



MultiMed-ST: Large-scale Many-to-many Multilingual Medical Speech Translation

Khai Le-Duc^{*1,2,3}, Tuyen Tran^{*3,4},

Bach Phan Tat⁵, Nguyen Kim Hai Bui⁶, Quan Dang⁴, Hung-Phong Tran⁴,
Thanh-Thuy Nguyen⁷, Ly Nguyen⁸, Tuan-Minh Phan⁹, Thi Thu Phuong Tran¹⁰,
Chris Ngo³, Nguyen X. Khanh^{♡11}, Thanh Nguyen-Tang^{♡†12}

¹University of Toronto, Canada ²University Health Network, Canada

³Knovel Engineering Lab, Singapore ⁴Hanoi University of Science and Technology, Vietnam

⁵KU Leuven, Belgium ⁶Eötvös Loránd University, Hungary

⁷HCMC Open University, Vietnam ⁸IESEG School of Management, France

⁹Technische Universität Dortmund, Germany ¹⁰University of Hertfordshire, United Kingdom

¹¹UC Berkeley, United States ¹²New Jersey Institute of Technology, United States

✉ duckhai.le@mail.utoronto.ca



<https://github.com/leduckhai/MultiMed-ST>

Abstract

Multilingual speech translation (ST) and machine translation (MT) in the medical domain enhances patient care by enabling efficient communication across language barriers, alleviating specialized workforce shortages, and facilitating improved diagnosis and treatment, particularly during pandemics. In this work, we present the *first* systematic study on medical ST, to our best knowledge, by releasing  **MultiMed-ST**, a large-scale ST dataset for the medical domain, spanning *all* translation directions in five languages: Vietnamese, English, German, French, and Simplified/Traditional Chinese, together with the models. With 290,000 samples, this is **the largest medical MT dataset** and **the largest many-to-many multilingual ST among all domains**. Secondly, we present **the most comprehensive ST analysis in the field's history**, to our best knowledge, including: empirical baselines, bilingual-multilingual comparative study, end-to-end vs. cascaded comparative study, task-specific vs. multi-task sequence-to-sequence comparative study, code-switch analysis, and quantitative-qualitative error analysis. All code, data, and models are available online: <https://github.com/leduckhai/MultiMed-ST>.

1 Introduction

Effective communication between healthcare providers and patients is a foundation of quality

medical care. However, linguistic barriers often hinder this communication, especially in multicultural and multilingual settings. These barriers can lead to misdiagnoses, improper treatment, and diminished patient satisfaction, ultimately compromising the overall quality of care (Al Shamsi et al., 2020; Woloshin et al., 1995; Cohen et al., 2005; Zhang and Gao, 2024).

Medical Speech Translation (ST), also known as ST in the medical domain, is a solution aimed at bridging these linguistic divides, by enabling (near) real-time communication between speakers of different languages. The demand for medical ST has grown significantly with the increasing globalization of healthcare (Karwacka, 2015; Khoong and Rodriguez, 2022). Whether addressing the needs of immigrant populations, international patients seeking specialized treatments, or global health crises requiring cross-border collaboration, these technologies have the potential to transform how medical professionals deliver care (Dempere, 2023; Swaminathan et al., 2023; Zhang et al., 2021b). Additionally, medical ST aligns with broader efforts to promote health equity and accessibility, ensuring that language differences do not impede the right to quality healthcare (Nurminen and Koponen, 2020; Dahal and Aoun, 2023).

Since the advent of large-scale pre-trained models adaptable to domain-specific tasks (Radford et al., 2022; Chu et al., 2023; Touvron et al., 2023), medical ST research has gained attention. However, the scarcity of such publicly available datasets and models, driven by privacy concerns, hinders real-world deployment. Existing publicly available

^(*)Equal contribution

^(♡)Equal advising

^(†)Done partly while at Johns Hopkins University

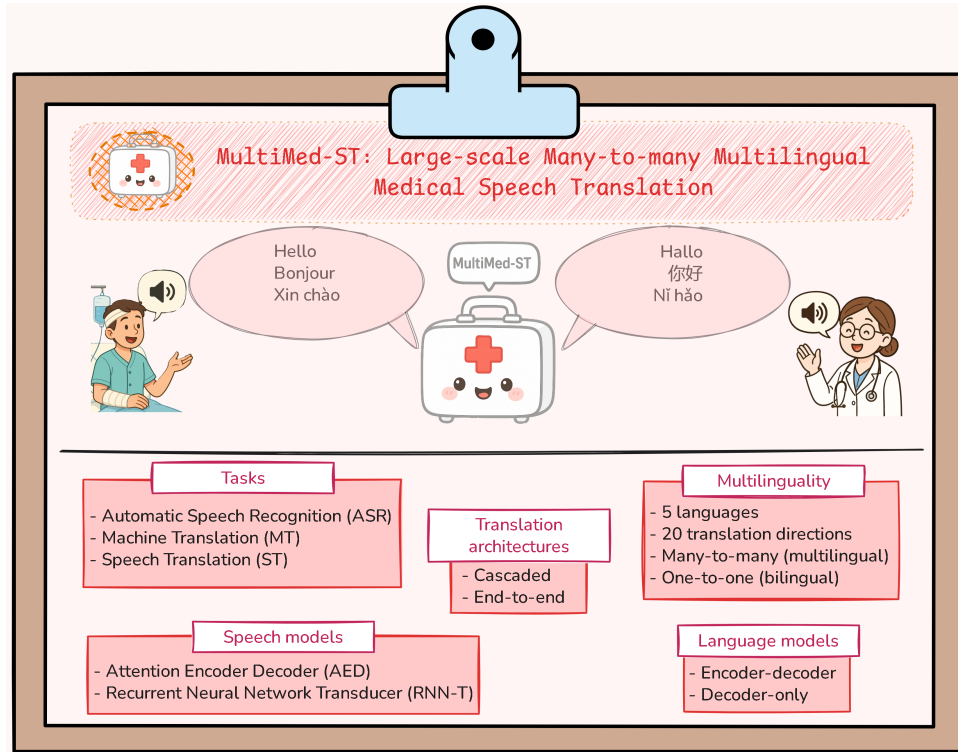


Figure 1: An overview of **MultiMed-ST** – A large-scale, many-to-many multilingual medical speech translation framework and dataset for facilitating cross-lingual communication in healthcare settings.

medical Machine Translation (MT)¹ datasets are text-only, small, and crawled from the Internet (see Appendix Table 12). For medical ST, previous works simply introduced the development of translation software without publishing datasets, models, or key findings, lacking a systematic and rigorous scientific approach (Bouillon et al., 2008; Marais et al., 2020; Xu et al., 2024).

To address the aforementioned issues, we introduce a large-scale high-quality, diverse dataset for many-to-many multilingual medical ST, supporting 5 languages: Vietnamese, English, German, French, and Mandarin Chinese. Our key contributions are:

- We present the *first* systematic medical ST study, to the best of our knowledge, by releasing **MultiMed-ST** - a large-scale many-to-many multilingual medical ST dataset for 5 languages, along with fine-tuned models. Built upon *MultiMed* Automatic Speech Recognition (ASR) dataset, our translation annotation is the largest medical MT dataset and the largest many-to-many multilingual ST

among all domains (see Section 2.3).

- We present the most extensive analysis ever conducted in ST research to date, only enabled by the large-scale, many-to-many nature of **MultiMed-ST** fine-tuning. It includes: (i) empirical baselines, (ii) task-specific vs. multi-task sequence-to-sequence (seq2seq) comparative study, (iii) end-to-end vs. cascaded comparative study, (iv) bilingual-multilingual comparative study, (v) code-switch analysis, and (vi) quantitative-qualitative error analysis. Our comprehensive analysis reveals guideline on how to build an effective many-to-many multilingual medical ST model from a task-centric, model-centric, data-centric, and linguistic-centric perspective (see Section 7 for the **five key findings**).

All code, data and models are published online.

2 Data

2.1 Data Collection

Speech data were sourced from the medical ASR dataset provided by Le-Duc et al. (2024), under

¹By definition, MT encompasses both text-to-text translation (text-only MT) and speech-to-text translation (ST). As such, ST is considered a subset of MT.

the scientific research license. This dataset comprises manually transcribed recordings of real-world multi-speaker medical conversations across five languages: Vietnamese, English, German, French, and Mandarin Chinese. As pointed out by the authors, it represents the largest and most diverse medical ASR resource, based on total duration (150 hours), number of recording conditions (10), number of accents (16), number of speaking roles (6), number of unique medical terms, and inclusion of all ICD-10 codes (see Table 11 in Appendix Section C).

2.2 Annotation Process and Data Quality Control

The data were initially translated from the source language into target languages (many-to-many) using the Gemini Large Language Model (LLM). Following the annotation process by Zheng et al. (2023), the LLM-generated translated transcripts were treated as outputs from a *real* human annotator. In the data quality process of the **test set**, five human annotators manually corrected and then cross-verified *all* these translations based on the context of the whole conversation. To remove bias from LLM-generated translations, only transcripts that received consensus approval from multiple annotators were retained, resulting in an **inter-annotator agreement of 100%**. Roughly 90% of LLM translations need correction by our annotators. The estimated² labor cost for the entire data quality process is 29k ~ 58k USD.

All human annotators possessed a professional language proficiency of C1 or higher (or HSK5 for Chinese) in their respective working languages. Additionally, each annotator had completed basic medical training and demonstrated substantial knowledge of medical terminology in their selected language. Furthermore, they were either currently pursuing or had completed undergraduate or graduate studies in countries where their chosen language is predominantly spoken.

The dataset was subsequently uploaded to the Hugging Face platform.

2.3 Data Statistics

The statistics of our data are described in Table 1. Our dataset has a total number of 290k samples for *all* directions.

²Based on the publicly available price provided by professional translation services like VerboLabs or GTE Localize. We are not permitted to provide the true amount.

Language		vi→X	en→X	de→X	fr→X	zh→X
#Samples	Train	4k5	25k5	1k4	1k4	1k2
	Dev	1k1	2k8	300	40	90
	Test	3k4	4k8	1k1	300	200
	All	9k1	33k1	2k8	1k8	1k6
Med. length	→vi	70	140	180	160	250
	→en	90	150	160	150	250
	→de	110	170	180	180	300
	→fr	100	160	200	140	290
	→zh	30	50	50	40	80

Table 1: **Statistics of our  MultiMed-ST dataset.**

In total, our dataset has **290k samples (utterances)** for **all directions of 5 languages**: Vietnamese (vi), English (en), German (de), French (fr), and both traditional and simplified Chinese (zh).

Median text length is calculated based on the number of characters.

Dataset	#Samples	Lang.	Direction
Neves (2017)	46k	2	one-to-one
ParaMed (Liu and Huang, 2021)	200k	2	one-to-one
Khresmoi (Pecina et al., 2017)	12k	8	many-to-many
WMT Biomed. (Bawden et al., 2020)	160k	9	one-to-one
YuQ (Yu et al., 2020)	130k	2	one-to-one
Bérard et al. (2020)	1k5	2	one-to-one
MedEV (Vo et al., 2024)	36k	2	one-to-one
 MultiMed-ST (ours)	290k	5	many-to-many

Table 2: **Dataset comparison with literature.** All publicly available datasets listed here are *text-only* medical MT. **Our  MultiMed-ST is the first medical ST dataset, and is the largest medical MT dataset.**

 Full details are shown in Table 12 in Appendix Section C.

To the best of our knowledge,  **MultiMed-ST is the largest medical MT dataset** when compared to existing medical MT datasets, as shown in Table 2, although speech data is much more difficult to collect and annotate.

Besides, in comparison with other large-scale ST datasets reported in the literature, the size of  **MultiMed-ST** is comparable ( see Table 13 in Appendix Section C). However,  **MultiMed-ST is the largest many-to-many multilingual ST among all domains.**

3 Problem Formulation

Informal definition: An ST model aims to convert an audio signal to a translated language sequence. A **cascaded ST** approach first transcribes speech to text (ASR) and then translates it using a separate MT model, while **end-to-end ST** directly converts speech in one language to text in another without intermediate transcription.

Formal definition: Given an audio signal $x_1^T := x_1, x_2, \dots, x_T$ of T audio frames, a source language sequence f_1^J of J words, and a target language se-

quence e_1^I of I words, the maximization of the posterior probability p of the target language sequence given the speech input is described as:

End-to-end approach:

$$x_1^T \rightarrow \hat{e}_1^I(x_1^T) = \arg \max_{I, e_1^I} p(e_1^I | x_1^T) \quad (1)$$

where $\hat{e}_1^I$ of length $\hat{I}$ words is the best target language sequence, and $\rightarrow$ is a mapping.

Cascaded approach:

$$x_1^T \rightarrow \hat{f}_1^J(x_1^T) = \arg \max_{J, f_1^J} p(f_1^J | x_1^T) \quad (2)$$

$$\hat{f}_1^J \rightarrow \hat{e}_1^I(\hat{f}_1^J) = \arg \max_{I, e_1^I} p(e_1^I | \hat{f}_1^J) \quad (3)$$

where Equation 2 is an ASR model that transcribes the speech signal into the best source language sequence $\hat{f}_1^J$, while Equation 3 is an MT model that generates the best target language sequence given the predicted source language sequence.

Further details of problem formulation are shown in Appendix Section B.

4 Experimental Setup

We first establish empirical baselines, then derive key insights from task (task-specific vs. multi-task), model (end-to-end vs. cascaded), data (bilingual vs. multilingual training), and linguistic (code-switching analysis) perspectives.

4.1 Training Setup

Training system: We employed two standard training systems, **cascaded** (ASR $\rightarrow$ MT) and **end-to-end**.

ASR models: We employed the 2 most state-of-the-art (SOTA) ASR architectures with varying model sizes.

- Attention Encoder Decoder (AED):
 - Whisper models (Radford et al., 2023): Whisper-small³, Whisper-large-v2⁴
 - Deepgram⁵
- Recurrent Neural Network Transducer (RNN-T): AssemblyAI⁶

³<https://huggingface.co/openai/whisper-small>

⁴<https://huggingface.co/openai/whisper-large-v2>

⁵<https://deepgram.com/>

⁶<https://www.assemblyai.com/>

MT models: We employed various SOTA open-source/closed-source, task-specific/multitask seq2seq architectures and data representations.

- Multilingual pre-trained models:
 - Encoder-decoder: mBART-large-50⁷ (Tang et al., 2020), M2M100-418M⁸ (Fan et al., 2020), Marian⁹ (Tiedemann and Thottingal, 2020)
 - Decoder: Llama-3.1-8B¹⁰ (Dubey et al., 2024), Qwen-2.5-7B¹¹ (Yang et al., 2024a), Mistral-v0.3-7B¹² (Jiang et al., 2023)
 - Commercial tool: Google Translate¹³
- Bilingual pre-trained models: VinAI Translate¹⁴ (Nguyen et al., 2022), EnViT5¹⁵ (Ngo et al., 2022)

End-to-end ST models: For direct speech-to-text translation, we employed Whisper, SeamlessM4T-large-v2¹⁶ (Communication et al., 2023b,a), Qwen2-Audio-7B-Instruct¹⁷ (Chu et al., 2024, 2023).

All ASR and MT models are general-domain since  **MultiMed-ST** is the first attempt to fine-tune medical domain ST. Full details of the training setup are shown in Appendix Section D.

4.2 Evaluation Metrics

Advantage/disadvantage discussion of automatic metrics is described in Appendix Section E.1.

Automatic MT metrics: To evaluate MT quality, two standard categories of evaluation metrics were utilized: **n-gram overlap metrics** (e.g., BLEU (Papineni et al., 2002), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005), ChrF (Popović, 2015), ROUGE (Lin, 2004)) and **embedding-based metrics** (e.g., BERTScore (Zhang et al.)).

⁷<https://huggingface.co/facebook/mbart-large-50>

⁸https://huggingface.co/facebook/m2m100_418M

⁹https://huggingface.co/docs/transformers/model_doc/marian

¹⁰<https://huggingface.co/meta-llama/Llama-3.1-8B>

¹¹<https://huggingface.co/Qwen/Qwen2.5-7B>

¹²<https://huggingface.co/mistralai/Mistral-7B-v0.3>

¹³<https://cloud.google.com/translate/docs>

¹⁴<https://huggingface.co/vinai/vinai-translate-vi2en>

<https://huggingface.co/vinai/vinai-translate-en2vi>

¹⁵<https://huggingface.co/VietAI/envit5-base>

¹⁶<https://huggingface.co/facebook/seamless-m4t-v2-large>

¹⁷<https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct>

MT	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
Decoder																					
Llama-3.1-8B	BLEU	53.44	48.24	37.50	40.49	23.16	15.57	16.09	11.61	50.18	39.63	29.25	27.46	49.44	40.01	33.45	31.16	28.21	23.49	18.87	13.07
	BERTSc	0.90	0.89	0.83	0.87	0.92	0.79	0.74	0.77	0.95	0.86	0.79	0.81	0.95	0.87	0.84	0.81	0.91	0.79	0.77	0.74
Qwen-2.5-7B	BLEU	54.50	49.63	28.61	38.75	26.21	19.25	29.06	14.44	49.69	40.67	20.97	33.91	52.10	43.73	40.72	23.26	35.63	32.95	24.05	16.95
	BERTSc	0.90	0.90	0.81	0.87	0.93	0.81	0.81	0.79	0.95	0.86	0.78	0.84	0.96	0.88	0.88	0.79	0.95	0.85	0.84	0.83
Mistral-v0.3-7B	BLEU	24.77	51.71	26.38	43.99	24.56	16.00	25.04	13.38	34.95	14.47	19.92	33.73	36.39	15.68	40.77	21.28	27.68	10.67	18.46	11.40
	BERTSc	0.82	0.89	0.81	0.88	0.92	0.79	0.78	0.78	0.91	0.79	0.78	0.85	0.92	0.79	0.86	0.78	0.93	0.75	0.80	0.76
Encoder-decoder																					
mBart-large-50	BLEU	59.73	56.23	44.77	46.48	16.48	12.61	22.97	10.43	39.58	36.17	24.63	28.73	41.45	41.12	40.48	30.43	15.03	14.26	15.70	10.67
	BERTSc	0.92	0.92	0.86	0.89	0.89	0.80	0.78	0.75	0.93	0.86	0.77	0.83	0.94	0.87	0.87	0.80	0.90	0.82	0.79	0.77
M2M100-418M	BLEU	62.31	57.49	46.38	49.36	23.01	21.10	24.95	16.72	43.73	35.04	29.41	34.72	44.76	43.83	43.53	30.42	21.65	27.69	21.88	15.17
	BERTSc	0.97	0.95	0.93	0.94	0.82	0.81	0.80	0.79	0.88	0.82	0.82	0.83	0.83	0.85	0.88	0.75	0.76	0.85	0.82	0.82
Marian	BLEU	58.22	53.84	38.67	45.81	17.63	15.97	15.56	12.84	39.97	33.41	17.13	32.62	42.74	38.26	39.59	18.11	11.44	16.14	11.33	6.24
	BERTSc	0.91	0.91	0.85	0.89	0.80	0.79	0.78	0.77	0.87	0.86	0.78	0.85	0.88	0.87	0.87	0.78	0.78	0.79	0.77	0.75
Commercial tool																					
Google Translate	BLEU	59.50	59.28	57.13	49.12	28.62	25.25	31.24	19.00	47.47	39.28	39.38	38.89	53.35	42.47	43.67	40.54	39.34	44.41	29.48	24.77
	BERTSc	0.91	0.91	0.90	0.89	0.84	0.83	0.83	0.81	0.88	0.86	0.85	0.86	0.90	0.88	0.88	0.86	0.88	0.87	0.85	0.85

Table 3: **Ground-truth MT baselines.** All MT models were fine-tuned monolingually (on each respective language pair separately) except Google Translate being recognized directly on test set. *en-vi* denotes translation from *en* to *vi*. Only BLEU (n-gram overlap metric) and BERTScore (embedding-based metric) were reported in this table.

📈 Google Translate leads overall, with Encoder-decoder MT models often surpassing LLMs on many language pairs.

📄 Full results for all evaluation metrics (including other n-gram overlap metrics) are shown in Table 19 (English to X), Table 20 (Vietnamese to X), Table 21 (French to X), Table 22 (German to X), and Table 23 (Chinese to X) in Appendix Section F.2.

Red highlight: best result. Blue highlight: second-best result (Encoder-decoder models outperform Decoder-only)

ASR metrics: In the context of ST, ASR performance influences translation quality; therefore, ASR was additionally assessed using Word Error Rate (WER) and Character Error Rate (CER).

Human evaluation: Human evaluators directly assess MT outputs by grading scores (0 to 10) based on three key criteria: *adequacy*, *fluency*, and *comprehensibility* (📄 see Appendix Section E.2).

LLM-as-a-judge: Unlike automated metrics, which rely on surface-level matching of n-grams, LLM-as-a-judge (Zheng et al., 2023) can assess translations based on deeper semantic understanding, contextual appropriateness, and syntactic correctness (📄 see Appendix Section E.3 and Figure 32).

5 Experimental Results

5.1 Automatic Speech Recognition Baselines

🔍 **What are trade-offs among model sizes, fine-tuning strategies, and performance of ASR models?** As shown in Table 4, the fine-tuned Whisper-small model achieved superior performance to larger pre-trained models, consistently outperforming all models across languages on the dev set. On test set, Whisper-small achieved the best WER for Vietnamese (29.60%) and CER for Chinese (31.3%), while Whisper-large-v2 excelled in English (WER 25.5%) and Chinese (CER 37.3%), and Deepgram outperformed others in French with

ASR	dev					test				
	vi	en	zh	de	fr	vi	en	zh	de	fr
Whisper-small-mono	21.2	24.4	25.9	17.5	35.8	29.6	33.8	31.3	26.3	45.7
+ SpecAugment	19.8	23.5	43.3	17.9	44.1	31.7	36.9	46.9	24.1	45.6
Whisper-small-multi	25.7	46.1	73.9	22.2	50.6	33.4	40.9	89.8	19.6	55.3
Whisper-large-v2-mono	57.7	26.9	39.0	23.7	52.9	62.6	25.5	37.3	24.2	41.7
Assembly	51.9	31.7	49.8	27.9	49.4	65.5	30.6	45.2	28.9	42.1
Deepgram	35.8	33.9	40.4	27.8	50.7	40.0	32.1	46.7	28.4	40.3

Table 4: **ASR baseline results.** Chinese (zh) is evaluated by CER (%), while other languages are evaluated by WER (%). Whisper is fine-tuned monolingually (each language separately) or multilingually (all languages simultaneously). SpecAugment (Park et al., 2019) is tested on Whisper-small-mono as data augmentation. Commercial models like Assembly and Deepgram only allows direct recognition.

