highlights three common types of critical errors made by the *QwenAudio-2-7B-Instruct* model. The first is numerical mistranslation. In the **en-zh** translation, "10万个" is incorrectly translated as "一百万个," causing an order-of-magnitude error that could significantly distort clinical information. In contrast, our MMST model correctly translates the number as "100,000个," accurately preserving the original meaning. The second issue involves a shift in sentiment polarity. In the **vi-zh** translation, the phrase expressing concern is mistranslated as a positive emotion ("高兴"), effectively reversing the intended emotional tone. MMST maintains the original sentiment and accurately represents the actual emotional state.

The third common issue is pragmatic misinterpretation. In the **fr-zh** translation, "现在能接受我的预约吗?" adds an unnecessary verb that alters the original intent. In contrast, MMST translates it as "现在预约一下这些时间," accurately maintaining the original tone and meaning. These examples show the limitations of shallow lexical alignment and highlight the need for discourse-aware modeling. By incorporating terminology alignment and contextual grounding, MMST significantly reduces such errors, providing translations that are both clinically reliable and pragmatically appropriate.

Table D.5: Case study of multilingual medical S2TT. "→" denotes translation source speech into target language text. Red highlights significant errors, while green indicates consistency with the ground truth.

Source	Target	Model	Text
en	→vi	Ground Truth	trong 100.000 tế bào là nguyên nhân gây bệnh.
		QwenAudio-2-7B-Instruct	Trong 100.000 tế bào, là nguyên nhân gây bệnh.
		MMST	trong 100.000 tế bào như nguyên nhân gây bệnh.
	→fr	Ground Truth	dans 100 000 cellules comme cause de la maladie.
		QwenAudio-2-7B-Instruct	Dans 100 000 cellules, en tant que cause de la maladie.
		MMST	dans 100 000 cellules comme cause de la maladie.
	→zh	Ground Truth	在10 万个细胞中作为疾病的致病原因。
		QwenAudio-2-7B-Instruct	在一百万个细胞中，作为疾病的原因。
		MMST	在100,000 个细胞中作为疾病的诱因。
	→de	Ground Truth	in 100.000 Zellen als Ursache der Krankheit.
		QwenAudio-2-7B-Instruct	In 100.000 Zellen als Erreger der Krankheit.
		MMST	in 100.000 Zellen als Ursache der Krankheit.
vi	→en	Ground Truth	Please listen to the doctor's advice, I am very concerned about you today
		QwenAudio-2-7B-Instruct	Please listen to the doctor's question, I am very busy today.
		MMST	Please listen to the doctor's advice together.
	→fr	Ground Truth	Veuillez écouter les conseils du médecin, aujourd'hui je suis très
		QwenAudio-2-7B-Instruct	Madame, pouvez-vous également écouter la question du médecin ? Aujourd'hui, je suis très occupé.
		MMST	Veuillez écouter les conseils du médecin. Madame, je suis très inquiet pour vous aujourd'hui.
	→zh	Ground Truth	请您听听医生的建议，今天我很关心您
		QwenAudio-2-7B-Instruct	请允许我听从医生的建议，今天我非常高兴
		MMST	请您一起听医生的建议，今天的您，我非常关心。
	→de	Ground Truth	Bitte hören Sie sich den Rat des Arztes an, ich bin heute sehr besorgt um Sie
		QwenAudio-2-7B-Instruct	Sprechen Sie mit der Schwester, hören Sie auf die Frage des Arztes? Heute möchte ich Ihnen helfen.
		MMST	Bitte hören Sie sich die Ratschläge des Arztes an. Frau, ich bin heute sehr besorgt um Sie.