📈 Monolingual Whisper-small leads overall, while larger models excel in high-resource languages.

a WER of 40.3%, highlighting the advantage of larger models for high-resource languages.

Also, results showed that monolingual fine-tuning consistently outperforms multilingual fine-tuning on both dev and test sets. Besides, SpecAugment (Park et al., 2019) does not help accuracy improvement.

5.2 Ground-truth Translation Baselines

🔍 **Task-specific models outperform multi-task models on ground-truth transcript:** The exper-

ASR	MT	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
Ground-truth	mBart-large-50	BLEU	59.73	56.23	44.77	46.48	16.48	12.61	22.97	10.43	39.58	36.17	24.63	28.73	41.45	41.12	40.48	30.43	15.03	14.26	15.70	10.67
	M2M100	BERTSc	0.92	0.92	0.86	0.89	0.89	0.80	0.78	0.75	0.93	0.86	0.77	0.83	0.94	0.87	0.87	0.80	0.90	0.82	0.79	0.77
	M2M100	BLEU	62.31	57.49	46.38	49.36	23.01	21.10	24.95	16.72	43.73	35.04	29.41	34.72	44.76	43.83	43.53	30.42	21.65	27.69	21.88	15.17
	-418M	BERTSc	0.97	0.95	0.93	0.94	0.82	0.81	0.80	0.79	0.88	0.82	0.82	0.83	0.83	0.85	0.88	0.75	0.76	0.85	0.82	0.82
Whisper-small-mono	mBart-large-50	BLEU	48.00	43.20	35.70	35.07	10.17	12.80	16.77	7.23	23.82	22.86	16.46	17.39	31.95	32.62	31.96	25.07	11.88	18.40	12.30	9.64
	M2M100	BERTSc	0.87	0.86	0.81	0.84	0.88	0.76	0.73	0.72	0.90	0.80	0.72	0.77	0.92	0.84	0.84	0.77	0.89	0.79	0.76	0.75
	M2M100	BLEU	48.21	43.16	36.94	36.55	15.64	13.95	16.99	11.10	25.65	21.88	18.44	19.98	33.66	34.70	34.67	24.31	16.65	21.83	16.94	13.06
	-418M	BERTSc	0.95	0.92	0.92	0.92	0.78	0.77	0.74	0.75	0.81	0.76	0.75	0.76	0.77	0.81	0.85	0.72	0.76	0.83	0.79	0.78
Whisper-small-multi	mBart-large-50	BLEU	47.99	43.22	36.02	34.93	10.16	12.18	16.95	6.88	25.53	25.34	19.15	19.26	34.6	35.2	34.45	27.19	12.39	18.4	11.5	9.31
	M2M100	BERTSc	0.87	0.86	0.81	0.84	0.88	0.76	0.73	0.72	0.91	0.81	0.73	0.78	0.93	0.85	0.85	0.78	0.89	0.78	0.75	0.74
	M2M100	BLEU	48.1	42.98	36.94	36.28	14.76	13.1	17.17	10.38	27.97	24.84	22.99	22.97	37.25	37.26	37.61	27.20	14.64	21.35	15.37	11.89
	-418M	BERTSc	0.95	0.92	0.92	0.91	0.78	0.77	0.74	0.75	0.82	0.77	0.77	0.76	0.79	0.81	0.85	0.74	0.74	0.85	0.78	0.76
Whisper-large-mono	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
	M2M100	BERTSc	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
	M2M100	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
	-418M	BERTSc	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
Assembly	mBart-large-50	BLEU	51.23	45.45	40.51	37.37	8.37	11.03	14.55	4.74	29.00	26.52	18.77	19.84	33.84	34.64	32.76	28.42	4.90	10.21	7.79	4.97
	M2M100	BERTSc	0.88	0.88	0.83	0.86	0.87	0.75	0.71	0.70	0.91	0.82	0.74	0.79	0.93	0.85	0.85	0.79	0.76	0.60	0.73	0.72
	M2M100	BLEU	51.30	45.85	41.91	38.38	13.60	12.56	13.95	9.71	31.20	27.12	22.84	22.93	35.89	37.15	35.83	30.17	14.90	19.79	13.11	10.23
	-418M	BERTSc	0.95	0.93	0.92	0.92	0.77	0.76	0.69	0.75	0.83	0.78	0.77	0.76	0.77	0.81	0.86	0.77	0.77	0.83	0.78	0.77
Deepgram	mBart-large-50	BLEU	50.93	45.37	39.93	37.30	9.44	12.05	15.48	5.88	28.95	27.39	18.82	20.52	33.99	35.37	33.49	27.90	4.90	7.79	07.03	3.31
	M2M100	BERTSc	0.88	0.88	0.83	0.85	0.88	0.67	0.65	0.70	0.91	0.82	0.73	0.79	0.93	0.85	0.85	0.79	0.80	0.60	0.71	0.69
	M2M100	BLEU	51.01	45.34	41.18	38.42	15.60	14.20	16.24	11.10	31.02	28.38	24.04	22.75	36.02	37.55	36.45	28.78	13.47	16.50	12.40	9.57
	-418M	BERTSc	0.95	0.93	0.92	0.92	0.76	0.76	0.73	0.74	0.82	0.77	0.78	0.74	0.79	0.82	0.86	0.75	0.76	0.82	0.76	0.76

Table 5: **Cascaded ST baseline results.** The effect of ASR models on MT quality is compared with MT on ground-truth text. Monolingual translation fine-tuning refers to fine-tuning MT models on each language pair separately, while multilingual translation fine-tuning refers to fine-tuning MT models on all language pairs simultaneously.

🔴 Whisper-large-v2 with M2M100-418M achieved the best overall ST performance, except for Vietnamese where Whisper-small-mono was superior.

📌 Extra results for all evaluation metrics and models are shown in Table 24 (English to X), Table 25 (Vietnamese to X), Table 26 (French to X), Table 27 (German to X), and Table 28 (Chinese to X) in Appendix Section F.3.

imental results for MT on ground-truth transcript are presented in Table 3. Overall, translations from Google Translate achieved the highest results and outperformed other models across most language pairs in both settings. Encoder-decoder models, particularly those with English as the source language, generally outperformed the decoder models (LLMs). Notably, the M2M100-418M model recorded higher BLEU scores than the LLMs on many language pairs. This demonstrates the effectiveness of models trained for specific MT tasks compared to multi-task models like LLMs.

5.3 Cascaded Speech Translation Baselines

🔴 **Multi-task models are on par with task-specific models in the ST setting.** We evaluated the impact of ASR models on text-to-text MT models, as shown in Table 5.

Specifically, Whisper-large-v2 - M2M100-418M achieved the highest performance on most language pairs (16/20), except for the language pair with Vietnamese as the source language, where Whisper-small-mono - M2M100-418M achieved the best performance. This outcome stems from two fac-

tors: Whisper-large-v2’s size and generalization enable more accurate transcripts for most languages, aiding MT model, while Whisper-small-mono outperforms it for Vietnamese.

ASR model performance differences reveal how ASR transcript quality impacts MT, with minor errors notably affecting complex languages like Vietnamese. Despite M2M100-418M’s robustness on ground-truth text, it is sensitive to ASR transcript quality. Also, M2M100-418M and mBart-large-50 do not significantly outperform LLMs in the cascaded ST setting, as shown in Table 6. Therefore, multi-task models (LLMs) still perform as well as task-specific models trained for MT task.

5.4 End-to-end and Cascaded Comparison

🔴 **MT accuracy is dropped on speech:** Table 3 and Table 6 show a significant decline in both BLEU and BERT scores due to the non-standard input text across all models, with the largest drop observed in the French-to-English from 50.18 to 30.15 with the LLama-3.1-8B model. This indicates that the ASR model’s poor performance for French significantly reduced translation accuracy.

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
Cascaded																					
Llama	BLEU	43.32	37.92	30.78	31.36	14.55	10.29	11.56	7.71	30.15	25.36	20.28	16.38	40.63	33.63	26.97	26.31	19.01	17.65	13.84	11.13
-3.1-8B	BERTSc	0.85	0.84	0.8	0.83	0.78	0.75	0.73	0.73	0.82	0.80	0.75	0.74	0.86	0.84	0.80	0.79	0.79	0.85	0.76	0.74
Qwen	BLEU	43.37	37.34	23.46	28.5	13.97	11.66	20.27	8.75	30.35	25.59	15.33	20.38	40.52	34.24	31.45	19.87	25.36	26.31	17.84	12.61
-2.5-7B	BERTSc	0.85	0.85	0.8	0.82	0.78	0.76	0.78	0.75	0.81	0.80	0.76	0.78	0.86	0.84	0.84	0.79	0.82	0.90	0.80	0.78
Mistral	BLEU	17.72	36.58	20.27	29.9	15.86	10.92	17.92	9.03	29.35	9.20	13.94	18.65	28.33	12.38	31.15	17.82	20.17	8.01	12.58	7.14
-v0.3-7B	BERTSc	0.77	0.83	0.77	0.81	0.78	0.75	0.77	0.75	0.79	0.74	0.76	0.78	0.78	0.77	0.83	0.78	0.81	0.86	0.78	0.72
mBart	BLEU	48.00	43.20	35.70	35.07	10.17	12.80	16.77	7.23	23.82	22.86	16.46	17.39	31.95	32.62	31.96	25.07	11.88	18.40	12.30	9.64
-large-50	BERTSc	0.87	0.86	0.81	0.84	0.88	0.76	0.73	0.72	0.90	0.80	0.72	0.77	0.92	0.84	0.84	0.77	0.89	0.79	0.76	0.75
M2M100	BLEU	48.21	43.16	36.94	36.55	15.64	13.95	16.99	11.10	25.65	21.88	18.44	19.98	33.66	34.70	34.67	24.31	16.65	21.83	16.94	13.06
-418M	BERTSc	0.95	0.92	0.92	0.92	0.78	0.77	0.74	0.75	0.81	0.76	0.75	0.76	0.77	0.81	0.85	0.72	0.76	0.83	0.79	0.78
Marian	BLEU	45.07	40.54	31.17	33.90	12.95	11.23	12.09	09.08	24.03	22.20	11.27	19.14	34.09	29.72	30.48	14.79	8.50	13.37	8.39	5.76
	BERTSc	0.87	0.86	0.82	0.84	0.77	0.76	0.75	0.74	0.81	0.80	0.74	0.79	0.85	0.83	0.84	0.76	0.75	0.77	0.74	0.73
Google	BLEU	46.21	44.77	44.74	36.29	18.79	16.42	21.63	12.54	27.82	24.18	24.49	22.38	40.74	32.69	33.15	31.89	27.74	30.70	20.71	19.11
Translate	BERTSc	0.86	0.86	0.85	0.84	0.78	0.78	0.78	0.76	0.81	0.80	0.79	0.79	0.86	0.85	0.84	0.83	0.83	0.82	0.80	0.81
End-to-end																					
SeamlessM4T	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
-large-v2	BERTSc	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
QwenAudio-2	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
-7B-Instruct	BERTSc	0.8	0.82	0.76	0.79	0.66	0.65	0.65	0.66	0.81	0.74	0.71	0.74	0.8	0.73	0.76	0.72	0.8	0.78	0.77	0.77
Whisper	BLEU					8.18				26.06				37.32				16.54			
	BERTSc					0.75				0.81				0.85				0.79			

Table 6: **End-to-end and cascaded comparison.** All cascaded models use *Whisper_{small-mono}* as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately), then MT models translate into target languages. End-to-end Whisper for ST is fine-tuned bilinguallly - on each language pair separately. End-to-end Whisper ST only supports X to English, thus no results for other translation directions were reported.

📈 Cascaded models significantly outperform end- to-end models.

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
Multilingual MT fine-tuning																					
Llama	BLEU	41.79	36.14	32.71	28.19	15.41	10.71	19.55	8.33	27.47	21.63	18.05	17.40	36.47	27.5	27.06	25.05	20.48	21.52	15.37	10.64
-3.1-8B	BERTSc	0.85	0.84	0.82	0.82	0.78	0.76	0.78	0.74	0.81	0.79	0.77	0.78	0.85	0.82	0.83	0.80	0.80	0.79	0.77	0.77
Qwen	BLEU	41.71	36.39	32.78	27.89	15.11	10.55	19.58	7.80	27.56	22.09	19.06	17.69	36.05	26.27	27.36	25.11	20.62	21.37	15.51	10.47
-2.5-7B	BERTSc	0.85	0.84	0.82	0.82	0.78	0.76	0.78	0.74	0.81	0.79	0.77	0.78	0.85	0.82	0.83	0.80	0.80	0.79	0.78	0.76
Mistral	BLEU	19.09	35.89	20.22	28.83	15.4	10.7	16.83	8.61	27.95	9.83	13.59	16.18	37.82	11.42	21.13	15.37	21.07	9.21	13.02	9.14
-v0.3-7B	BERTSc	0.8	0.84	0.79	0.83	0.78	0.75	0.77	0.74	0.82	0.75	0.76	0.78	0.86	0.77	0.81	0.72	0.81	0.73	0.76	0.76
Bilingual MT fine-tuning																					
Llama	BLEU	43.32	37.92	30.78	31.36	14.55	10.29	11.56	7.71	30.15	25.36	20.28	16.38	40.63	33.63	26.97	26.31	19.01	17.65	13.84	11.13
-3.1-8B	BERTSc	0.85	0.84	0.8	0.83	0.78	0.75	0.73	0.73	0.82	0.80	0.75	0.74	0.86	0.84	0.80	0.79	0.79	0.85	0.76	0.74
Qwen	BLEU	43.37	37.34	23.46	28.5	13.97	11.66	20.27	8.75	30.35	25.59	15.33	20.38	40.52	34.24	31.45	19.87	25.36	26.31	17.84	12.61
-2.5-7B	BERTSc	0.85	0.85	0.8	0.82	0.78	0.76	0.78	0.75	0.81	0.80	0.76	0.78	0.86	0.84	0.84	0.79	0.82	0.90	0.80	0.78
Mistral	BLEU	17.72	36.58	20.27	29.9	15.86	10.92	17.92	9.03	29.35	9.20	13.94	18.65	28.33	12.38	31.15	17.82	20.17	8.01	12.58	7.14
-v0.3-7B	BERTSc	0.77	0.83	0.77	0.81	0.78	0.75	0.77	0.75	0.79	0.74	0.76	0.78	0.78	0.77	0.83	0.78	0.81	0.86	0.78	0.72

Table 7: **Bilingual-multilingual fine-tuning comparison.** All ST results are from cascaded ST models with ASR transcript generated by Whisper Small fine-tuned monolingually on source language.

📈 Overall, Bilingual fine-tuning outperforms multilingual MT fine-tuning.

A similar trend was also observed in in-context learning experiments (📌 see Appendix Section F.1).

📌 **Cascaded models significantly outperform end-to-end models:** Table 6 compares cascaded models with end-to-end models. The results show a significant performance gap, with most cascaded models significantly outperforming end-to-end models. For a fair comparison with general-domain ST in the literature, our findings align with prior insights that end-to-end models require extensive data (probably thousands of hours) and numerous parameters to match the accuracy of cascaded models (Sperber and Paulik, 2020; Sperber et al., 2019; Xue et al., 2022).

5.5 Bilingual-Multilingual Fine-tuning Comparison

📌 **Bilingual fine-tuning outperforms multilingual MT fine-tuning:** As shown in Table 7, fine-tuning MT models on all language pairs simultaneously resulted in a degradation of performance for most language pairs compared to fine-tuning on each language pair separately. When fine-tuning on multiple language pairs simultaneously, the shared parameters of the model must allocate their representational capacity across all pairs. This leads to interference between language pairs, especially when their linguistic structures or vocabularies differ significantly, as also observed in general-domain MT (Dabre et al., 2020; Blackwood et al.,

MT	Metrics	Ground-truth		ASR	
		en-vi	vi-en	en-vi	vi-en
Bilingual pre-trained MT					
VinAI	BLEU	65.85	28.55	50.79	15.46
	BERTSc	0.93	0.84	0.88	0.77
EnViT5	BLEU	20.72	23.46	17.26	15.16
	BERTSc	0.83	0.82	0.80	0.78
Multilingual pre-trained MT					
mBart -large-50	BLEU	59.73	16.48	48.00	10.17
	BERTSc	0.92	0.89	0.87	0.88
M2M100 -418M	BLEU	62.31	23.01	48.21	15.64
	BERTSc	0.97	0.82	0.95	0.78

Table 8: **Bilingual-multilingual pre-training comparison.** All ST results are from cascaded ST models with ASR transcript generated by Whisper Small fine-tuned monolingually on source language.  Multilingual MT models perform on par with bilingual ones.

2018).

5.6 Bilingual-Multilingual Pre-training Comparison

 **Multilingual pre-trained MT models match bilingual accuracy without needing multiple language-pair variants:** As shown in Table 8, the VinAI model achieved the highest BLEU score (50.79) for English-to-Vietnamese, while the M2M100-418M model excelled in BERTScore (0.95 vs. 0.88 for VinAI). For Vietnamese-to-English, M2M100-418M slightly outperformed VinAI with BLEU scores of 15.64 and 15.46, respectively. The EnViT5 model performed poorly for both translation directions.

These results show that bilingual pre-trained MT models do not consistently outperform multilingual ones. This findings underscores the advantage of multilingual ones in leveraging diverse language pairs to achieve acceptable overall performance across metrics without requiring multiple variants for each language pair, as also observed in general-domain MT (Dabre et al., 2020; Team et al., 2024; Maillard et al., 2023).

5.7 Code-Switch Analysis

In the medical domain, it is common for English terms or keywords to be retained in their original form when translated into other languages, a phenomenon referred to as code-switching. In Table 9, this study filtered code-switched sentences for Vietnamese, German, French, and Chinese, evaluating model performance with BLEU and BERTScore metrics for each language pair.

 **Multilingual pre-trained MT models could**

MT	Metrics	Ground-truth				ASR			
		en-vi	en-fr	en-zh	en-de	en-vi	en-fr	en-zh	en-de
Decoder									
Llama -3.1-8B	BLEU	51.92	51.12	39.42	39.96	41.68	38.21	33.02	30.49
	BERTSc	0.90	0.90	0.83	0.87	0.85	0.85	0.80	0.82
Qwen -2.5-7B	BLEU	51.60	50.00	29.62	37.39	41.81	36.13	24.18	27.18
	BERTSc	0.90	0.90	0.82	0.87	0.85	0.85	0.80	0.82
Mistral -v0.3-7B	BLEU	26.31	52.74	25.48	44.75	18.51	37.80	19.11	30.99
	BERTSc	0.83	0.90	0.81	0.88	0.78	0.85	0.76	0.82
Encoder-decoder									
mBart -large-50	BLEU	60.69	56.47	49.20	45.67	46.20	40.91	38.02	33.78
	BERTSc	0.92	0.92	0.88	0.89	0.87	0.86	0.84	0.84
M2M100 -418M	BLEU	61.11	57.07	52.06	48.75	46.26	41.91	39.70	35.95
	BERTSc	0.92	0.92	0.88	0.90	0.87	0.86	0.84	0.85
Marian	BLEU	56.70	53.20	43.00	43.86	43.54	38.59	33.42	32.48
	BERTSc	0.91	0.91	0.86	0.89	0.87	0.86	0.82	0.84

Table 9: **Code-switch analysis.** All ST results are from cascaded ST models with ASR transcript generated by Whisper Small fine-tuned monolingually on source language. The original dataset shows code-switching percentages of 11.2%, 7%, 7.9%, and 12.8% for Vietnamese, French, Chinese, and German, respectively.

handle orthographic differences in code-switch

ST: Generally, results from code-switching in Table 9 are not consistently lower or higher than ground-truth baselines (Table 3) and cascaded monolingual fine-tuning ST baselines (Table 5). The results show that multilingual pre-trained MT models can process multiple languages simultaneously within a single context, even with large orthographic differences like English-Chinese or smaller orthographic differences like English-Vietnamese/German.

6 Error Analysis

6.1 Quantitative Error Analysis

 **Strong correlation between n-gram overlap, contextual-embedding and subjective evaluation:** As shown in Table 10, for most language pairs and MT models, there was a strong correlation between n-gram overlap metric and embedding-based metric and the evaluation outcomes obtained from both subjective LLM-as-a-judge and subjective human evaluations in ST quality. This alignment suggests that traditional automatic metrics remain reliable indicators of ST quality, even as evaluation methodologies evolve. The consistency across these metrics reinforces their validity in assessing *adequacy*, *fluency* and *comprehensibility* of medical ST - the phenomenon is sometimes seen in general-domain MT (Zheng et al., 2023; Zhang et al.; Bavaresco et al., 2024). LLM-as-a-judge is a newly explored research trend, thus we found no reference for ST, to our best knowledge.

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
Llama -3.1-8B	BLEU	41.79	36.14	32.71	28.19	15.41	10.71	19.55	8.33	27.47	21.63	18.05	17.40	36.47	27.50	27.06	25.05	20.48	21.52	15.37	10.64
	BERTSc	0.85	0.84	0.82	0.82	0.78	0.76	0.78	0.74	0.81	0.79	0.77	0.78	0.85	0.82	0.83	0.80	0.80	0.79	0.77	0.77
	LLM-judge	5.14	4.64	4.45	4.63	3.88	3.49	3.15	3.41	4.38	4.01	3.43	3.44	5.81	5.39	4.52	4.06	3.88	3.61	3.78	3.69
	Human	6.85	6.47	4.31	8.53	6.54	5.64	4.12	7.24	5.19	5.45	4.04	6.42	6.15	8.05	6.64	4.14	4.08	3.58	5.64	6.54
Qwen -2.5-7B	BLEU	41.71	36.39	32.78	27.89	15.11	10.55	19.58	7.80	27.56	22.09	19.06	17.69	36.05	26.27	27.36	25.11	20.62	21.37	15.51	10.47
	BERTSc	0.85	0.84	0.82	0.82	0.78	0.76	0.78	0.74	0.81	0.79	0.77	0.78	0.85	0.82	0.83	0.80	0.80	0.79	0.78	0.76
	LLM-judge	4.93	4.91	3.46	4.52	4.04	3.91	4.05	3.48	4.39	4.10	3.62	3.96	6.17	5.51	5.52	3.99	4.97	4.50	4.67	4.36
	Human	7.97	6.56	4.41	8.55	6.72	5.73	4.17	7.42	7.50	5.51	4.06	7.39	8.30	7.71	6.57	4.22	7.69	5.09	5.86	7.58
Mistral -v0.3-7B	BLEU	19.09	35.89	20.22	28.83	15.40	10.70	16.83	8.61	27.95	9.83	13.59	16.18	37.82	11.42	21.13	15.37	21.07	9.21	13.02	9.14
	BERTSc	0.80	0.84	0.79	0.83	0.78	0.75	0.77	0.74	0.82	0.75	0.76	0.78	0.86	0.77	0.81	0.72	0.81	0.73	0.76	0.76
	LLM-judge	2.40	4.54	3.20	4.50	3.57	3.18	3.41	3.09	4.49	2.30	3.56	3.86	4.91	2.73	5.02	3.53	4.46	2.0	3.97	2.94
	Human	6.36	6.52	3.43	6.19	6.00	5.73	5.08	4.74	7.55	5.51	3.69	4.64	7.78	4.07	6.57	3.91	7.33	2.53	6.14	5.20

Table 10: **LLM-as-a-judge and human evaluation results.** All ST results are from cascaded ST models with ASR transcript generated by Whisper Small fine-tuned monolingually on source language. A BERTScore of > 0.8 is often seen as good translation quality. while > 0.9 is excellent translation quality.

 Automatic metrics (BLEU, BERTScore) strongly correlate with both LLM-as-a-judge and human evaluations across most language pairs.

6.2 Qualitative Error Analysis

We analyzed recurring translation errors in medical content across English, Vietnamese, German, Chinese, and French, identifying key areas for improvement.

With English as the source, common issues included sentence fragmentation (notably in Chinese and Vietnamese), literal idiom translation, inconsistent medical terminology, and errors in proper noun handling. Vietnamese source texts led to grammatical errors in word order, verb tense, and articles, along with imprecise word choice, omissions, and register inconsistencies. German sources showed frequent word order errors, literal idiom translations, and issues with case, gender, and verb conjugation, especially in French and Vietnamese. Chinese texts often resulted in unnatural word-for-word translations, tense inaccuracies, missing grammatical elements, and misused measure words. French exhibited similar challenges to English, including sentence fragmentation, literal idiom translation, inconsistent terminology, and Vietnamese grammar errors in word order and verb conjugation.

 More qualitative results are shown in Appendix Section F.4.

7 Conclusion

In this work, we aim to remove language barriers in healthcare by presenting the first systematic study on medical ST, to our best knowledge. Specifically, we release  **MultiMed-ST**, a large-scale ST dataset in the medical domain, covering *all* translation directions in five languages: Vietnamese, English, German, French, Simplified/Traditional

Chinese, together with the models. With 290,000 samples, our dataset is the world’s largest medical MT dataset and the largest many-to-many multilingual ST among all domains.

 Our key findings are: (1) Although task-specific models surpass multi-task models when evaluated on ground-truth transcripts, both exhibit comparable performance in the medical ST setting. (2) Cascaded models still significantly outperform end-to-end models. (3) In the medical cascaded ST, multilingual pre-trained MT models should be selected for bilingual fine-tuning on each language pair for two primary reasons: first, multilingual pre-trained MT models achieve bilingual accuracy without the need for multiple separate language-pair variants; second, bilingual fine-tuning has been shown to outperform multilingual MT fine-tuning. (4) Multilingual pre-trained MT models are capable of handling orthographic differences in code-switching with comparable effectiveness to non-code-switching in medical ST. (5) In medical ST, n-gram overlap evaluation exhibits a strong correlation with both contextual embedding-based evaluation and subjective assessment.

8 Limitations

Science and religion always go hand in hand. Carelessness in science can lead to serious consequences - not to mention the karmic repercussions researchers may face under the law of karma in Buddhism. Despite our best efforts to minimize human errors, mistakes in data, experiments, and processes are inevitable and often beyond our understanding or control.

Medical research is a matter of great importance,

as it can have direct negative impacts on human health. Given the critical nature of medical transcription (see Appendix Section G.3), errors in ASR and ST outputs and annotation can lead to serious implications, potentially affecting patient diagnoses and treatment decisions (Adane et al., 2019). Therefore, **we earnestly urge readers to independently verify our hypotheses and experimental results using their own medical data.** We also strongly recommend conducting pilot tests in a simulated doctor-patient environment before full-scale deploying them in real-world applications.

 Further limitations are extensively discussed in each Appendix Section.

Acknowledgement

This work was initiated as part of a bachelor thesis by Khai Le-Duc at RWTH Aachen University under the supervision of Prof. Hermann Ney and PD. Ralf Schlüter.

Most of the theoretical formulations in this work were borrowed from lectures by Hermann Ney, Ralf Schlüter, and Albert Zeyer, as well as from PhD dissertations at the Machine Learning and Human Language Technology Group at RWTH Aachen University.

We would like to thank other contributors, Long Vo-Dang, Nhut Huy Pham, and Viet Thanh Duy Nguyen for their precious initial efforts in this work.

References

- Kasaw Adane, Mucheye Gizachew, and Semalegne Kendie. 2019. The role of medical data in efficient patient care delivery: a review. *Risk management and healthcare policy*, pages 67–73.
- Hilal Al Shamsi, Abdullah G Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for healthcare: a systematic review. *Oman medical journal*, 35(2):e122.
- Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91.
- Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, et al. 2020. Findings of the iwslt 2020 evaluation campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34.
- Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roei Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. [The missing ingredient in zero-shot neural machine translation.](#)
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations.](#)
- Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On using specaugment for end-to-end speech translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. [Neural machine translation by jointly learning to align and translate.](#)
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.](#) In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. [Disc-medllm: Bridging general large language models and real-world medical consultation.](#)
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. LLMs instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.
- Rachel Bawden, Giorgio Maria Di Nunzio, Christian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névél, Mariana Neves, Maite Oronoz, et al. 2020. Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *Fifth Conference on Machine Translation*, pages 658–685. Association for Computational Linguistics (ACL).
- Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. [Bias mitigation in machine translation quality estimation.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

- L. Bentivogli, M Cettolo, M Gaido, A Karakanta, A Martinelli, M Negri, M Turchi, et al. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *ACL-IJCNLP 2021-59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2873–2887. Association for Computational Linguistics (ACL).
- Alexandre Bérard, Zae Myung Kim, Vassilina Nikoulina, Eunjeong Lucy Park, and Matthias Gallé. 2020. A multilingual neural machine translation model for biomedical data. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.
- Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.
- Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122.
- Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6486–6493.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). Association for Computational Linguistics.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics.
- Pierrette Bouillon, Glenn Flores, Maria Georgescu, Sonia Halimi, Beth Ann Hockey, Hitoshi Isahara, Kyoko Kanzaki, Yukie Nakao, Manny Rayner, Marianne Santaholma, et al. 2008. Many-to-many multilingual medical speech translation on a pda. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*.
- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.
- 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*.
- Ozgür Cetin and Elizabeth Shriberg. 2006. Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap. In *2006 IEEE international conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.
- Xuankai Chang, Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. 2021. Hypothesis stitcher for end-to-end speaker-attributed asr on long-form multi-talker recordings. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6763–6767. IEEE.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.
- Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur. 2015. Pronunciation and silence probability modeling for asr. In *Interspeech*, pages 533–537.
- Xi Chen, Songyang Zhang, Qibing Bai, Kai Chen, and Satoshi Nakamura. 2024b. Llast: Improved end-to-end speech translation system leveraged by large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6976–6987.
- Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023. M 3 st: Mix at three levels for speech translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

- Shun-Po Chuang, Yung-Sung Chuang, Chih-Chiang Chang, and Hung-Yi Lee. 2021. Investigating the reordering capability in ctc-based non-autoregressive end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1068–1077.
- Kenneth W Church and Eduard H Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8:239–258.
- Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.
- Adam L Cohen, Frederick Rivara, Edgar K Marcuse, Heather McPhillips, and Robert Davis. 2005. Are language barriers associated with serious medical events in hospitalized pediatric patients? *Pediatrics*, 116(3):575–579.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora 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 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, Changan Wang, Jeff Wang, and Skyler Wang. 2023a. [SeamlessM4: Massively multilingual & multimodal machine translation](#).
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady ElSahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta Ruiz Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine T. Kao, Ann Lee, Xutai Ma, Alexandre Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Y. Sun, Paden Tomasello, Changan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023b. [Seamless: Multilingual expressive and streaming speech translation](#). *ArXiv*, abs/2312.05187.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.
- Suresh Budha Dahal and Muhammad Aoun. 2023. Exploring the role of machine translation in improving health information access for linguistically diverse populations. *Advances in Intelligent Information Systems*, 8(2):1–13.
- Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.
- Juan Dempere. 2023. Deploying ai-driven natural language processing systems for public health crisis management. In *The HCT International Conference on Advancements in Health Sciences*, pages 131–153. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Long Doan, Nguyen Luong Tran, Thai Hoang, Dat Quoc Nguyen, et al. 2021. Phomt: A high-quality and large-scale benchmark dataset for vietnamese-english

machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Raden-ovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gre-goire Mialon, Guan Pang, Guillem Cucurell, Hai-ley Nguyen, Hannah Korevaar, Hu Xu, Hugo Tou-vron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jong-soo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshi-tiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pa-supuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoy-chev, Niladri Chatterji, Olivier Duchenne, Onur Ågelebi, Patrick Alrassy, Pengchuan Zhang, Peng-wei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srimi-vasan, Raj Ganapathy, Ramon Calderer, Ricardo Sil-veira Cabral, Robert Stojnic, Roberta Raileanu, Ro-hit Girdhar, Rohit Patel, Romain Sauvestre, Ron-nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa-hana Chennabasappa, Sanjay Singh, Sean Bell, Seo-hyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sha-ran Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Van-denhen-de, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Syd-ney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias

Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Vir-ginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenxin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xin-feng Xie, Xuchao Jia, Xuwei Wang, Yaelle Gold-schlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yi-wen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengx-ing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Sha-jnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boe-senberg, Alex Vaughan, Alexei Baevski, Allie Fein-stein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, An-drew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yaz-dan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Han-cock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Da-mon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Tes-tuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Hol-land, Edward Dowling, Eissa Jamil, Elaine Mont-gomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Flo-rez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han-wen Zha, Haroun Habeeb, Harrison Rudolph, He-len Suk, Henry Aspegren, Hunter Goldman, Ibrahim Dam-laj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khan-delwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsim-poukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L.

- Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Her-moso, Mo Metanat, Mohammad Rastegari, Mun-ish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pa-van Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratan-chandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Mah-eswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lind-say, Shaam Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agar-wal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poe-naru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojuan Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. [The llama 3 herd of models](#).
- Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. [Improving statistical machine translation in the medi-cal domain using the unified medical language sys-tem](#). In *Proceedings of the 20th International Confer-ence on Computational Linguistics*, COLING ’04, page 792–800, USA. Association for Computa-tional Linguistics.
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. [Data augmentation for low-resource neural machine translation](#). In *Proceedings of the 55th An-nual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vi-taliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. [Beyond english-centric multilingual machine translation](#).
- Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.
- Patrick Fernandes, Carolina Scarton, and Lucia Specia. 2023. [Assessing the Role of Context in Chat Transla-tion Evaluation: Is Context Helpful and Under What Conditions?](#) *Transactions of the Association for Computational Linguistics*.
- Glenn Flores, Margaret B. Laws, Susan J. Mayo, Barry Zuckerman, Marcia Abreu, Lillian Medina, and Eliz-abeth J. Hardt. 2003. [Errors in medical interpretation and their potential clinical consequences in pediatric encounters](#). *Pediatrics*, 111(1):6–14.
- Robert Flynn and Anton Ragni. 2023. How much con-text does my attention-based asr system need? *arXiv preprint arXiv:2310.15672*.
- Markus Freitag and Yaser Al-Onaizan. 2016. [Fast do-main adaptation for neural machine translation](#).
- Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein. 2020. Rnn-transducer with stateless prediction network. In *ICASSP 2020-2020 IEEE International Confer-ence on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7049–7053. IEEE.
- Pavel Golik. 2020. *Data-driven deep modeling and training for automatic speech recognition*. Ph.D. the-sis, Dissertation, RWTH Aachen University, 2020.
- Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. [Universal neural machine translation for extremely low resource languages](#).
- M. Harishbhai Tilala, P. Kumar Chenchala, A. Chop-padandi, J. Kaur, S. Naguri, R. Saoji, and B. Devagup-tapu. 2024. [Ethical considerations in the use of arti-ficial intelligence and machine learning in health care: A comprehensive review](#). *Cureus*, 16(6):e62443.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shum-ing Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.
- Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. [Sockeye: A toolkit for neural machine translation](#).
- Hay Man Htun, Ye Kyaw Thu, Hutchatai Chanlekha, Kotaro Funakoshi, and Thepchai Supnithi. 2024. mymedicon: End-to-end burmese automatic speech

- recognition for medical conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12032–12039.
- Hui Huang, Shuangzhi Wu, Kehai Chen, Hui Di, Muyun Yang, and Tiejun Zhao. 2023. [Improving translation quality estimation with bias mitigation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2175–2190. Association for Computational Linguistics.
- Jaeyoung Huh, Sangjoon Park, Jeong Eun Lee, and Jong Chul Ye. 2023. [Improving medical speech-to-text accuracy with vision-language pre-training model](#).
- Hirofumi Inaguma, Yosuke Higuchi, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2021. Orthros: Non-autoregressive end-to-end speech translation with dual-decoder. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7503–7507. IEEE.
- Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.
- Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 13 indian languages. *arXiv preprint arXiv:2411.04699*.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google's multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in C++](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451.
- Wioleta Karwacka. 2015. Medical translation. *Ways to translation*, pages 271–298.
- Elaine C Khoong and Jorge A Rodriguez. 2022. A research agenda for using machine translation in clinical medicine. *Journal of General Internal Medicine*, 37(5):1275–1277.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. [OpenNMT: Open-source toolkit for neural machine translation](#). In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.
- Philipp Koehn and Rebecca Knowles. 2017. [Six challenges for neural machine translation](#).
- Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Primock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598.
- Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, and Paulius Micikevicius. 2018. [OpenSeq2Seq: Extensible toolkit for distributed and mixed precision training of sequence-to-sequence models](#). In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 41–46, Melbourne, Australia. Association for Computational Linguistics.
- Khai Le-Duc. 2024. Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370.
- Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, Bach Phan Tat, Minh-Huong Ngo, and Truong-Son Hy. 2024. Multimed: Multilingual medical speech recognition via attention encoder decoder. *arXiv preprint arXiv:2409.14074*.
- Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. 2001. [Julius - an open source real-time large vocabulary recognition engine](#). In *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001*, pages 1691–1694. ISCA.

- Chia-ying Lee, Yu Zhang, and James Glass. 2013. Joint learning of phonetic units and word pronunciations for asr. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 182–192.
- Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. 2022. xformers: A modular and hackable transformer modelling library. <https://github.com/facebookresearch/xformers>.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. *Llava-med: Training a large language-and-vision assistant for biomedicine in one day*.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Boxiang Liu and Liang Huang. 2021. Paramed: a parallel corpus for english–chinese translation in the biomedical domain. *BMC Medical Informatics and Decision Making*, 21(1):258.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. *Multilingual denoising pre-training for neural machine translation*.
- Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020b. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8417–8424.
- Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020c. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*.
- Minh-Thang Luong and Christopher Manning. 2015. *Stanford neural machine translation systems for spoken language domains*. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.
- Xutai Ma, Anna Sun, Siqi Ouyang, Hirofumi Inaguma, and Paden Tomasello. 2023. Efficient monotonic multihead attention. *arXiv preprint arXiv:2312.04515*.
- Andrew L Maas, Quoc V Le, Tyler M O’Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng. 2012. Recurrent neural networks for noise reduction in robust asr. In *Interspeech*, volume 2012, pages 22–25.
- Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756.
- Kavya Manohar and Leena Pillai. 2024. What is lost in normalization? exploring pitfalls in multilingual asr model evaluations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10864–10869.
- Laurette Marais, Johannes A Louw, Jaco Badenhorst, Karen Calteaux, Ilana Wilken, Nina Van Niekerk, and Glenn Stein. 2020. Awezamed: A multilingual, multimodal speech-to-speech translation application for maternal health care. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.
- Vikram C Mathad, Tristan J Mahr, Nancy Scherer, Kathy Chapman, Katherine C Hustad, Julie Liss, and Visar Berisha. 2021. The impact of forced-alignment errors on automatic pronunciation evaluation. In *Interspeech*, pages 1922–1926.
- Nikki Mirghafori, Eric Fosler, and Nelson Morgan. 1996. Towards robustness to fast speech in asr. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 335–338. IEEE.
- Lauren Neergard. 2003. Hospitals struggle with growing language barrier. *Associated Press, The Charlotte Observer Sept*, 2:2003.
- Mariana Neves. 2017. A parallel collection of clinical trials in portuguese and english. In *Proceedings of the 10th workshop on building and using comparable corpora*, pages 36–40.
- Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. 2022. *Mtct: Multi-domain translation for english and vietnamese*.
- Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. 2022. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*.
- Toan Q. Nguyen and David Chiang. 2017. *Transfer learning across low-resource, related languages for neural machine translation*.

- Mary Nurminen and Maarit Koponen. 2020. Machine translation and fair access to information. *Translation Spaces*, 9(1):150–169.
- Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#).
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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.
- Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. *Interspeech*.
- Pavel Pecina, Ondřej Dušek, Jan Hajič, Jindřich Libovický, and Zdeňka Uřešová. 2017. [Khresmoi query translation test data 2.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*.
- Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Luká Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. [The kaldi speech recognition toolkit](#).
- Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. [Wav2letter++: A fast open-source speech recognition system](#). In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6460–6464. IEEE.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#).
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Alec Radford and Karthik Narasimhan. 2018. [Improving language understanding by generative pre-training](#).
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- David Rybach, Stefan Hahn, Patrick Lehnert, David Nolden, Martin Sundermeyer, Zoltán Tóth, Simon Wiesler, Ralf Schlüter, and Hermann Ney. 2011. Rasr - the rwth aachen university open source speech recognition toolkit.
- Sreelekha S. 2017. [Statistical vs rule based machine translation; a case study on indian language perspective](#).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: Lms for intralingual cultural adaptation. *arXiv preprint arXiv:2406.14504*.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Benjamin LW Sobel. 2017. Artificial intelligence’s fair use crisis. *Colum. JL & Arts*, 41:45.
- Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325.
- 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.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.
- Zewei Sun, Qingnan Jiang, Shujian Huang, Jun Cao, Shanbo Cheng, and Mingxuan Wang. 2022. [Zero-shot domain adaptation for neural machine translation with retrieved phrase-level prompts](#).
- Akshay Swaminathan, Iván López, Rafael Antonio García Mar, Tyler Heist, Tom McClintock, Kaitlin Caoili, Madeline Grace, Matthew Rubashkin, Michael N Boggs, Jonathan H Chen, et al. 2023. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. *NPJ Digital Medicine*, 6(1):213.
- Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213. IEEE.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. [Multilingual translation with extensible multilingual pretraining and finetuning](#).

- NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.
- Jörg Tiedemann and Santhosh Thottingal. 2020. Opusmt—building open translation services for the world. In *Proceedings of the 22nd annual conference of the European Association for Machine Translation*, pages 479–480.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Vivek Tyagi, Hervé Bourlard, and Christian Wellekens. 2006. On variable-scale piecewise stationary spectral analysis of speech signals for asr. *Speech communication*, 48(9):1182–1191.
- Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. [Tensor2tensor for neural machine translation](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Neha Verma, Kenton Murray, and Kevin Duh. 2022. [Strategies for adapting multilingual pre-training for domain-specific machine translation](#). In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 31–44, Orlando, USA. Association for Machine Translation in the Americas.
- Nhu Vo, Dat Quoc Nguyen, Dung D Le, Massimo Piccardi, and Wray Buntine. 2024. Improving vietnamese-english medical machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8955–8962.
- Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. 2004. Sphinx-4: a flexible open source framework for speech recognition. Technical report, USA.
- Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2022. [fairseq s2t: Fast speech-to-text modeling with fairseq](#).
- Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *Inter-speech*.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012.
- John S White, Theresa A O’Connell, and Francis E O’Mara. 1994. The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.
- Steven Woloshin, Nina A Bickell, Lisa M Schwartz, Francesca Gany, and H Gilbert Welch. 1995. Language barriers in medicine in the united states. *Jama*, 273(9):724–728.
- Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. [How far can 100 samples go? unlocking overall zero-shot multilingual translation via tiny multi-parallel data](#).
- Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. [Generalized data augmentation for low-resource translation](#).
- Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023a. Recent advances in direct speech-to-text translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6796–6804.
- Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023b. Efficient sequence transduction by jointly predicting tokens and durations. In *International Conference on Machine Learning*, pages 38462–38484. PMLR.
- Zhe-Jia Xu, Yeou-Jiunn Chen, and Qian-Bei Hong. 2024. Multilingual speech translator for medical consultation. In *2024 27th Conference of the Oriental COCOSA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSA)*, pages 1–5. IEEE.