fr	→en	Ground Truth	Okay, thank you. Would you like to schedule the next appointment now?
		QwenAudio-2-7B-Instruct	Okay, thank you. Would you like to take my phone call now?
		MMST	Thank you. Would you like us to book the following appointment now?
	→vi	Ground Truth	Được rồi, cảm ơn. Bạn có muốn đặt lịch hẹn tiếp theo ngay bây giờ không?
		QwenAudio-2-7B-Instruct	Cảm ơn bạn. Bạn có muốn chúng tôi đưa ông ấy đến phòng khám hiện tại không?
		MMST	Cảm ơn. Bạn có muốn chúng tôi tiếp nhận lịch hẹn sau này không?
	→zh	Ground Truth	好的，谢谢。您想现在预约下一个就诊时间吗?
		QwenAudio-2-7B-Instruct	谢谢。希望您现在能接受我的预约吗?
		MMST	好的，谢谢，请您现在预约一下这些时间。
	→de	Ground Truth	Okay, danke. Möchten Sie den nächsten Termin jetzt vereinbaren?
		QwenAudio-2-7B-Instruct	Danke. Möchten Sie, dass ich Ihnen jetzt den Termin mit Herrn Sylvestre geben kann?
		MMST	Danke. Sollen wir jetzt den folgenden Termin vereinbaren?
de	→en	Ground Truth	Yes, please come back to me in two days. We will see if the symptoms persist after these two days.
		QwenAudio-2-7B-Instruct	Yes, please come back to me in two days. We will see if the symptoms persist after these two days.
		MMST	Yes, please come back to me in two days. We will see if the symptoms persist after these two days.
	→vi	Ground Truth	Vâng, xin vui lòng quay lại gặp tôi trong hai ngày. Chúng tôi sẽ xem liệu các triệu chứng có còn kéo dài sau hai ngày này hay không.
		QwenAudio-2-7B-Instruct	Vâng, xin hãy quay lại với tôi trong hai ngày. Chúng ta sẽ xem liệu các triệu chứng của bạn có còn tiếp tục sau hai ngày không?
		MMST	Vâng, hãy quay lại tôi sau hai ngày. Chúng tôi sẽ xem liệu các triệu chứng sau hai ngày này còn kéo dài hay không.
	→fr	Ground Truth	Oui, veuillez revenir me voir dans deux jours. Nous verrons si les symptômes persistent après ces deux jours.
		QwenAudio-2-7B-Instruct	Oui, veuillez revenir me voir dans deux jours. Nous verrons si les symptômes persistent après ces deux jours.
		MMST	Oui, veuillez revenir me voir dans deux jours. Nous verrons si les symptômes persistent après ces deux jours.
	→zh	Ground Truth	是的，请您两天后再来找我。我们将看看两天后您的症状是否还会持续。
		QwenAudio-2-7B-Instruct	是的，请您在两天后再来找我。我们将看看这些症状是否在接下来的两天内仍然存在。
		MMST	是的，两天后再来看我。我们会看看两天后症状是否会持续。
zh	→en	Ground Truth	Besides work, we also need to take care of other patients.
		QwenAudio-2-7B-Instruct	That is, after work, we also take care of other patients.
		MMST	Besides working, we also have to take care of other patients.
	→vi	Ground Truth	ngoài việc đi làm còn phải chăm sóc những bệnh nhân khác.
		QwenAudio-2-7B-Instruct	Đó là bạn đi làm ngoài giờ, chúng tôi còn phải chăm sóc cho các bệnh nhân khác sao?
		MMST	Chúng tôi phải chăm sóc bệnh nhân khác ngoài việc làm việc.
	→fr	Ground Truth	En plus du travail, nous devons aussi prendre soin d'autres patients.
		QwenAudio-2-7B-Instruct	C'est-à-dire que vous travaillez en dehors du travail, nous devons également prendre soin d'autres patients.
		MMST	C'est en plus de notre travail, nous devons également prendre soin d'autres patients.
	→de	Ground Truth	Neben der Arbeit müssen wir auch andere Patienten betreuen.
		QwenAudio-2-7B-Instruct	Wenn Sie außerhalb der Arbeit sind, müssen wir auch andere Patienten betreuen.
		MMST	Es ist einfach so, dass wir außerhalb der Arbeit auch andere Patienten behandeln müssen.