- Jian Xue, Peidong Wang, Jinyu Li, Matt Post, and Yashesh Gaur. 2022. Large-scale streaming end-to-end speech translation with neural transducers. *Inter-speech*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mt5: A massively multilingual pre-trained text-to-text transformer](#).
- Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xi-anchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Qimin Yang, Rongsheng Wang, Jiexin Chen, Runqi Su, and Tao Tan. 2024b. [Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise](#).
- Xinyi Yang, Runzhe Zhan, Derek F Wong, Junchao Wu, and Lidia S Chao. 2023. Human-in-the-loop machine translation with large language model. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 88–98.
- Leon R Yankwich. 1954. What is fair use? *The University of Chicago Law Review*, 22(1):203–215.
- Steve Young, D. Kershaw, James Odell, D. Ollason, V. Valtchev, and Philip Woodland. 2000. [The htk book version](#).
- Qing Yu, Zhe Li, Jiabao Sheng, Jing Sun, and Wushour Slamu. 2020. Yuq: A chinese-uyghur medical-domain neural machine translation dataset towards knowledge-driven. In *Machine Translation: 16th China Conference, CCMT 2020, Hohhot, China, October 10-12, 2020, Revised Selected Papers 16*, pages 37–54. Springer.
- Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021a. Bstc: A large-scale chinese-english speech translation dataset. In *Proceedings of the Second Workshop on Automatic Simultaneous Translation*, pages 28–35.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Yongle Zhang, Dennis Asamoah Owusu, Emily Gong, Shaan Chopra, Marine Carpuat, and Ge Gao. 2021b. Leveraging machine translation to support distributed teamwork between language-based subgroups: The effects of automated keyword tagging. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6.
- Yongle Zhang and Ge Gao. 2024. Assisting international migrants with everyday information seeking: From the providers’s lens. In *International Conference on Information*, pages 338–355. Springer.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. [Transfer learning for low-resource neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Contents

1	Introduction	1
2	Data	2
2.1	Data Collection	2
2.2	Annotation Process and Data Quality Control	3
2.3	Data Statistics	3
3	Problem Formulation	3
4	Experimental Setup	4
4.1	Training Setup	4
4.2	Evaluation Metrics	4
5	Experimental Results	5
5.1	Automatic Speech Recognition Baselines	5
5.2	Ground-truth Translation Baselines	5
5.3	Cascaded Speech Translation Baselines	6
5.4	End-to-end and Cascaded Comparison	6
5.5	Bilingual-Multilingual Fine-tuning Comparison	7
5.6	Bilingual-Multilingual Pre-training Comparison	8
5.7	Code-Switch Analysis	8
6	Error Analysis	8
6.1	Quantitative Error Analysis	8
6.2	Qualitative Error Analysis	9
7	Conclusion	9
8	Limitations	9
A	Related Works	23
A.1	Neural Machine Translation	23
A.2	Cascaded Speech Translation	23
A.3	End-to-end Speech Translation	23
A.4	Medical Machine Translation	24
A.5	Domain Adaptation for Machine Translation	24
A.6	Multilingual Machine Translation	25
B	Theoretical Formulation	26
B.1	Mel-Frequency Cepstral Coefficients (MFCCs)	26
B.2	Attention Encoder Decoder (AED)	27
B.3	SpecAugment	29
C	Dataset Comparison with Literature	30
D	Details of Experimental Setup	33
D.1	Training Setup: Whisper	33
D.2	Training Setup: Deepgram	34
D.3	Training Setup: AssemblyAI	34
D.4	Training Setup: mBART	35
D.5	Training Setup: M2M100	37
D.6	Training Setup: Marian	37

D.7	Training Setup: Llama	37
D.8	Training Setup: Qwen	38
D.9	Training Setup: Mistral	38
D.10	Training Setup: Google Translate	40
D.11	Training Setup: VinAI Translate	41
D.12	Training Setup: EnViT5	42
D.13	Training Setup: SeamlessM4T	42
D.14	Training Setup: Qwen-Audio	43
D.15	In-context Learning Prompt	45
E	Details of Evaluation Metrics	54
E.1	Discussion about Automatic Evaluation Metrics	54
E.2	Details of Human Evaluation	56
E.3	Details of LLM-as-a-judge	56
F	Extra Experimental Results	58
F.1	In-context Learning Results	58
F.2	Full Results: Ground-truth Translation Baselines	78
F.3	Extra Results: Cascaded Speech Translation Baselines	89
F.4	Qualitative Results	100
F.4.1	Vietnamese to German Speech Translation	100
F.4.2	Vietnamese to English Speech Translation	101
F.4.3	Vietnamese to French Speech Translation	102
F.4.4	Vietnamese to Chinese Speech Translation	103
F.4.5	English to Vietnamese Speech Translation	104
F.4.6	English to German Speech Translation	105
F.4.7	English to French Speech Translation	106
F.4.8	English to Chinese Speech Translation	107
F.4.9	German to Vietnamese Speech Translation	108
F.4.10	German to English Speech Translation	109
F.4.11	German to French Speech Translation	110
F.4.12	German to Chinese Speech Translation	111
F.4.13	French to Vietnamese Speech Translation	112
F.4.14	French to German Speech Translation	113
F.4.15	French to English Speech Translation	114
F.4.16	French to Chinese Speech Translation	115
F.4.17	Chinese to Vietnamese Speech Translation	116
F.4.18	Chinese to English Speech Translation	117
F.4.19	Chinese to French Speech Translation	118
F.4.20	Chinese to German Speech Translation	119
G	Ethical Statements	120
G.1	Fair Use	120
G.1.1	Fair Use Considerations	120
G.1.2	Ensuring Fair Use Compliance	121
G.2	Data Consent	121
G.3	Annotation Problem for Long-form Speech	122
H	List of Abbreviations	124
I	Contribution Statements	126

A Related Works

A.1 Neural Machine Translation

Neural Machine Translation (NMT) has experienced substantial advancements with the development of Transformer-based models, such as the Transformer architecture (Vaswani et al., 2017). The Transformer represents the first seq2seq model solely reliant on the attention mechanism, wherein the recurrent layers of traditional models are replaced by multi-headed self-attention within the encoder-decoder framework. This architectural innovation has significantly accelerated training speeds in comparison to Recurrent Neural Network (RNN) and Convolution Neural Network (CNN), resulting in superior performance. BERT (Devlin et al., 2019), a pre-trained model designed to address the unidirectional constraints of earlier language models (such as the left-to-right processing in Transformers), incorporates a masked language model (MLM) to enable bidirectional representation, thereby enhancing machine translation tasks. Building on BERT and other pre-training paradigms, BART (Lewis et al., 2019) generalizes these techniques, achieving competitive results in various NMT applications.

The GPT series, which demonstrates the efficacy of generative pre-training followed by fine-tuning for the MT task in GPT-1 (Radford and Narasimhan, 2018), exhibits remarkable performance in text generation and zero-shot tasks. GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) further scale the model's size and training data, facilitating state-of-the-art performance in few-shot and zero-shot tasks, including translation. GPT-4 (OpenAI et al., 2024) further improves capabilities in multilingual and domain-specific MT tasks.

Several NMT frameworks, such as OpenNMT (Klein et al., 2017), have been developed to facilitate the integration of custom deep learning models for translation tasks. These frameworks provide tools that optimize the efficiency of training, inference, and deployment in NMT systems. Marian-NMT (Junczys-Dowmunt et al., 2018) emphasizes speed and scalability, enabling the implementation of state-of-the-art NMT models with minimal computational overhead. OpenSeq2Seq (Kuchaiev et al., 2018) offers reference implementations designed for efficient distributed and mixed-precision training. Tensor2Tensor (Vaswani et al., 2018) and Sockeye (Hieber et al., 2018) prioritize the secu-

rity, reliability, and production-level performance of their software components. Fairseq (Ott et al., 2019) is a fast, extensible toolkit for sequence modeling that offers scalability and is versatile across numerous applications.

A.2 Cascaded Speech Translation

ST traditionally contains two components: ASR (to convert audio into text) and NMT (to translate text-to-text). The success in the ASR technology starts with HTK (Young et al., 2000) - a toolkit for manipulating Hidden Markov Models (HMM) provides comprehensive facilities for speech analysis, training, and recognition. Later success includes Julius (Lee et al., 2001) - an open-source, high-performance, two-pass large vocabulary continuous speech recognition (LVCSR) decoder; Sphinx-4 (Walker et al., 2004) - a flexible, modular, and plug-gable framework for ASR written entirely in Java; RWTH ASR (Rybach et al., 2011) - an open-source ASR decoding system which includes state-of-the-art ASR capabilities. Furthermore, Kaldi model (Povey et al., 2011) provides a hybrid ASR system based on finite-state transducers. Recent state-of-the-art framework was wav2vec 2.0 (Baevski et al., 2020) - a framework for self-supervised learning of speech representations which masks latent representations of the raw waveform and solves a contrastive task over quantized speech representations; and Whisper model (Radford et al., 2022) - which suggests that scaling weakly supervised pre-training has been underestimated in ASR research. Other novel frameworks are from Facebook AI's end-to-end ASR research, including wave2letter++ (Pratap et al., 2019) - the fastest open-source deep learning ASR framework, and Fairseq S2T (Wang et al., 2022) - which bypassed traditional transcription steps, improving both latency and accuracy.

A.3 End-to-end Speech Translation

The development of end-to-end ST models, which eliminate intermediary stages like ASR outputs and lattices, has significantly reduced error propagation (Chen et al., 2024b). Research shows end-to-end ST models achieve performance comparable to cascaded models (Sperber et al., 2019; Ansari et al., 2020; Bentivogli et al., 2021). Moreover, these models offer benefits like reduced latency and applicability to unwritten languages. (Bérard et al., 2016).

Some researchers have modified the multi-task encoder-decoder architecture (Weiss et al., 2017)

by splitting the decoder into two components (Liu et al., 2020b; Anastasopoulos and Chiang, 2018): one used to transcribe and the other one used to translate. Parallel research initiatives have likewise separated the encoder (Liu et al., 2020c; Cheng et al., 2023), with subsequent studies demonstrating that a shared encoder can be independently segmented to optimize the utilization of ASR data (Tang et al., 2021; Xu et al., 2023a). Furthermore, non-autoregressive (NAR) modeling has been investigated as a method to reduce latency (Inaguma et al., 2021; Chuang et al., 2021).

Recent advancements have notably explored multitasking within the framework of large-scale training, yielding remarkable performance on ST benchmarks, like Whisper (Radford et al., 2022), SeamlessM4T (Communication et al., 2023a). Another predominant approach involves the integration of an LLM at the backend with a speech encoder at the frontend, like LauraGPT (Chen et al., 2024b), Qwen-Audio (Chu et al., 2023).

A.4 Medical Machine Translation

The translation of medical texts poses distinct challenges owing to the use of specialized terminology, frequent abbreviations, and the imperative requirement for precision (Neergard, 2003; Flores et al., 2003). Early methodologies predominantly utilized Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT), both of which were tailored to medical language corpora (Eck et al., 2004). RBMT utilizes predefined rules and lexical databases to translate texts by analyzing their grammatical and lexical structures. It is particularly adept at managing medical terminology, provided that the dictionaries are up-to-date and comprehensive. However, RBMT has limitations, including an inability to resolve ambiguity, interpret idiomatic expressions, and account for variations in language use. Additionally, RBMT requires substantial human effort for the creation and ongoing maintenance of the rules and dictionaries specific to each language pair (S, 2017). SMT, in contrast, depends on large parallel corpora-collections of aligned texts in two languages - to estimate the probability of translation equivalents (Brown et al., 1993). In contrast to rule-based or dictionary-based systems, SMT relies on data-driven algorithms to produce translations. This characteristic enables SMT to be highly adaptable across different domains and genres, including specialized fields such as medical texts, by utilizing domain-specific cor-

pora customized for both the source and target languages, as well as their respective contexts. However, SMT is not without limitations. It often encounters challenges in generating fluent or grammatically accurate translations, particularly when dealing with low-resource languages or rare terminology, resulting in outputs that may be unnatural or imprecise (Koehn and Knowles, 2017). The occurrence of NMT allowed for vast improvements, particularly with encoder-decoder architectures enhanced by attention mechanisms (Bahdanau et al., 2016). Recent studies have demonstrated that domain adaptation techniques, such as fine-tuning LLMs on domain-specific datasets, can enhance the performance of medical tasks, including translation. (Bao et al., 2023; Yang et al., 2024b).

A.5 Domain Adaptation for Machine Translation

Medical MT for low-resource languages continues to present a significant challenge, primarily due to the absence of multilingual medical databases. Strategies such as data augmentation, which involves generating synthetic data to expand existing datasets (Fadaee et al., 2017; Xia et al., 2019), back-translation, where target-to-source translations are utilized to create additional source-to-target pairs (Sennrich et al., 2016), and transfer learning (Zoph et al., 2016; Nguyen and Chiang, 2017; Gu et al., 2018), which capitalizes on knowledge from high-resource languages to enhance performance in low-resource languages, have been proposed to address this issue.

Multilingual NMT models such as mBART (Liu et al., 2020a), XLM-R (Conneau et al., 2020), M2M-100 (Fan et al., 2020), and mT5 (Xue et al., 2021) have demonstrated significant potential in overcoming the challenges associated with low-resource or domain-specific settings. This is achieved through the use of cross-lingual transfer learning, which allows the model to leverage shared linguistic representations across multiple languages. Consequently, this approach markedly improves the model's ability to generalize, even in the presence of limited training data in the target language.

Ethical considerations are also an essential problem in the context of medical MT, given its potential implications for patient care (Harishbhai Tilala et al., 2024).

Future research is further centered on the integration of multimodal data, such as the combina-

tion of textual and audio-visual inputs, to improve translation accuracy within medical contexts (Huh et al., 2023; Li et al., 2023). Furthermore, fine-tuning pre-trained models on multilingual medical datasets, such as the Unified Medical Language System (UMLS), has shown promise in enhancing model performance while addressing the unique challenges associated with medical domains. However, these research directions lie beyond the scope of the our present study.

A.6 Multilingual Machine Translation

Recent research has increasingly focused on multilingual translation. For instance, studies by Luong and Manning (2015) and Freitag and Al-Onaizan (2016) have demonstrated that pre-training models on a diverse dataset, followed by fine-tuning on a smaller target dataset, yields effective results. Liu et al. (2020a) extended the BART model with mBART and showed that multilingual denoising pre-training leads to significant performance improvements across a variety of MT benchmarks. Additionally, Verma et al. (2022) highlighted the effectiveness of multilingual pre-training in domain adaptation scenarios. Research by Johnson et al. (2017) further indicated that a trained multilingual NMT system could perform zero-shot translation between previously unseen language pairs without direct supervision, provided that both source and target languages were included in the training process. (Arivazhagan et al., 2019) observed that the cosine similarity between the pooled encoder outputs of sentence pairs decreased during multilingual training. Meanwhile, Sun et al. (2022) addressed domain adaptation by constructing bilingual phrase-level databases and retrieving contextually relevant prompts, which improved translation quality in unseen domains. On a different note, (Wu et al., 2024) proposed an approach that fine-tuned models with a minimal amount of multi-parallel data, finding that a small, randomly sampled set of fine-tuning directions was sufficient for achieving comparable improvements.

B Theoretical Formulation

B.1 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCC serves as a compact representation of the audio signal’s spectral properties. The computation of MFCCs begins by dividing the input signal $x_1^T := x_1, x_2, \dots, x_T$ into overlapping frames, as visualized in Figure 2¹⁸.

Pre-emphasis: The audio signal, sampled at 16 kHz with a step size of 10 ms, is processed by extracting 160 consecutive samples from the Pulse Code Modulation (PCM) waveform for each frame. These 10 ms frames are non-overlapping, ensuring that stacking adjacent vectors avoids discontinuities. The 16-bit quantized samples, which span the integer range from -2^{15} to $+2^{15}$, must be normalized to a numerically stable range. This normalization is achieved by applying mean and variance normalization, either globally across the entire training dataset or on a per-utterance basis. A commonly employed processing technique, known as high-frequency pre-emphasis, can be implemented by computing the differences between adjacent samples, as illustrated below:

$$x'_t = x_t - x_{t-1} \in \mathbb{R} \quad (4)$$

A sequence of $16 \text{ kHz} \times 10 \text{ ms} = 160$ pre-emphasized waveform samples can then be considered a feature vector:

$$\hat{x}_t = x'^t_{t-160+1} \in \mathbb{R}^{160} \quad (5)$$

Amplitude spectrum - Fast Fourier Transform (FFT): The Short-Time Fourier Transform (STFT) is applied to overlapping windows with a duration of 25 ms. Given a sampling rate of 16 kHz, this window length corresponds to $25 \text{ ms} \times 16 \text{ kHz} = 400$ samples. To facilitate computation using the FFT, the sample count is zero-padded to the next power of two, resulting in $2^9 = 512$.

$$\begin{aligned} \mathbf{z}_t &\in \mathbb{R}^{512} \\ &= \left[\begin{array}{cccc} x'^{t'}_{t-400+1} & x'^{t'}_{t-400+2} & \dots & x'^{t'}_{t'} \end{array} \quad \underbrace{0 \dots 0}_{\text{zero-padding}} \right] \end{aligned} \quad (6)$$

The extended sample vector is weighted using a Hann window, which exhibits smaller side lobes in

the amplitude spectrum compared to a rectangular window:

$$\begin{aligned} w^{(n)} &= 0.5 - 0.5 \cos \left(\frac{2\pi(n-1)}{512-1} \right), \\ 1 &\leq n \leq 512 \end{aligned} \quad (7)$$

$$d_t^{(n)} = \mathbf{z}_t^{(n)} \cdot w^{(n)} \quad (8)$$

While the discrete STFT could be done directly by evaluating the sum

$$\begin{aligned} \mathcal{S}_t^{(\mathbb{F})} &= \sum_{n=0}^{512-1} d_t^{(n)} \cdot \exp \left(-j \frac{2\pi}{512} \mathbb{F} n \right), \\ 1 &\leq \mathbb{F} \leq 512 \end{aligned} \quad (9)$$

the complexity can be reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$ by applying the fast Fourier transform.

The 512-FFT results in a 257-dimensional vector because of the symmetry of the amplitude spectrum of a real-valued signal. The phase spectrum is removed.

$$\begin{aligned} \hat{x}_t &= \left[|\mathcal{S}_t^{(0)}| \quad |\mathcal{S}_t^{(1)}| \quad \dots \quad |\mathcal{S}_t^{(512/2)}| \right] \\ &\in \mathbb{R}^{512/2+1} \end{aligned} \quad (10)$$

MFCC: The MFCC feature extraction is based on the STFT of the pre-emphasized speech signal (Davis and Mermelstein, 1980). It considers the nonlinear sensitivity of human auditory perception to variations in frequency. This is evidenced that the filter bank used to integrate the magnitude spectrum $|\mathcal{S}_t^{(\mathbb{F})}|$ consists of $\mathbb{I}$ filters equidistantly spaced on the mel scale. The mel scale is a logarithmically scaled frequency axis. The k -th frequency bin of the FFT centered around $\mathbb{F}_k$ Hz is then mapped to $\tilde{\mathbb{F}}_k$ on the mel scale:

$$\mathbb{F}_k = \frac{k}{512} \cdot \mathbb{F}_\delta \quad (11)$$

$$\tilde{\mathbb{F}}_k = 2595 \cdot \log_{10} \left(1 + \frac{\mathbb{F}_k}{700 \text{ Hz}} \right) \quad (12)$$

The filter center $\tilde{\mathbb{F}}_c^{(i)}$ of the i -th triangular filter is then placed at $i \cdot \tilde{\mathbb{F}}_b$, where the bandwidth $\tilde{\mathbb{F}}_b$ corresponds to $\tilde{\mathbb{F}}_{512}/\mathbb{I}$. With these parameters, the coefficients of the i -th triangular filter can be calculated explicitly as a piecewise linear function and stored in a weight vector $\mathbf{v}_i \in \mathbb{R}^{N/2+1}$.

¹⁸Golik (2020)’s Dissertation at RWTH Aachen University described MFCC more comprehensively. MFCC visualization image is retrieved from Pytorch library.

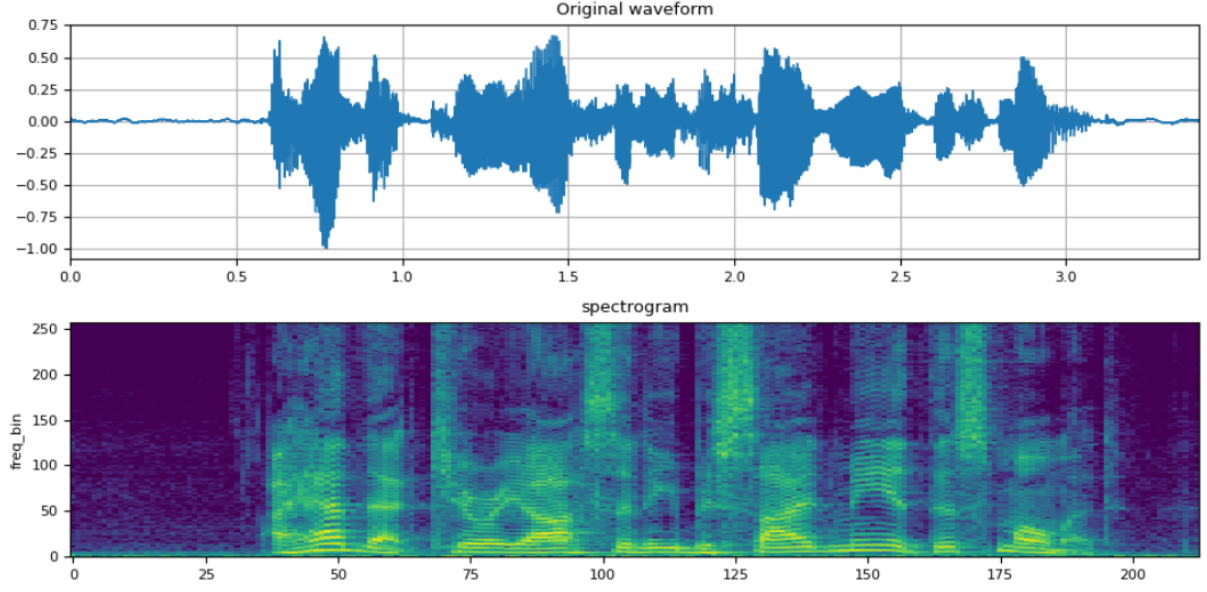


Figure 2: **Mel-Frequency Cepstral Coefficients (MFCC) visualization.** The computation of MFCCs begins by dividing the original waveform into overlapping 20ms frames.

By applying discrete cosine transform (DCT), the MFCC features are extracted from the logarithm filter outputs:

$$\mathcal{X}_t^{(i)} = \log_{10} \left(\sum_{\mathbb{F}=0}^{512} |\mathcal{S}_t^{(\mathbb{F})}| \mathcal{V}_i^{(\mathbb{F})} \right) \quad (13)$$

$$\mathcal{C}_{m,i} = \cos \left(\frac{\pi m(i + 0.5)}{\mathbb{I}} \right) \quad (14)$$

$$\mathcal{C}_t^{(m)} = \sum_{i=0}^{\mathbb{I}-1} \mathcal{C}_{m,i} \mathcal{X}_t^{(i)} \quad (15)$$

$$\hat{\mathbf{x}}_t = [\mathcal{C}_t^{(0)} \mathcal{C}_t^{(1)} \dots \mathcal{C}_t^{(\mathbb{I}-1)}] \in \mathbb{R}^{\mathbb{I}} \quad (16)$$

B.2 Attention Encoder Decoder (AED)

As for AED models, Whisper architecture is shown in Figure 3, and Deepgram architecture is shown in Figure 4.

An ASR model is used to transcribe speech into text by mapping an audio signal $x_1^T := x_1, x_2, \dots, x_T$ of length T to the most likely word sequence w_1^N of length N . The word sequence probability is described as:

$$p(w_1^N | x_1^T) = \prod_{n=1}^N p(w_n | w_1^{n-1}, x_1^T). \quad (17)$$

In the ASR encoder-decoder architecture, given D as the feature dimension size, the input audio signal matrix could be described as $x_1^T \in \mathbb{R}^{T \times D_{input}}$.

When simplified, downsampling before or inside the encoder - conducted by a fixed factor, such as striding in a CNN - is removed. Thus, the encoder output sequence is as follows:

$$h_1^T = \text{Encoder}(x_1^T) \in \mathbb{R}^{T \times D_{encoder}}. \quad (18)$$

Using a stack of Transformer ($\mathcal{T}$) blocks (Vaswani et al., 2017), the encoder output sequence is described as function composition:

$$h_1^T = \mathcal{T}_0 \circ \dots \circ \mathcal{T}_{N_{EncLayers}}(x_1^T). \quad (19)$$

In the decoder, the probability for each single word is defined as:

$$\begin{aligned} p(w_n | w_1^{n-1}, x_1^T) &= p(w_n | w_1^{n-1}, h_1^T(x_1^T)) \\ &= p(w_n | w_1^{n-1}, h_1^T). \end{aligned} \quad (20)$$

Based on Equation 17, the word sequence probability given the output of encoder is described as:

$$p(w_1^N | x_1^T) = \prod_{n=1}^N p(w_n | w_1^{n-1}, h_1^T). \quad (21)$$

Then, decoder hidden state is formulated as:

$$g_n = \mathcal{F}(g_{n-1}, w_{n-1}, c_n) \in \mathbb{R}^{D_g}, \quad (22)$$

where $\mathcal{F}$ is neural network; D_g is hidden state dimension; and c_n is context vector, e.g. weighted sum of encoder outputs via attention mechanism.

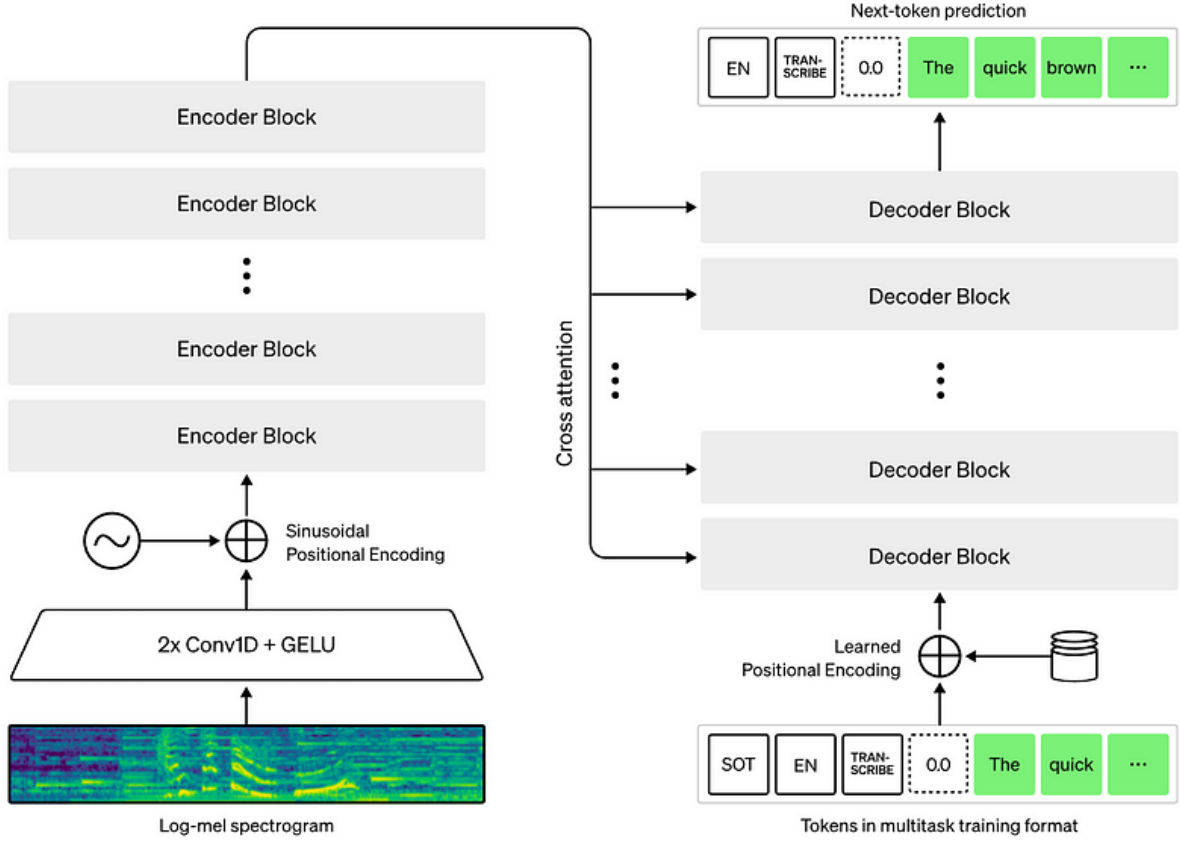


Figure 3: **OpenAI's Whisper architecture.** Whisper is a Transformer-based AED architecture, using MFCC features as input.

The attention mechanism in the decoder is described via 3 components: context vector c_n , attention weights $\alpha_{n,t}$, and attention energy $e_{n,t}$:

$$\begin{aligned}
 c_n &= \sum_{t=1}^T \alpha_{n,t} h_t \in \mathbb{R}^{D_{\text{encoder}}}, \\
 \alpha_{n,t} &= \frac{\exp(e_{n,t})}{\sum_{t'=1}^T \exp(e_{n,t'})} \\
 &= \text{Softmax}_T(\exp(e_{n,t})) \in \mathbb{R}, \\
 e_{n,t} &= \text{Align}(g_{n-1}, h_t) \in \mathbb{R} \\
 &= W_2 \cdot \tanh(W_1 \cdot [g_{n-1}, h_t]),
 \end{aligned} \tag{23}$$

where n is decoder step; t is encoder frame; $\alpha \in \mathbb{R}^{T \times N}$ is attention weight matrix; $\alpha_n \in \mathbb{R}^T$ is normalized probability distribution over t ; Softmax_T is Softmax function over spatial dimension T , not feature dimension; $W_1 \in \mathbb{R}^{(D_g + D_{\text{encoder}}) \times D_{\text{key}}}$; $W_2 \in \mathbb{R}^{D_{\text{key}}}$.

In the decoding, the output probability distribu-

tion over vocabulary is defined as:

$$\begin{aligned}
 p(w_n = * | w_1^{n-1}, h_1^T) \\
 = \text{Softmax}(\text{MLP}(w_{n-1}, g_n, c_n)) \in \mathbb{R}^N,
 \end{aligned} \tag{24}$$

where MLP is Multi-layer Perceptron.

To train an AED model, sequence-level frame-wise cross-entropy loss is employed:

$$\begin{aligned}
 \mathcal{L}_{\text{AED}} &= - \sum_{(x_1^T, w_1^N)} \log p(w_1^N | x_1^T) \\
 &= - \sum_{(x_1^T, w_1^N)} \sum_{n=1}^N \log p(w_n | w_1^{n-1}, x_1^T).
 \end{aligned} \tag{25}$$

During beam search, the auxiliary quantity for each unknown partial string (tree of partial hypotheses) w_1^n is defined as:

$$\begin{aligned}
 Q(n; w_1^n) &:= \prod_{n'=1}^n p(w_{n'} | w_0^{n'-1}, x_1^T) \\
 &= p(w_n | w_0^{n-1}, x_1^T) \cdot Q(n-1, w_1^{n-1}).
 \end{aligned} \tag{26}$$

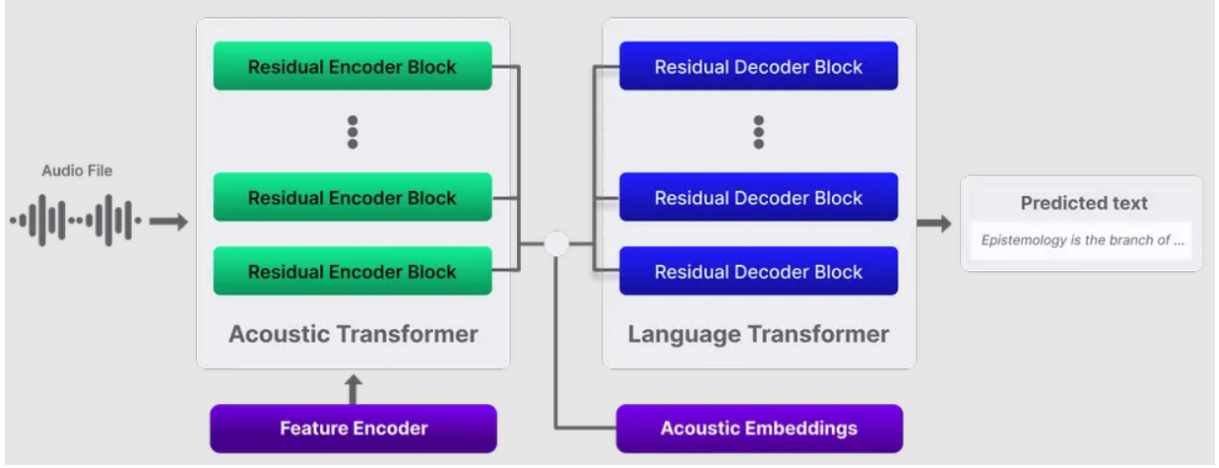


Figure 4: **Deepgram’s Nova-2 architecture.** To our best understanding of Deepgram’s documentation, Deepgram’s Nova-2 is a Transformer-based AED architecture, using raw waveform as input instead of MFCC like Whisper. Feature extraction from raw waveform is probably conducted by a learnable feature encoder, e.g. a block of CNNs like wav2vec 2.0. Between encoder-decoder space, (unknown) acoustic embeddings are probably added as cross-attention.

After discarding the less likely hypotheses in the beam search, the word sequence probability is calculated by the best hypothesis:

$$p(w_1^N | x_1^T) = Q(N; w_1^N). \quad (27)$$

B.3 SpecAugment

SpecAugment (Park et al., 2019) is a data augmentation technique for ASR that manipulates spectrograms to improve model robustness by randomly applying masking in consecutive frames in the time axis as well as consecutive dimensions in the feature axis. It performs three main transformations¹⁹: time warping, frequency masking, and time masking.

Figure 5 shows examples of the individual augmentations applied to a single input.

Time Masking: Given an audio signal $x_1^T := x_1, x_2, \dots, x_T$ of length T . Time masking is masking of τ successive time steps $[t, t + \tau)$, where we set:

$$(x_t, \dots, x_{t+\tau}) := 0 \quad (28)$$

where τ is the masking window selected from a uniform distribution from 0 to the maximum time mask parameter TMM. The time position t is picked from another uniform distribution over $[0, T)$ such that the maximum sequence length T is not exceeded (i.e. if $t + \tau > T$, we set it to T).

Frequency Masking: Frequency masking is applied such that ϕ consecutive frequency channels

$[f, f + \phi)$ are masked, where ϕ is selected from a uniform distribution from 0 to the frequency mask parameter FMM, and f is chosen from $[0, \nu)$, where ν is the input feature dimension, e.g. the number of MFCC channels. For raw waveform as input, $\nu = 1$. Similar to time masking, if $f + \phi > \nu$, we set it to $f = \nu$.

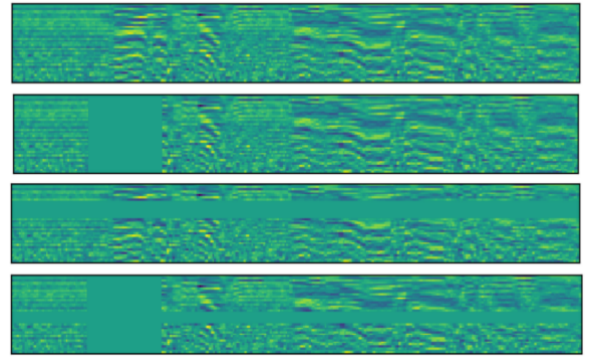


Figure 5: **SpecAugment visualization.** From top to bottom, the figures show the spectrogram of the input audio with no data augmentation, time masking, frequency masking and both masking applied.

¹⁹Bahar et al. (2019) analyzed deeply in end-to-end ST. Park et al. (2019) stated that time warping is the most expensive and the least influential, we do not include it here

C Dataset Comparison with Literature

Dataset	Dur.	Language	Nature	#Rec. Cond.	#Spk	#Acc	#Roles
VietMed (Le-Duc, 2024)	16h	Vietnamese	Real-world	8	61	6	6
PriMock57 ² (Korfiatis et al., 2022)	9h	English	Simulated	1	64	4	2
Fareez et al. (2022) ³	55h	English	Simulated	1	N/A	1	2
AfriSpeech-200 ⁴ (Olatunji et al., 2023)	≈123h	African English	Read speech	1	N/A	N/A	1
myMediCon ⁵ (Htun et al., 2024)	11h	Burmese	Read speech	1	12	5	2
 MultiMed-ST ¹ (ours)	150h	Multiling.	Real-world	10	198	16	6

Table 11: **Dataset comparison with literature: A list of all publicly available medical ASR datasets.**

Our  MultiMed-ST is the largest and most diverse medical ASR dataset.

From left to right: Total duration in hours (h), language, nature of speech, number of recording conditions, number of speakers, number of accents, speaking roles.

¹In our dataset, only the number of recording conditions, speakers, accents and speaking roles for Vietnamese and English are identified because of technical and privacy issues. Therefore, the exact number of speakers and accents must be much larger than the currently reported number. 10 recording conditions include: Documentary, Interview, Lecture, News, Podcast, Webinar, Speech, Talk, Vlog, Workshop. 10 English accents include: Main US, Southern US, UK, Australian, Indian, Mexican, European, Japanese, Uzbekistan, Russian. 6 Vietnamese accents include: North, South Central Coast, South East, South West, Central Highland, North Central Coast.

²Speech collected by simulated medical conversations between 2 speaking roles - clinicians and actors/actresses. 4 English accents include: British English, European, other English, and other non-English.

³Speech was recorded as patient-physician interviews (counted as 1 recording condition and 2 speaking roles) by West England speakers (counted as 1 accent)

⁴AfriSpeech-200 dataset is a mix of general-domain and medical-domain speech. To our best understanding of the paper, we estimate the total duration of medical-domain speech to be around 123 hours. Recordings were collected by crowd-sourced workers to read aloud the medical transcripts (also known as read speech), thus both the number of recording conditions and speaking roles are counted as 1.

⁵myMediCon dataset hired speakers to read aloud the translated medical transcripts from English corpus (thus known as read speech). 5 speakers' accents include: Native Burmese, Pa'O, Kachin, Dawei, and Mon. 2 speaking roles are patients and doctors.

Dataset	Size	Domain	Language	Direction
Neves (2017) ¹	23k pairs = 46k samples	clinical trials	en-pt	one-to-one
ParaMed ² (Liu and Huang, 2021)	100k pairs = 200k samples	medical documents	en-zh	one-to-one
Khresmoi ³ (Pecina et al., 2017)	1k5 pairs = 12k samples	medical documents	8 EU lang.	many-to-many
WMT Biomedical Task ⁴ (Bawden et al., 2020)	160k samples	medical documents	9 lang.	one-to-one
YuQ ⁵ (Yu et al., 2020)	65k pairs = 130k samples	medical articles	en-ug	one-to-one
Bérard et al. (2020) ⁶	1500 samples	COVID-19	en-kr	one-to-one
MedEV ⁷ (Vo et al., 2024)	18k pairs = 36k samples	medical documents	en-vi	one-to-one
 MultiMed-ST (ours)	48k pairs = 290k samples	medical conversations	5 lang.	many-to-many

Table 12: **Dataset comparison with literature: A list of all publicly available medical MT datasets.**

Our  MultiMed-ST is the first medical ST dataset, and is the largest medical MT dataset, to the best of our knowledge, given the fact that speech data is much more difficult to collect compared to medical text data.

¹ Text-only medical MT dataset for English - Portuguese

² Text-only medical MT dataset for English - Chinese crawled from the New England Journal of Medicine, thus leading to low diversity

³ Text-only medical MT dataset for 8 European languages: Czech, English, French, German, Hungarian, Polish, Spanish, and Swedish. The dataset requires users' costly payment.

⁴ Text-only medical MT dataset for 9 European languages: English, Basque, Chinese, French, German, Italian, Portuguese, Spanish, Russian

⁵ Text-only medical MT dataset for Chinese-Uyghur, covering seven clinical disciplines and five sense organs science

⁶ Text-only medical MT test set for Korean-English, collected from official COVID-19 guidelines and recent papers

⁷ Text-only medical MT dataset for English - Vietnamese, containing 18k high-quality sentence pairs as dev and test set. The rest training data was not quality-controlled by human annotators.

Dataset	Size	Domain	Language	Direction
BhasaAnuvaad ¹ (Jain et al., 2024)	47k samples	general-domain, spontaneous	en-13 Indic lang.	one-to-many
Europarl-ST ² (Iranzo-Sánchez et al., 2020)	200k samples	parliamentary debates	6 EU lang.	many-to-many
MaSS ³ (Boito et al., 2020)	8k samples	bible	8 lang.	many-to-many
Fisher & Callhome ⁴ (Post et al., 2013)	170k samples	telephone, spontaneous	en-es	one-to-one
BSTC ⁵ (Zhang et al., 2021a)	40k samples	various TED-like domains	en-zh	one-to-one
 MultiMed-ST (ours)	290k samples	medical conversations	5 lang.	many-to-many

Table 13: **Dataset comparison with literature: A list of some of the largest publicly available medical ST datasets.**

Although medical ST data is widely known to be very difficult to collect, **our  MultiMed-ST is as large as popular large-scale general-domain ST datasets.** Although ours is not the largest among all existing ST datasets, **our  MultiMed-ST is the largest many-to-many multilingual ST datasets.**

¹ Bidirectional ST dataset from English into 13 Indian languages, known as the largest Indic language ST dataset

² Many-to-many multilingual ST dataset, covering English, German, French, Spanish, Italian and Portuguese. The domain is about parliamentary debates, thus leading to low diversity

³ Clean ST dataset extracted from the Bible, covering English, Spanish, Basque, Finnish, French, Hungarian, Romanian, and Russian

⁴ Crowd-sourced Spanish-English ST dataset derived from two costly ASR datasets Fisher (Cieri et al., 2004) and Callhome

⁵ The first large-scale Chinese-English ST dataset, containing 68 hours of mandarin speeches from three TED-like content producers

D Details of Experimental Setup

D.1 Training Setup: Whisper

Whisper, a Transformer-based AED (see Appendix Section B.2), is an end-to-end multitask ASR and ST model pre-trained on 680k hours of labeled data. Approximately 65% of the data (equivalent to 438,000 hours) consists of English-language audio paired with English transcripts. Around 18% (or 126,000 hours) comprises non-English audio with English transcripts, while the remaining 17% (or 117,000 hours) includes non-English audio along with their corresponding transcripts. The non-English data encompasses 98 distinct languages.

For ASR, we performed a full fine-tuning (both encoder and decoder) monolingually (each language separately) and multilingually (all languages simultaneously). For ST, we performed a full fine-tuning bilingually (each language pair separately) and multilingually (all language pairs simultaneously).

Whisper variants: We employed 2 variants of Whisper models: Whisper-small²⁰ (244M parameters) and Whisper-large-v2²¹ (1550M parameters). Figure 6 and Figure 7 show the fine-tuning configuration of Whisper-small model and Whisper-large-v2 model respectively.

```
1. MODEL_NAME="openai/whisper-small"
2. SAMPLING_RATE=16000
3. NUM_PROC=2
4. TRAIN_STRATEGY="steps"
5. LEARNING_RATE=1e-5
6. WARMUP=500
7. TRAIN_BATCHSIZE=8
8. EVAL_BATCHSIZE=8
9. NUM_STEPS=5000
```

Figure 6: Fine-tuning configuration of **Whisper-small** model

Pre-processing setup: To preprocess data for the Whisper models, you must prepare audio files and their corresponding text transcriptions in a format suitable for training or fine-tuning. Begin by converting audio files to a consistent format (e.g., 16 kHz, mono-channel WAV files) to ensure compatibility. Use libraries like ffmpeg²² or librosa²³ for this purpose. Normalize and clean the tran-

```
1. MODEL_NAME="openai/whisper-large-v2"
2. SAMPLING_RATE=16000
3. NUM_PROC=2
4. TRAIN_STRATEGY="steps"
5. LEARNING_RATE=1e-5
6. WARMUP=500
7. TRAIN_BATCHSIZE=8
8. EVAL_BATCHSIZE=8
9. NUM_STEPS=4000
```

Figure 7: Fine-tuning configuration of **Whisper-large-v2** model

scriptions by removing extraneous characters, correcting spelling, and aligning timestamps with the audio. Tokenize the text using Whisper’s tokenizer, ensuring it matches the pre-trained model’s vocabulary. Additionally, segment long audio files into smaller chunks with overlapping windows to fit the model’s input length constraints while preserving context. Finally, package the processed audio-text pairs into a dataset format such as JSON, which includes metadata like file paths, transcription text, and optional timestamps for alignment.

Training setup:

- **Whisper-small:** As shown in Figure 6, the training configuration for the openai/whisper-small model is detailed as follows. The model was trained using a sampling rate of 16,000 Hz (SAMPLING_RATE) to process audio data effectively. The training utilized 2 processors (NUM_PROC) to parallelize computations. A step-based training strategy (TRAIN_STRATEGY="steps") was adopted, where the model was trained for 5,000 steps (NUM_STEPS). The learning rate was set to 1×10^{-5} (LEARNING_RATE), with a warmup period of 500 steps (WARMUP) to stabilize training. Both the training and evaluation batch sizes were configured as 8 (TRAIN_BATCHSIZE and EVAL_BATCHSIZE, respectively) to ensure efficient memory usage while maintaining model performance.
- **Whisper-large-v2:** As shown in Figure 7, the training setup for fine-tuning the Whisper model leverages the openai/whisper-large-v2 architecture. The audio inputs are resampled to a sampling rate of 16,000 Hz (SAMPLING_RATE=16000) for consistency with the model’s requirements. Training is distributed across two processing units (NUM_PROC=2) using a step-based train-

²⁰<https://huggingface.co/openai/whisper-small>

²¹<https://huggingface.co/openai/whisper-large-v2>

²²<https://www.ffmpeg.org/>

²³<https://librosa.org/>

```

AUDIO_URL = {
    "url": "https://static.deepgram.com/examples/Bueller-Life-moves-pretty-fast.wav"
}

## STEP 1 Create a Deepgram client using the API key from environment variables
deepgram: DeepgramClient = DeepgramClient("", ClientOptionsFromEnv())

## STEP 2 Call the transcribe_url method on the prerecorded class
options: PrerecordedOptions = PrerecordedOptions(
    model="nova-2",
    smart_format=True,
)
response = deepgram.listen.rest.v("1").transcribe_url(AUDIO_URL, options)
print(f"response: {response}\n\n")

```

Figure 8: **Deepgram API call.** The model employed in our experiments is Nova-2 version. We used the API to directly recognize the audio files instead of fine-tuning.

ing strategy (TRAIN_STRATEGY="steps"). The learning rate is set to a modest value of 1×10^{-5} (LEARNING_RATE=1e-5) with a warm-up phase spanning 500 steps (WARMUP=500) to stabilize the optimization process. The training batch size and evaluation batch size are both configured to 8 (TRAIN_BATCHSIZE=8 and EVAL_BATCHSIZE=8, respectively), balancing computational efficiency with memory constraints. The total number of training steps is capped at 4,000 (NUM_STEPS=4000), ensuring effective model convergence without overfitting.

D.2 Training Setup: Deepgram

Deepgram²⁴ is also a Transformer-based AED architecture. We employed the Nova-2 version, which supports over 30 languages. Since this is a commercial API, we could only employ direct recognition on audio files instead of fine-tuning, as shown in Figure 8.

D.3 Training Setup: AssemblyAI

AssemblyAI²⁵ is a Conformer-based RNN-T architecture. We used Universal-2 with 600M parameters, pre-trained on from 150,000 hours to 300,000 hours of supervised multilingual data.

Special Tokenization²⁶: RNN-T demonstrates a constrained capacity to generate consecutive identical tokens. Previous studies (Ghodsí et al., 2020; Xu et al., 2023b) have indicated that RNN-T pos-

sesses a pronounced inductive bias that inhibits the prediction of identical tokens in succession. Therefore, Universal-2 incorporates a unique `<repeat_token>` within its tokenization scheme, which is inserted between repeated tokens in the target sequences of the training data. This modification eliminates the need for the RNN-T model to predict the same token multiple times consecutively. Consequently, it enables accurate recognition of repeated tokens without deletions, addressing a limitation of the RNN-T architecture. During inference, the `<repeat_token>` is removed from the final ASR output.

RNN-T: The RNN-T encoder was pre-trained on 12.5 million hours of diverse, multilingual audio data. Following the pre-training phase, the encoder was integrated with a randomly initialized decoder, and the complete model underwent fine-tuning utilizing a combination of the aforementioned supervised dataset and a pseudo-labeled dataset.

Text Formatting: The Text Formatting module processes raw transcripts into well-structured text by incorporating Punctuation Restoration, Truecasing, and Inverse Text Normalization (ITN), ensuring the final output is both highly readable and adaptable for diverse applications, as shown in Figure 9.

Text Formatting Architecture: Figure 10 shows the visualization of Universal-2 Text Formatting Architecture. The architecture is described as below:

- **Token-based Truecasing:** Universal-1 employed a character-based model for Truecasing, which exhibited susceptibility to hallu-

²⁴<https://deepgram.com/>

²⁵<https://www.assemblyai.com/>

²⁶<https://www.assemblyai.com/research/universal-2>

before winning her council seat in 1975, to which she was reelected in 1970, 919 80, 319 80, 719 90, 119 95 and 1999, coming in first each time and we will return at 07:26 p.m. 54321. Happy holidays. The history That's Denver YmCa.org Slash University Dash hills dash redevelopment. It's going to be a major project Donna wins both the showcases. Sixty five thousand, \$661 worth. Prizes today for Donna. Not bad. what are you gonna bid? Two thousand dollars. Two thousand dollars. Mark, 2250 dollars. Two thousand 250 dollars. Robin? Two thousand 200 fifty, one dollars. Two thousand 200 fifty, one dollars. Donna? Dollar. One dollar. Dollar one. Donna bids a dollar. Bottom trawling and bottom gill nets a year ago in regulation no. Nineteen 50 four, 2000 three. or www. Dot stopbribes.org or www. Dot preventcorruption dot org. Fifteen thousand dollars. Twenty 121 thousand dollars. Twenty 1000 dollars. Yeah. No, no, no, no, no, wait. No, I mean, three, 300 dollars. Twenty one thousand, 300 dollars. Twenty one thousand, 300 dollars. How much should I pay? Twelve dollars 75 cents, please. Thanks a lot. Here's 15 dollars.

before winning her council seat in 1975, to which she was reelected in 1979, 1983, 1987, 1991, 1995 and 1999, coming in first each time and we will return at 7:26pm 5, 4, 3, 2, 1. Happy Holidays. The history That's Denver ymca.org/university-hills-redevelopment It's going to be a major project Donna wins both the showcases. \$65,661 worth prizes today for Donna. Not bad. what are you gonna bid? \$2,000. \$2,000. Mark. \$2,250. \$2,250. Robin. \$2,251. \$2,251. Donna. \$1. \$. \$1. Donna bids a dollar. Bottom trawling and bottom gill nets a year ago in Regulation No. 1954, 2003. or www.stopbribes.org or www.PreventCorruption.org. \$15,000. 21. \$21,000. \$21,000? Yeah. No, no, no, no, no, wait. No, I mean three. \$300. \$21,300. \$21,300. How much should I pay? \$12.75, please. Thanks a lot. Here's \$15.

Figure 9: An example of text formatting in AssemblyAI’s Universal-2. Green text is the final ASR output and red text is the ASR output before it is processed by text formatting module.

cination errors and incurred increased computational overhead. Universal-2 switched to token-based modeling resulting in more accurate Truecasing with reduced computational demands.

- seq2seq Modeling for ITN: Universal-2 employs a seq2seq model, which more effectively captures contextual information for ITN compared to a rule-based approach.
- Multi-objective tagging model: The model comprises a shared Transformer encoder, followed by three distinct classification heads designed to perform specific tasks: (1) post-punctuation prediction, (2) token-level truecasing to address all-uppercase, all-lowercase, word capitalization, and mixed-case word identification, and (3) textual span detection for ITN processing.
- Text span conversion model: The seq2seq model employs a Transformer-based encoder-decoder architecture and is utilized to process normalized mixed-case and ITN spans identified by the multi-objective tagging model, generating their corresponding formatted representations.

D.4 Training Setup: mBART

mBART-50 (Tang et al., 2020) is a multilingual seq2seq model designed to demonstrate the feasibility of creating multilingual MT models through multilingual fine-tuning. Rather than fine-tuning the model for a single translation direction, it is

fine-tuned across multiple translation directions simultaneously. The mBART-50 model extends the original mBART framework by incorporating 25 additional languages, enabling support for multilingual MT across 50 languages. The pre-trained model, mBART-large-50, is primarily optimized for fine-tuning on MT but can also be adapted for other multilingual seq2seq applications.

Pre-processing setup: Due to the multilingual nature of the model, it requires input sequences to adhere to a specific format. A unique language identifier token is employed as a prefix in both the source and target texts. The format for the text is $[lang_code]X[eos]$, where X represents the source or target text, and $lang_code$ corresponds to the source language code for the source text and the target language code for the target text. The beginning-of-sequence (bos) token is not utilized. Once the examples are formatted in this manner, the model can be trained as a standard seq2seq model. Pre-processing might also involve cleaning data (removing noise, handling encoding issues), truncating or padding sentences to the maximum sequence length supported by the model, and batching data for efficient processing.

Training setup: As shown Figure 11, the training setup for the mBART-large-50 model is designed to optimize performance with a range of hyperparameters. The learning rate is set to a relatively low value of $1e-5$, ensuring fine-tuning without overshooting optimal solutions. The learning rate schedule follows a cosine annealing with restarts strategy, allowing for periodic adjustments to prevent overfitting as training progresses. A

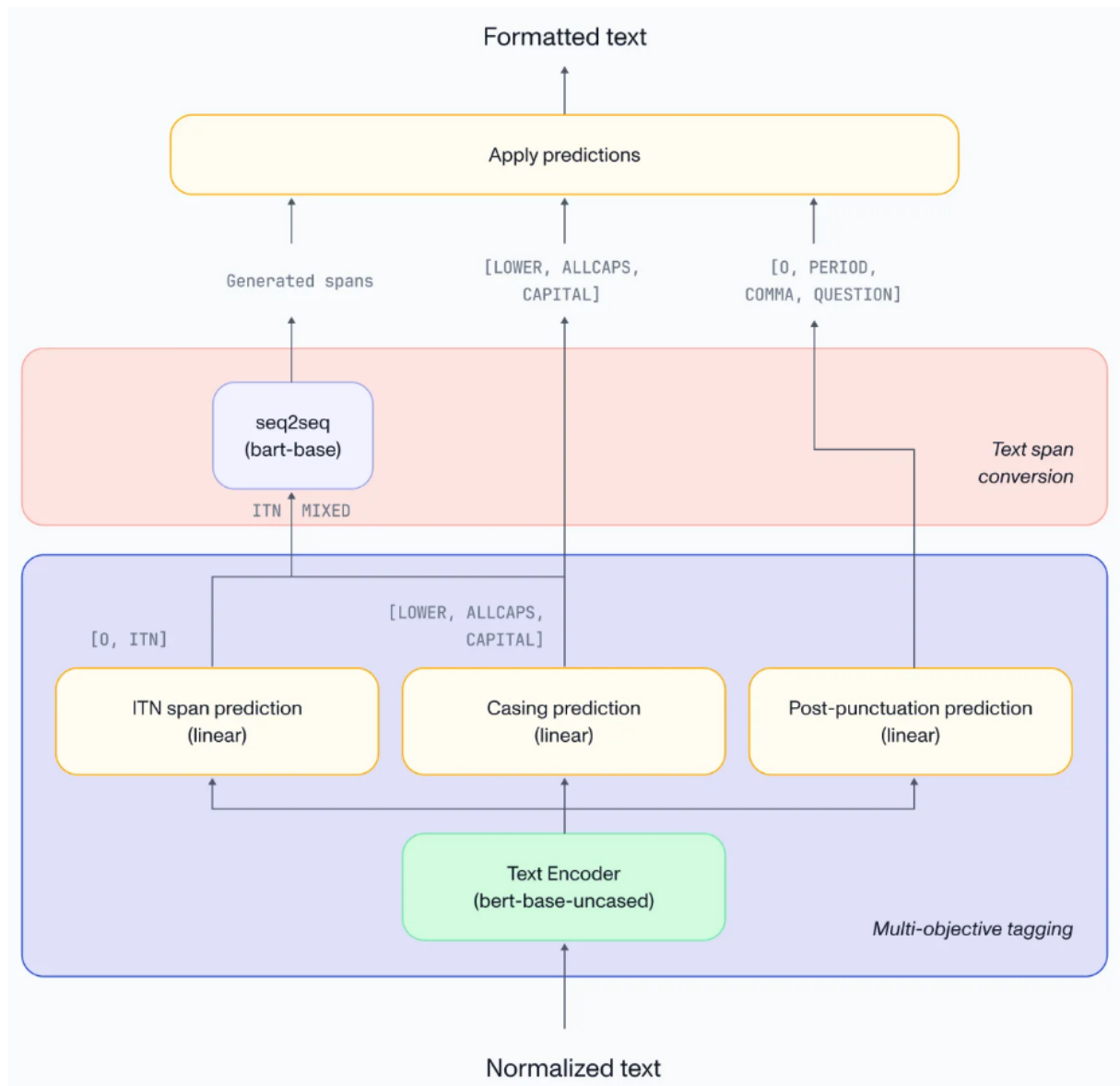


Figure 10: AssemblyAI's Universal-2 Text Formatting Architecture.

```

1. learning_rate=1e-5,
2. lr_scheduler_type="cosine_with_restarts",
3. per_device_train_batch_size=32,
4. per_device_eval_batch_size=32,
5. weight_decay=0.03,
6. label_smoothing_factor=0.1,
7. num_train_epochs=10,
8. warmup_steps=225,
9. max_input_length=256

```

Figure 11: Fine-tuning configuration of **mBART-large-50** model

batch size of 32 for both training and evaluation is chosen to balance computational efficiency and model convergence. Regularization is applied with a weight decay of 0.03, and label smoothing of 0.1 is used to help the model generalize better by softening the target labels. Training runs for 10 epochs, providing ample opportunity for the model to adapt to the dataset. A warmup phase with 225 steps is included to gradually ramp up the learning rate and avoid instability at the start of training. The maximum input sequence length is capped at 256 tokens, optimizing memory usage while accommodating most text sequences.

```

1. learning_rate=2e-5,
2. lr_scheduler_type="cosine_with_restarts",
3. per_device_train_batch_size=32,
4. per_device_eval_batch_size=32,
5. weight_decay=0.03,
6. label_smoothing_factor=0.1,
7. num_train_epochs=10,
8. warmup_steps=225,
9. max_input_length=256

```

Figure 12: Fine-tuning configuration of **M2M100-148M** model

D.5 Training Setup: M2M100

M2M100-418M is a multilingual encoder-decoder model designed for many-to-many multilingual MT. This model is capable of directly translating across 9,900 translation directions involving 100 languages.

Preprocessing setup: To translate into a target language, the target language identifier (id) is designated as the first generated token. This can be achieved by specifying the *forced_bos_token_id* parameter in the *generate* method. The M2M100Tokenizer relies on SentencePiece²⁷ (SPM). All datasets must undergo detokenization prior to the application of SPM during the data pre-processing phase. Following the download of raw data, it is necessary to post-process the data, apply SPM, and then binarize the dataset.

Training setup: As shown in Figure 12, the learning rate was set to 2×10^{-5} , with a *cosine_with_restarts* learning rate scheduler employed to dynamically adjust the learning rate during training. A per-device batch size of 32 was used for both training and evaluation to balance computational efficiency and memory usage. Weight decay was applied with a factor of 0.03 to regularize the model and prevent overfitting. To further enhance generalization, a label smoothing factor of 0.1 was introduced. The model was trained for 10 epochs, with the first 225 steps dedicated to warmup to allow a gradual ramp-up of the learning rate. Additionally, the maximum input length for sequences was capped at 256 tokens to ensure efficient processing of data. This training configuration was chosen to achieve optimal performance on multilingual MT tasks.

D.6 Training Setup: Marian

Marian is an encoder-decoder fine-tuned on one-to-one translation task, which is built upon BART

architecture. The original Marian is a highly efficient and open-source NMT framework, implemented in pure C++ with minimal external dependencies. Its development is primarily led by the Microsoft Translator team. All models are transformer encoder-decoders with 6 layers in each component.

```

1. learning_rate=2e-5,
2. lr_scheduler_type="cosine_with_restarts",
3. per_device_train_batch_size=32,
4. per_device_eval_batch_size=32,
5. weight_decay=0.03,
6. label_smoothing_factor=0.1,
7. num_train_epochs=10,
8. warmup_steps=225,
9. max_input_length=256

```

Figure 13: Fine-tuning configuration of **Marian** model

We employed the Python Hugging Face version for training instead of C++ version, as shown in Figure 14.

Training setup: As shown in Figure 13, the learning rate was set to 2×10^{-5} , and a cosine learning rate scheduler with restarts (*cosine_with_restarts*) was utilized to adjust the learning rate dynamically during training. The batch size for both training and evaluation was fixed at 32 samples per device (*per_device_train_batch_size* and *per_device_eval_batch_size*). A weight decay value of 0.03 was applied to mitigate overfitting, and label smoothing with a factor of 0.1 was incorporated to improve model generalization. The model was trained for a total of 10 epochs (*num_train_epochs*), with 225 warmup steps (*warmup_steps*) to stabilize the optimization process. Additionally, the maximum input sequence length was restricted to 256 tokens (*max_input_length*) to efficiently handle the computational requirements.

D.7 Training Setup: Llama

The Meta Llama 3.1 series comprises a collection of multilingual LLMs designed for text-based input and output. These pre-trained and instruction-tuned generative models, specifically the 8B parameter variant, are optimized for multilingual dialogue applications. The Llama 3.1 instruction-tuned models demonstrate superior performance compared to numerous open-source and proprietary conversational models across standard industry benchmarks. In our experiments, we employed Llama-3.1-8B model.

²⁷<https://pypi.org/project/sentencepiece/>

```

from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

tokenizer = AutoTokenizer.from_pretrained("Helsinki-NLP/opus-mt-en-de")

model = AutoModelForSeq2SeqLM.from_pretrained("Helsinki-NLP/opus-mt-en-de")

```

Figure 14: **Python Hugging Face version.** Marian C++ version is much more efficient for training but it is more difficult to train and deploy. Thus in the scope of our experiments, we only used Hugging Face implementation.

Model architecture: Llama-3.1-8B is an autoregressive language model (decoder-only model) designed with an optimized transformer architecture. The fine-tuned variants employ Supervised Fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to enhance alignment with human preferences for both helpfulness and safety.

Training setup: As shown in Figure 16, the training setup for the Llama-3.1-8B model was configured with a per-device training batch size of 4 and a gradient accumulation of 4 steps to effectively utilize hardware resources. The training process involved a warmup phase consisting of 1 step, followed by a single training epoch. The maximum number of training steps was set to 200, with an extended English configuration allowing up to 300 steps. A learning rate of 2×10^{-4} was used, alongside mixed-precision training, where FP16 was enabled if BF16 support was unavailable. Conversely, BF16 was activated on supported devices. The AdamW optimizer in its 8-bit variant was employed, with a weight decay of 0.01 to mitigate overfitting. A linear learning rate scheduler was adopted, and logging was performed at every step to ensure detailed progress tracking throughout the training process.

D.8 Training Setup: Qwen

The Qwen2.5 LLMs have been pre-trained on a newly developed large-scale dataset comprising up to 18 trillion tokens, representing a substantial expansion compared to Qwen2. This enhanced pre-training has endowed Qwen2.5 with significantly improved capabilities, including advanced instruction-following, the ability to generate extended texts exceeding 8,000 tokens, improved comprehension of structured data (e.g., tables), and enhanced generation of structured outputs, particularly in JSON format. Qwen2.5 supports a context length of up to 128,000 tokens and can produce out-

puts of up to 8,000 tokens. Additionally, these models maintain multilingual functionality, encompassing more than 29 languages, such as Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, and Vietnamese.

We employed Qwen-2.5-7B version. Its model card is shown in Figure 17 and its Hugging Face implementation is shown in Figure 18. Qwen-2.5-7B could also be run locally via Ollama²⁸ service. However, in the scope of our experiments, we only used Hugging Face for training.

Training setup: As shown in Figure 16, the training setup for the Qwen-2.5-7B model was configured with a per-device training batch size of 4 and a gradient accumulation of 4 steps to effectively utilize hardware resources. The training process involved a warmup phase consisting of 1 step, followed by a single training epoch. The maximum number of training steps was set to 200, with an extended English configuration allowing up to 300 steps. A learning rate of 2×10^{-4} was used, alongside mixed-precision training, where FP16 was enabled if BF16 support was unavailable. Conversely, BF16 was activated on supported devices. The AdamW optimizer in its 8-bit variant was employed, with a weight decay of 0.01 to mitigate overfitting. A linear learning rate scheduler was adopted, and logging was performed at every step to ensure detailed progress tracking throughout the training process.

D.9 Training Setup: Mistral

Mistral 7B is an LLM consisting of 7 billion parameters, developed and released by Mistral AI. This model has been meticulously engineered to offer a balance of computational efficiency and high performance, making it suitable for practical applications. Upon its release, Mistral 7B demonstrated

²⁸<https://github.com/ollama/ollama>

	Training Data	Params	Input modalities	Output modalities	Context length	GQA	Token count	Knowledge cut off
Llama 3.1 (text only)	A new mix of publicly available online data.	8B	Multilingual Text	Multilingual Text and code	128k	Yes	15T+	December 2023
		70B	Multilingual Text	Multilingual Text and code	128k	Yes		
		405B	Multilingual Text	Multilingual Text and code	128k	Yes		

Figure 15: **Model card of Llama 3.1 family.** The Llama 3.1 family of models was pre-trained on approximately 15 trillion tokens sourced from publicly available datasets, with token counts reflecting pre-training data exclusively. All versions of Llama 3.1 utilize Grouped Query Attention (GQA) to enhance inference scalability. Fine-tuning was conducted using a combination of publicly available instruction datasets and over 25 million synthetically generated examples. The pre-training dataset has a cutoff date of December 2023.

```

1. per_device_train_batch_size = 4,
2. gradient_accumulation_steps = 4,
3. warmup_steps = 1,
4. num_train_epochs = 1,
5. max_steps = 200, (with english max_steps=300)
6. learning_rate = 2e-4,
7. fp16 = not is_bfloat16_supported(),
8. bf16 = is_bfloat16_supported(),
9. logging_steps = 1,
10. optim = "adamw_8bit",
11. weight_decay = 0.01,
12. lr_scheduler_type = "linear"

```

Figure 16: Fine-tuning configuration of **LLM** model

superior performance across all evaluated benchmarks, surpassing the leading open-source 13B-parameter model, Llama 2. We employed Mistral-v0.3-7B²⁹ version.

Model architecture: The model incorporates attention mechanisms such as

- GQA: which enhances inference speed and reduces memory usage during decoding
- Sliding Window Attention (SWA) (Child et al., 2019; Beltagy et al., 2020): enabling the processing of sequences of arbitrary length while minimizing inference cost, in which each layer attends to the previous 4,096 hidden states. The primary advancement, and

the primary motivation for the initial investigation, is the linear computational cost of $O(\text{sliding_window.seq_len})$.

Sliding window attention leverages the hierarchical structure of transformer layers to extend the receptive field beyond the fixed window size. Specifically, a token i at layer k attends to the tokens in the range $[i - \text{sliding_window}, i]$ at layer $k - 1$. These attended tokens, in turn, have attended to tokens in the range $[i - 2 \times \text{sliding_window}, i]$ at layer $k - 2$. As a result, higher layers are able to access information from tokens further in the past than what the local attention pattern of the current layer suggests.

Finally, a constrained attention span allows for the limitation of the cache size to that of a sliding window of tokens, facilitated by the use of rotating buffers. This approach reduces the cache memory requirement by 50% for inference on sequences of length 8192, without compromising model performance.

- FlashAttention (Dao et al., 2022; Dao, 2024) and xFormers (Lefaudeux et al., 2022): In practice, changes made to FlashAttention and xFormers yield a 2x speed improvement for sequence length of 16k with a window of 4k.

²⁹<https://huggingface.co/mistralai/Mistral-7B-v0.3>

Models	# Params (B)	# Non-Emb Params (B)	# Layers	# Head (Q / KV)	# Tie Embedding	Context Length	Generation Length	License
Qwen2.5-0.5B	0.49	0.36	24	14/2	Yes	32K	8K	Apache 2.0
Qwen2.5-1.5B	1.5	1.3	28	12/2	Yes	32K	8K	Apache 2.0
Qwen2.5-3B	3.1	2.8	36	16/2	Yes	32K	8K	Qwen Research
Qwen2.5-7B	7.6	6.5	28	28/4	No	128K	8K	Apache 2.0
Qwen2.5-14B	14.7	13.1	48	40/8	No	128K	8K	Apache 2.0
Qwen2.5-32B	32.5	31.0	64	40/8	No	128K	8K	Apache 2.0
Qwen2.5-72B	72.7	70.0	80	64/8	No	128K	8K	Qwen
Qwen2.5-Coder-1.5B	1.5	1.3	28	12/2	Yes	128K	2K	Apache 2.0
Qwen2.5-Coder-7B	7.6	6.5	28	28/4	No	128K	2K	Apache 2.0
Qwen2.5-Math-1.5B	1.5	1.3	28	12/2	Yes	4K	2K	Apache 2.0
Qwen2.5-Math-7B	7.6	6.5	28	28/4	No	4K	2K	Apache 2.0
Qwen2.5-Math-72B	72.7	70.0	80	64/8	No	4K	2K	Qwen

Figure 17: **Model card of Qwen2.5 family.** The training setup for Qwen-2.5-7B features a causal language model architecture. This model employs transformers with various advanced components, including RoPE (Rotary Positional Embeddings), SwiGLU activation functions, RMSNorm normalization, and Attention QKV bias. With a total of 7.61 billion parameters, the Qwen2.5 model has 6.53 billion parameters dedicated to non-embedding components. The model consists of 28 layers, with attention heads configured as 28 for the Query (Q) and 4 for Key-Value (KV). The model is designed for a context length of 131,072 tokens, allowing for processing of long-range dependencies in text sequences during pre-training.

Training setup: As shown in Figure 16, the training setup for the Mistral-v0.3-7B model was configured with a per-device training batch size of 4 and a gradient accumulation of 4 steps to effectively utilize hardware resources. The training process involved a warmup phase consisting of 1 step, followed by a single training epoch. The maximum number of training steps was set to 200, with an extended English configuration allowing up to 300 steps. A learning rate of 2×10^{-4} was used, alongside mixed-precision training, where FP16 was enabled if BF16 support was unavailable. Conversely, BF16 was activated on supported devices. The AdamW optimizer in its 8-bit variant was employed, with a weight decay of 0.01 to mitigate overfitting. A linear learning rate scheduler was adopted, and logging was performed at every step to ensure detailed progress tracking throughout the training process.

D.10 Training Setup: Google Translate

Using Google Translate as a MT model in a cascaded ST system can provide a powerful and scalable solution for real-time multilingual communica-

tion. In a cascaded ST setup, the process typically involves two stages: first, ASR module converts audio into text, and then an MT model like Google Translate is used to render that text into the desired language.

By leveraging Google Translate, which is backed by advanced NMT techniques, the system can provide high-quality, context-aware translations. The integration of Google Translate into the ST system offers several benefits, including the ability to handle a wide range of language pairs, rapid updates, and continuous improvements due to the vast data the system processes. Additionally, Google Translate has been trained on massive multilingual corpora, which helps it deal with diverse linguistic nuances and idiomatic expressions.

However, this approach also comes with challenges. One potential issue is that the quality of the ASR output plays a critical role in the overall effectiveness of the MT. If the ASR system produces too many errors or misinterprets the audio, Google Translate will likely propagate these errors, leading to inaccuracies in the final trans-

```

1. from transformers import AutoModelForCausalLM, AutoTokenizer
2.
3. model_name = "Qwen/Qwen2.5-7B-Instruct"
4.
5. model = AutoModelForCausalLM.from_pretrained(
6.     model_name,
7.     torch_dtype="auto",
8.     device_map="auto"
9. )
10. tokenizer = AutoTokenizer.from_pretrained(model_name)
11.
12. prompt = "Give me a short introduction to large language model."
13. messages = [
14.     {"role": "system", "content": "You are Qwen, created by Alibaba Cloud. You are a helpful assistant."},
15.     {"role": "user", "content": prompt}
16. ]
17. text = tokenizer.apply_chat_template(
18.     messages,
19.     tokenize=False,
20.     add_generation_prompt=True
21. )
22. model_inputs = tokenizer([text], return_tensors="pt").to(model.device)
23.
24. generated_ids = model.generate(
25.     **model_inputs,
26.     max_new_tokens=512
27. )
28. generated_ids = [
29.     output_ids[len(input_ids):] for input_ids, output_ids in zip(model_inputs.input_ids, generated_ids)
30. ]
31.
32. response = tokenizer.batch_decode(generated_ids, skip_special_tokens=True)[0]

```

Figure 18: **Hugging Face implementation of Qwen-2.5-7B model.** The implementation for Qwen-2.5-7B is also conducted via Ollama to run locally. We can also access the Ollama service via its OpenAI-compatible API. However, in the scope of our experiments, we only used Hugging Face for training.

lated output. Furthermore, Google Translate may struggle with medical-domain language or highly medical content, which may require fine-tuning or customization to ensure higher MT accuracy.

Despite these challenges, using Google Translate in a cascaded ST system remains a viable and practical solution for multilingual communication, especially when quick deployment and ease of integration are paramount. It is also an ideal solution when working with a wide array of languages, as Google Translate supports over 100 languages, making it adaptable to diverse linguistic needs.

D.11 Training Setup: VinAI Translate

The pre-trained VinAI Translate models represent state-of-the-art systems for Vietnamese-to-English and English-to-Vietnamese text translation. The platform features a user-friendly, interactive interface and incorporates advanced models for ASR, MT, and text-to-speech (TTS). Experimental results demonstrate that the system achieves state-of-the-art performance, surpassing Google Translate in both automated and human evaluations on publicly available Vietnamese-English translation benchmarks.

In our experiments, we only leveraged the MT module of VinAI Translate for our bilingual Vietnamese-English cascaded ST systems.

```

1. {
2.     "name_or_path": "vinai/vinai-translate-en2vi",
3.     "activation_function": "gelu",
4.     "add_bias_logits": false,
5.     "add_final_layer_norm": true,
6.     "architectures": [
7.         "MBartForConditionalGeneration"
8.     ],
9.     "d_model": 1024,
10.    "decoder_attention_heads": 16,
11.    "decoder_ffn_dim": 4096,
12.    "decoder_layers": 12,
13.    "encoder_attention_heads": 16,
14.    "encoder_ffn_dim": 4096,
15.    "encoder_layers": 12,
16.    "is_encoder_decoder": true,
17.    "max_length": 1024,
18.    "max_position_embeddings": 1024,
19.    "model_type": "mbart",
20.    "normalize_before": true,
21.    "normalize_embedding": true,
22.    "num_beams": 5,
23.    "num_hidden_layers": 12,
24.    "output_past": true,
25.    "pad_token_id": 1,
26.    "static_position_embeddings": false,
27.    "vocab_size": 91408
28. }

```

Figure 19: Fine-tuning configuration of **VinAI Translate** model

The pre-trained MT component: The pre-trained seq2seq model mBART is first fine-tuned using 3M high-quality English-Vietnamese sentence pairs from the PhoMT dataset (Doan et al., 2021) for English-to-Vietnamese translation. Subsequently, the fine-tuned model is employed to translate English sentences from "noisy" datasets

into Vietnamese. Sentence pairs with BLEU scores in the range of 0.15 to 0.95 are selected, resulting in an additional 6M pairs. Combining these with the initial 3M pairs yields a total of 9M high-quality sentence pairs. To simulate ASR output, this dataset is augmented for each translation direction (English-to-Vietnamese and Vietnamese-to-English) by applying lowercase conversion and punctuation removal to the source sentences while keeping the target sentences unchanged. This augmentation adds another 9M sentence pairs for each direction, resulting in 18M sentence pairs per direction. The mBART model is then fine-tuned for each translation direction using the full 18M sentence pairs to develop the machine translation MT component.

Training setup: As shown in Figure 19, we adopted pre-trained MT model from VinAI Translate to fine-tune on our own dataset. The model utilizes the MBartForConditionalGeneration framework with a transformer-based encoder-decoder structure. The encoder and decoder each consist of 12 layers, with 16 attention heads per layer, and a Feed-Forward Network (FFW) dimension of 4096. The hidden layer size is set to 1024, with a maximum sequence length and position embeddings capped at 1024. The activation function employed is gelu, and both embedding and layer normalization are applied before each layer. The model includes a vocabulary size of 91,408 tokens and does not use static position embeddings. During generation, the beam search decoding strategy is employed with 5 beams. Additional features include the use of a pad token ID of 1, final layer normalization, and bias-free logits.

D.12 Training Setup: EnViT5

The EnViT5 model is a Text-to-Text Transformer based on the encoder-decoder architecture introduced within the T5 framework proposed by Raffel et al. (2020). For pre-training, the model utilizes the CC100 dataset, a monolingual dataset derived from web crawl data, as described by Wenzek et al. (2020). This corpus comprises monolingual data for over 100 languages. Subsequently, the model is fine-tuned using MTet, the largest publicly available parallel corpus for English-Vietnamese translation. MTet, as published along with EnViT5 model, contains 4.2 million high-quality training sentence pairs and includes a multi-domain test set curated by the Vietnamese research community.

Training setup: As shown in Figure 20, the

```

1. {
2.   "architectures": [
3.     "T5ForConditionalGeneration"
4.   ],
5.   "d_ff": 2048,
6.   "d_kv": 64,
7.   "d_model": 768,
8.   "is_encoder_decoder": true,
9.   "layer_norm_epsilon": 1e-06,
10.  "model_type": "t5",
11.  "num_decoder_layers": 12,
12.  "num_heads": 12,
13.  "num_layers": 12,
14.  "relative_attention_max_distance": 128,
15.  "relative_attention_num_buckets": 32,
16.  "tie_word_embeddings": false,
17.  "vocab_size": 50048
18. }
```

Figure 20: Fine-tuning configuration of **EnViT5** model

training setup utilizes a model architecture based on the T5ForConditionalGeneration class, designed for tasks requiring a Transformer-based encoder-decoder structure. The model configuration includes a hidden dimensionality (d_model) of 768, with FFW sublayers of size 2048 (d_ff) and key-value dimensionality (d_kv) of 64. It consists of 12 encoder layers and 12 decoder layers (num_layers and num_decoder_layers), each employing 12 attention heads (num_heads). Relative position embeddings are implemented with a maximum attention distance of 128 (relative_attention_max_distance) and 32 buckets (relative_attention_num_buckets). Layer normalization is applied with an epsilon value of 10^{-6} (layer_norm_epsilon). The model does not tie word embeddings (tie_word_embeddings = false) and supports a vocabulary size of 50,048 tokens (vocab_size).

D.13 Training Setup: SeamlessM4T

SeamlessM4T is a foundational, all-in-one massively multilingual and multimodal MT model designed to provide high-quality translations for both speech and text across nearly 100 languages. The SeamlessM4T framework supports the following tasks: Speech-to-speech translation (S2ST), Speech-to-text translation (S2TT), Text-to-speech translation (T2ST), Text-to-text translation (T2TT), ASR.

Key language support capabilities include: 101 languages for speech input, 96 languages for text input and output, 35 languages for speech output. As the first model of its kind, SeamlessM4T enables simultaneous S2ST and S2ST for multiple source and target languages.

```

1. {
2.   "activation_function": "relu",
3.   "adaptor_kernel_size": 8,
4.   "adaptor_stride": 8,
5.   "add_adapter": true,
6.   "architectures": [
7.     "SeamlessM4Tv2Model"
8.   ],
9.   "char_vocab_size": 10943,
10.  "conv_depthwise_kernel_size": 31,
11.  "decoder_attention_heads": 16,
12.  "decoder_ffn_dim": 8192,
13.  "decoder_layerdrop": 0.05,
14.  "decoder_layers": 24,
15.  "decoder_start_token_id": 3,
16.  "encoder_attention_heads": 16,
17.  "encoder_ffn_dim": 8192,
18.  "encoder_layerdrop": 0.05,
19.  "encoder_layers": 24,
20.  "feature_projection_input_dim": 160,
21.  "hidden_size": 1024,
22.  "is_encoder_decoder": true,
23.  "lang_embed_dim": 256,
24.  "layer_norm_eps": 1e-05,
25.  "max_position_embeddings": 4096,
26.  "num_adapter_layers": 1,
27.  "num_attention_heads": 16,
28.  "num_hidden_layers": 24,
29.  "position_embeddings_type": "relative_key",
30.  "sampling_rate": 16000,
31.  "speech_encoder_attention_heads": 16,
32.  "speech_encoder_chunk_size": 20000,
33.  "speech_encoder_hidden_act": "swish",
34.  "speech_encoder_intermediate_size": 4096,
35.  "speech_encoder_layers": 24,
36.  "spkr_embed_dim": 256,
37.  "vocab_size": 256102,
38. }

```

Figure 21: Fine-tuning configuration of **SeamlessM4T-large-v2** model

The latest version of SeamlessM4T incorporates multitask-UnitY2, featuring a non-autoregressive unit decoder and hierarchical upsampling to enhance data efficiency in predicting translation units. Additionally, the model includes the w2v-BERT 2.0 speech encoder, pre-trained on 4.5 million hours of unlabeled audio data. The multitask model has been fine-tuned with increased supervision using automatically aligned data pairs to improve performance, particularly for low-resource languages.

SeamlessM4T leverages the Efficient Monotonic Multihead Attention (EMMA) mechanism (Ma et al., 2023), allowing for low-latency generation of target translations without requiring complete source utterances, thereby enabling real-time MT capabilities.

Training setup: As shown in Figure 21, the SeamlessM4T-large-v2 model is designed to leverage a robust architecture with an encoder-decoder framework, incorporating 24 encoder and 24 decoder layers. The encoder and decoder use 16 attention heads, a hidden size of 1024, and FFW network dimensions of 8192. The activation function is set to ReLU, with layer_norm_eps config-

ured at 10^{-5} . The maximum position embeddings extend up to 4096, utilizing a relative key-based position embedding type. The system integrates a speech encoder with 24 layers, 16 attention heads, an intermediate size of 4096, and a swish activation function, operating on a sampling rate of 16 kHz and a chunk size of 20,000. Adapter layers are added with one adapter per layer, featuring a kernel size of 8, a stride of 8, and a depthwise convolution kernel size of 31. The model, based on the SeamlessM4Tv2Model architecture, supports a vocabulary size of 256,102 for text and 10,943 for characters. Additional features include speaker embeddings and language embeddings, both of dimension 256. To ensure stability, a dropout rate of 0.05 is applied to both encoder and decoder layers.

D.14 Training Setup: Qwen-Audio

```

1. {
2.   "architectures": [
3.     "Qwen2AudioForConditionalGeneration"
4.   ],
5.   "audio_config": {
6.     "model_type": "qwen2_audio_encoder",
7.     "num_mel_bins": 128,
8.     "encoder_layers": 32,
9.     "encoder_attention_heads": 20,
10.    "encoder_ffn_dim": 5120,
11.    "d_model": 1280,
12.    "activation_function": "gelu",
13.    "scale_embedding": false,
14.  },
15.  "audio_token_index": 151646,
16.  "model_type": "qwen2_audio",
17.  "text_config": {
18.    "intermediate_size": 11008,
19.    "max_position_embeddings": 8192,
20.    "model_type": "qwen2",
21.    "rope_theta": 10000,
22.    "rms_norm_eps": 1e-5,
23.    "sliding_window": 32768,
24.    "vocab_size": 156032
25.  },
26.  "vocab_size": 156032
27. }

```

Figure 22: Fine-tuning configuration of **Qwen2-Audio-7B-Instruct** model

The Qwen2-Audio series represents the latest advancements in the Qwen large audio-language model framework. This model is designed to process diverse audio signal inputs, enabling comprehensive audio analysis and generating direct textual responses based on spoken instructions. Qwen2-Audio supports over eight languages and dialects, including but not limited to Chinese, English, Cantonese, French, Italian, Spanish, German, and Japanese.

The Qwen language model and an audio encoder

serve as the foundational models. Multi-task pre-training is subsequently applied to achieve audio-language alignment, followed by supervised fine-tuning and direct preference optimization (DPO). These steps are designed to enhance the model’s performance on downstream tasks and align with human preferences.

Training setup: As shown in Figure 22, we fine-tuned Qwen2-Audio-7B-Instruct in an end-to-end ST manner. The training setup leverages the Qwen2AudioForConditionalGeneration architecture, which combines an advanced audio encoder and a text-based generative model. The audio encoder configuration (qwen2_audio_encoder) includes 128 mel bins for the MFCC, 32 encoder layers, 20 attention heads per layer, a FFW dimension of 5120, and a model dimension (d_model) of 1280. The encoder employs the GELU activation function and does not scale embeddings. For token-level alignment, the audio token index is set to 151646. The text model (qwen2) features an intermediate size of 11008, supports a maximum of 8192 positional embeddings, and uses rotary position encoding with a scaling factor (rope_theta) of 10000. Additionally, the sliding window size is 32768 to handle long-context text processing. Both the audio and text models share a vocabulary size of 156032.

D.15 In-context Learning Prompt

We present our prompt templates used in in-context learning experiments.

Prompt template for *SFT* on the entire dataset is shown in Figure 23 for Llama-3.1-8B, Figure 26 for Qwen-2.5-7B, and Figure 29 for Mistral-v0.3-7B.

Prompt template for *few-shot learning* is shown in Figure 24 for Llama-3.1-8B, Figure 27 for Qwen-2.5-7B, and Figure 30 for Mistral-v0.3-7B.

Prompt template for *zero-shot learning* is shown in Figure 25 for Llama-3.1-8B, Figure 28 for Qwen-2.5-7B, and Figure 31 for Mistral-v0.3-7B.



You are **Llama-3.1-8B**, a useful assistant by **Meta**.

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: Translate the following sentence from English to Vietnamese:

Input: As already said, I'm a wife, a mother, and a full-time finance professional with a daily three-hour round-trip commute. I'm an ME/CFS patient and I am one of the lucky ones.

Response: Tôi là một người vợ, một người mẹ và một chuyên gia tài chính toàn thời gian, với quãng đường di chuyển hàng ngày là ba tiếng đồng hồ đi và về. Tôi là bệnh nhân ME/CFS và tôi là một trong những người may mắn.

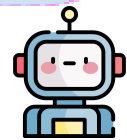


Figure 23: Prompt template for **SFT on the entire dataset** using **Llama-3.1-8B** model. This prompt is used for cascaded ST system.



You are **Llama-3.1-8B**, a useful assistant by **Meta**.

Translate the following sentence from English to Vietnamese:

Example 1:

###Input: I get a few, I get a, I get a couple of different lasers I get to do, but, um, uh, they, they do a lot more. So shout out those retina doctors with their lasers, their fancy lasers.

###Response: Tôi sử dụng một vài loại laser khác nhau, tôi có thể sử dụng một số trong số chúng, nhưng chúng làm được nhiều hơn thế. Vì vậy, hãy dành một lời cảm ơn đến các bác sĩ võng mạc với laser của họ, những laser tuyệt vời của họ.

Now, translate this text:

###Input: is activated, it also generates heat. And so this is also being used. But while these studies have been really interesting and intriguing in humans, in the end, if you really want to

Response:

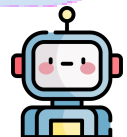


Figure 24: Prompt template for **few-shot learning** using **Llama-3.1-8B** model. This prompt is used for cascaded ST system.



You are **Llama-3.1-8B**, a useful assistant by **Meta**.

Translate the following sentence from English to Vietnamese:

###Input: is activated, it also generates heat. And so this is also being used. But while these studies have been really interesting and intriguing in humans, in the end, if you really want to

Response:

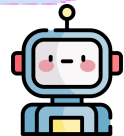


Figure 25: Prompt template for **zero-shot learning** using **Llama-3.1-8B** model. This prompt is used for cascaded ST system.



You are **Qwen-2.5-7B**, a useful assistant by **Alibaba**.

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: Translate the following sentence from English to Vietnamese:

Input: As already said, I'm a wife, a mother, and a full-time finance professional with a daily three-hour round-trip commute. I'm an ME/CFS patient and I am one of the lucky ones.

Response: Tôi là một người vợ, một người mẹ và một chuyên gia tài chính toàn thời gian, với quãng đường di chuyển hàng ngày là ba tiếng đồng hồ đi và về. Tôi là bệnh nhân ME/CFS và tôi là một trong những người may mắn.



Figure 26: Prompt template for **SFT on the entire dataset** using **Qwen-2.5-7B** model. This prompt is used for cascaded ST system.



You are **Qwen-2.5-7B**, a useful assistant by **Alibaba**.

Translate the following sentence from English to Vietnamese:

Example 1:

###Input: I get a few, I get a, I get a couple of different lasers I get to do, but, um, uh, they, they do a lot more. So shout out those retina doctors with their lasers, their fancy lasers.

###Response: Tôi sử dụng một vài loại laser khác nhau, tôi có thể sử dụng một số trong số chúng, nhưng chúng làm được nhiều hơn thế. Vì vậy, hãy dành một lời cảm ơn đến các bác sĩ võng mạc với laser của họ, những laser tuyệt vời của họ.

Now, translate this text:

###Input: is activated, it also generates heat. And so this is also being used. But while these studies have been really interesting and intriguing in humans, in the end, if you really want to

Response:

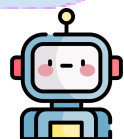


Figure 27: Prompt template for **few-shot learning** using **Qwen-2.5-7B** model. This prompt is used for cascaded ST system.



You are **Qwen-2.5-7B**, a useful assistant by **Alibaba**.

Translate the following sentence from English to Vietnamese:

###Input: is activated, it also generates heat. And so this is also being used. But while these studies have been really interesting and intriguing in humans, in the end, if you really want to

Response:

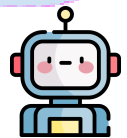


Figure 28: Prompt template for **zero-shot learning** using **Qwen-2.5-7B** model. This prompt is used for cascaded ST system.



You are **Mistral-v0.3-7B**, a useful assistant by **Mistral AI**.

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: Translate the following sentence from English to Vietnamese:

Input: As already said, I'm a wife, a mother, and a full-time finance professional with a daily three-hour round-trip commute. I'm an ME/CFS patient and I am one of the lucky ones.

Response: Tôi là một người vợ, một người mẹ và một chuyên gia tài chính toàn thời gian, với quãng đường di chuyển hàng ngày là ba tiếng đồng hồ đi và về. Tôi là bệnh nhân ME/CFS và tôi là một trong những người may mắn.

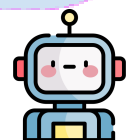


Figure 29: Prompt template for **SFT on the entire dataset** using **Mistral-v0.3-7B** model. This prompt is used for cascaded ST system.



You are **Mistral-v0.3-7B**, a useful assistant by **Mistral AI**.

Translate the following sentence from English to Vietnamese:

Example 1:

###Input: I get a few, I get a, I get a couple of different lasers I get to do, but, um, uh, they, they do a lot more. So shout out those retina doctors with their lasers, their fancy lasers.

###Response: Tôi sử dụng một vài loại laser khác nhau, tôi có thể sử dụng một số trong số chúng, nhưng chúng làm được nhiều hơn thế. Vì vậy, hãy dành một lời cảm ơn đến các bác sĩ võng mạc với laser của họ, những laser tuyệt vời của họ.

Now, translate this text:

###Input: is activated, it also generates heat. And so this is also being used. But while these studies have been really interesting and intriguing in humans, in the end, if you really want to

Response:



Figure 30: Prompt template for **few-shot learning** using **Mistral-v0.3-7B** model. This prompt is used for cascaded ST system.



You are **Mistral-v0.3-7B**, a useful assistant by **Mistral AI**.

Translate the following sentence from English to Vietnamese:

###Input: is activated, it also generates heat. And so this is also being used. But while these studies have been really interesting and intriguing in humans, in the end, if you really want to

Response:

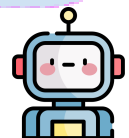


Figure 31: Prompt template for **zero-shot learning** using **Mistral-v0.3-7B** model. This prompt is used for cascaded ST system.

E Details of Evaluation Metrics

E.1 Discussion about Automatic Evaluation Metrics

In this section, we discuss the advantages and disadvantages of two types of automatic evaluation metrics in MT: n-gram overlap metrics (e.g. BLEU, METEOR, etc.) and embeddings-based metrics (e.g. BERTScore)

N-gram overlap metrics:

- Advantages:
 - Simplicity and widespread use: N-gram overlap metrics are widely used in MT, especially BLEU, making them a standard for benchmarking models and enabling easy comparison across studies.
 - Efficient computation: N-gram overlap metrics are computationally efficient and works well for quick assessments of translation quality.
 - Word n-gram matching: By focusing on n-gram overlaps, these metrics capture the degree of lexical similarity between the hypothesis and reference translations.
- Disadvantages:
 - Insensitive to semantics: N-gram overlap metrics rely solely on surface-level word matches, failing to account for semantic similarity or paraphrasing.
 - Context ignorance: N-gram overlap metrics do not account for context, which is crucial in capturing the nuances of MT.
 - Reliance on references: The quality of n-gram overlap metrics heavily depends on the availability of high-quality reference translations, limiting its reliability in low-resource scenarios.
 - Bias towards short phrases: N-gram overlap metrics may over-penalize longer, valid translations due to brevity penalties or under-represented n-grams.

Embeddings-based metrics:

- Advantages:
 - Semantic sensitivity: By leveraging contextual embeddings from models like BERT (Devlin et al., 2019) for BERTScore, embeddings-based metrics

capture semantic similarity and accounts for paraphrasing better than n-gram overlap metrics.

- Robust to variations: Embeddings-based metrics are more robust to word order and phrasing differences, making it suitable for languages with flexible syntactic structures.
 - Handles low-resource scenarios: Embeddings-based metrics perform well even with a limited number of reference translations by emphasizing meaning over exact matches.
- Disadvantages:
 - Higher computational cost: Calculating embeddings-based metrics requires the use of pre-trained transformer models, making it more resource-intensive.
 - Dependency on pre-trained models: The quality of embeddings-based metrics depends on the pre-trained embeddings, which might not always align well with the target language or domain.
 - Less established: Embeddings-based metrics are relatively newer and less standardized, which may hinder direct comparisons across different studies.
 - Overemphasis on semantic similarity: While beneficial, embeddings-based metrics may overlook syntactic errors or stylistic mismatches that are critical in ST.

Both n-gram overlap metrics and embeddings-based metrics have their merits and limitations, and their effectiveness often depends on the specific requirements of the ST task. Combining them or using them in tandem with human evaluation can provide a more comprehensive assessment.

BLEU (Bilingual Evaluation Understudy):

The BLEU score measures the similarity between a candidate translation Can and one or more reference translations Ref by calculating the precision of n-gram matches, penalized for brevity.

Modified n-gram precision: The modified precision $Prec_n$ is calculated for n-grams of size n as below

$$Prec_n = \frac{\sum_{g \in Can} \min(\text{count}(g, Can), \text{count}(g, Ref))}{\sum_{g \in Can} \text{count}(g, Can)} \quad (29)$$

where:

- g represents an n-gram
- $\text{count}(g, \text{Can})$ is the frequency of g in the candidate
- $\text{count}(g, \text{Ref})$ is the frequency of g in the reference

Brevity Penalty (BP): A brevity penalty accounts for the candidate translation being shorter than the reference:

$$BP = \begin{cases} 1, & \text{if } \text{len}(\text{Can}) > \text{len}(\text{Ref}), \\ e^{1 - \frac{\text{len}(\text{Ref})}{\text{len}(\text{Can})}}, & \text{otherwise.} \end{cases} \quad (30)$$

BLEU Score: The BLEU score is computed as the geometric mean of the n-gram precisions, weighted by a constant w_n :

$$\text{BLEU} = BP \cdot \exp \left(\sum_{n=1}^N w_n \log \text{Prec}_n \right) \quad (31)$$

Typically, $N = 4$ (up to 4-grams), and $w_n = \frac{1}{N}$.

BERTScore: BERTScore evaluates the semantic similarity between the candidate Can and reference Ref by computing cosine similarities of their token embeddings.

Token Embeddings: Let $\mathbf{E}(\text{Can}) = \{\mathbf{can}_i\}$ and $\mathbf{E}(\text{Ref}) = \{\mathbf{ref}_j\}$ be the token embeddings of Can and Ref, obtained from a pre-trained model like BERT.

Cosine Similarity Matrix: Compute the cosine similarity between all pairs of token embeddings as below

$$\text{Sim}_{ij} = \frac{\mathbf{can}_i \cdot \mathbf{ref}_j}{\|\mathbf{can}_i\| \|\mathbf{ref}_j\|} \quad (32)$$

Precision, Recall, and F1 Score:

- Precision:

$$\text{Prec} = \frac{1}{|\text{Can}|} \sum_{i=1}^{|\text{Can}|} \max_j \text{Sim}_{ij} \quad (33)$$

- Recall:

$$\text{Rec} = \frac{1}{|\text{Ref}|} \sum_{j=1}^{|\text{Ref}|} \max_i \text{Sim}_{ij} \quad (34)$$

- F1 Score (BERTScore):

$$\text{BERTScore} = 2 \cdot \frac{\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}} \quad (35)$$

In practice, BERTScore can be averaged across a dataset to produce a final evaluation score.

WER: Both WER and CER are common metrics for evaluating the quality of ASR for ST. They compare the output sequence (hypothesis) with a reference sequence and compute the number of errors in terms of word or character differences. WER measures the ratio of the total number of word-level errors (insertions, deletions, and substitutions) to the total number of words in the reference.

Definition:

$$\text{WER} = \frac{\text{Sub} + \text{Del} + \text{Ins}}{\text{Num}} \quad (36)$$

where:

- Sub: Number of word substitutions
 - Del: Number of word deletions
 - Ins: Number of word insertions
 - Num: Total number of words in the reference
- Steps:

1. Align the hypothesis and reference sequences using dynamic programming (e.g., Levenshtein distance)
2. Count the number of substitutions, deletions, and insertions

CER: CER operates similarly to WER but at the character level, making it suitable for scripts where words are not clearly delineated, such as Chinese or languages with agglutination.

Definition:

$$\text{CER} = \frac{\text{SUB} + \text{DEL} + \text{INS}}{\text{NUM}} \quad (37)$$

Where:

- SUB: Number of character substitutions.
- DEL: Number of character deletions.
- INS: Number of character insertions.
- NUM: Total number of characters in the reference.

Steps:

1. Align the hypothesis and reference sequences character by character
2. Count substitutions, deletions, and insertions

These formulations highlight the alignment-based approach to calculating WER and CER, which can be implemented using algorithms like dynamic programming to find the minimal edit distance between the hypothesis and the reference.

E.2 Details of Human Evaluation

Various approaches exist for eliciting judgments from informants regarding the quality of machine-translated sentences. Human evaluators may be tasked with directly assessing MT outputs by assigning scores to specific indicators on a predefined scale (0 to 10) for the same source sentence. These evaluations are typically based on three key criteria: adequacy, fluency, and comprehensibility.

- **Adequacy:** Measures how well the meaning of the source text is conveyed in the translation
- **Fluency:** Evaluates the grammatical and stylistic quality of the translated text, irrespective of the source text
- **Comprehensibility:** Assesses how easily a human reader can understand the translated text without referring to the source.

Direct assessments of MT ranking serve as the standard evaluation methods in recent biomedical MT shared task campaigns conducted by WMT (Bojar et al., 2016, 2017), as well as in MT research from the 1990s led by the Advanced Research Projects Agency (ARPA) (Church and Hovy, 1993; White et al., 1994).

E.3 Details of LLM-as-a-judge

The concept of using LLMs as a "judge" in MT has emerged as a promising method for evaluating translation quality. Unlike traditional evaluation metrics such as BLEU, ROUGE, or METEOR, which rely on n-gram overlap between machine-generated translations and reference texts, LLM-based evaluators leverage their advanced understanding of language semantics, context, and grammar. This approach allows for a more nuanced assessment of translation fidelity, fluency, and adequacy.

Advantages:

- **Contextual understanding:** LLMs excel at evaluating translations by considering the broader context and subtle nuances in language use, which traditional metrics often overlook (Karpinska and Iyyer, 2023; Fernandes et al., 2023).
- **Reference-free evaluation:** LLMs can perform evaluations without requiring reference translations, which can reduce biases introduced by specific linguistic choices in reference texts (Chen et al., 2024a; Stureborg et al., 2024).
- **Scalability and automation:** LLMs enable scalable, automated evaluation pipelines, reducing the dependency on human annotators for large-scale MT tasks (Yang et al., 2023; He et al., 2024).

Disadvantages:

- **Biases in LLMs:** LLMs may inherit biases from their training data, potentially influencing their judgment of translation quality (Behnke et al., 2022; Huang et al., 2023).
- **Generalization:** Ensuring that LLMs perform reliably across languages, domains, and translation styles remains a significant challenge (Yan et al., 2024; Singh et al., 2024).

Figure 32 shows the LLM-as-a-judge prompt template we used for ST transcript evaluation.



evaluation_prompt = f"""

Please act as an impartial translation quality assessment expert. Evaluate the Assistant's Translation by analyzing these aspects:

1. ***Correctness*** (Faithfulness to source):
 - Compare with Source Text ({source_lang})
 - Check for omissions/additions/distortions
 - Preservation of semantic meaning
2. ***Fluency*** (Target language quality):
 - Naturalness in {target_lang}
 - Grammatical correctness
 - Idiomatic expression
3. ***Terminology*** (Domain consistency):
 - Specialized term consistency
 - Comparison with Reference Translation
 - Proper noun handling

Source Text ({source_lang}) {row['source sentence']}

Reference Translation ({target_lang}) {row['reference sentence']}

Assistant's Translation ({target_lang}) {row['prediction sentence']}

Evaluation Steps

1. Analyze errors in each category
2. Classify error severity (minor/major/critical)
3. Provide specific examples of errors
4. Assign numerical score (1-10 scale):
 - 1-4: Poor (meaning distorted)
 - 5-6: Fair (meaning preserved but with issues)
 - 7-8: Good (minor errors)
 - 9-10: Excellent (near-perfect)

Format your response as:

***Errors:**
- [Category] Description (severity)

Example: "Original: X | Translation: Y"

***Rating:** {[N]} (Score between 1-10)

Provide only the formatted response starting with ***Errors:**

"""

Response:



Figure 32: LLM-as-a-judge prompt template we used for ST transcript evaluation.

F Extra Experimental Results

F.1 In-context Learning Results

This section presents in-context learning results, comparing full fine-tuning, zero-shot, few-shot on both ground-truth transcript and ASR transcript in the cascaded setting.

Full results are shown in Table 14 (English to X), Table 15 (Vietnamese to X), Table 16 (French to X), Table 17 (German to X), and Table 18 (Chinese to X) below.

For LLMs, we conducted few-shot learning experiments to assess their strength in MT tasks. On ground-truth transcripts, fine-tuned models significantly outperformed LLMs on most language pairs for both Qwen-2.5-7B and Mistral-v0.3-7B. Notably, the Llama-3.1-8B model only showed better results than few-shot versions for source languages such as English, French, and German. Furthermore, as the number of few-shot examples increased, performance improved across all three LLMs, as shown in the tables. For ASR transcripts, a similar trend was observed, although there was a notable exception: when the source language was English, few-shot models maintained BLEU scores comparable to those of ground-truth text, despite a slight drop in BERTScores, which were still higher than those of fine-tuned LLMs.

Model	Metrics	Ground-truth				ASR			
		en-vi	en-fr	en-zh	en-de	en-vi	en-fr	en-zh	en-de
Llama-3.1-8B - Ft.	BLEU	53.44	48.24	37.5	40.49	43.32	37.92	30.78	31.36
	BERTScore	0.9	0.89	0.83	0.87	0.78	0.76	0.73	0.74
	TER	39.42	46.9	58.37	53.12	53.71	60.93	66.54	67.05
	METEOR	0.77	0.72	0.6	0.66	0.63	0.59	0.51	0.53
	ChrF	67.74	70.97	32.39	66.08	57.23	60.63	26.34	56.58
	ROUGE-1	0.83	0.73	0.15	0.67	0.76	0.63	0.13	0.57
	ROUGE-2	0.68	0.57	0.13	0.47	0.58	0.47	0.11	0.37
	ROUGE-L	0.76	0.7	0.15	0.63	0.67	0.59	0.13	0.53
Llama-3.1-8B - 0 Shot	BLEU	42.24	41.25	24.22	33.89	14.59	11.01	10.01	8.84
	BERTScore	0.94	0.92	0.88	0.91	0.75	0.75	0.63	0.74
	TER	47.12	48.21	90.92	55.95	95.01	110.53	119.62	119.14
	METEOR	0.66	0.65	0.51	0.59	0.32	0.32	0.25	0.28
	ChrF	57.44	63.84	25.17	58.71	30.2	38.57	10.25	35.75
	ROUGE-1	0.77	0.68	0.17	0.61	0.5	0.34	0.09	0.29
	ROUGE-2	0.59	0.51	0.15	0.4	0.28	0.18	0.07	0.13
	ROUGE-L	0.69	0.66	0.16	0.58	0.39	0.31	0.09	0.26
Llama-3.1-8B - 1 Shot	BLEU	47.77	46.22	31.69	38.34	14.57	11.33	13.75	8.77
	BERTScore	0.95	0.93	0.88	0.92	0.77	0.77	0.69	0.74
	TER	40.66	42.69	59.98	50.31	101.77	115.8	83.28	129.49
	METEOR	0.72	0.7	0.56	0.64	0.36	0.34	0.31	0.3
	ChrF	62.4	68.12	27.1	63.19	33.13	40.18	12.97	37.59
	ROUGE-1	0.81	0.73	0.17	0.66	0.54	0.34	0.11	0.28
	ROUGE-2	0.64	0.56	0.16	0.44	0.3	0.19	0.09	0.13
	ROUGE-L	0.74	0.7	0.17	0.63	0.41	0.31	0.1	0.25
Llama-3.1-8B - 8 Shot	BLEU	46.7	45.58	31.28	38.0	16.14	12.38	13.97	9.38
	BERTScore	0.95	0.93	0.88	0.92	0.77	0.77	0.69	0.75
	TER	41.73	43.32	59.24	50.92	95.97	109.27	80.78	123.43
	METEOR	0.7	0.7	0.56	0.64	0.37	0.34	0.32	0.3
	ChrF	61.5	67.39	26.85	62.46	34.01	40.42	13.29	37.53
	ROUGE-1	0.81	0.73	0.17	0.66	0.55	0.35	0.11	0.29
	ROUGE-2	0.63	0.56	0.15	0.44	0.31	0.19	0.09	0.13
	ROUGE-L	0.73	0.7	0.17	0.62	0.43	0.32	0.1	0.26
	BLEU	47.77	46.22	43.62	38.34	15.45	12.04	13.94	9.17
	BERTScore	0.95	0.93	0.94	0.92	0.77	0.75	0.69	0.74

Table 13 continued from previous page

Model	Metrics	Ground-truth				ASR			
		en-vi	en-fr	en-zh	en-de	en-vi	en-fr	en-zh	en-de
Llama-3.1-8B - 16 Shot	TER	40.66	42.69	46.76	50.31	101.83	109.92	81.05	124.29
	METEOR	0.72	0.7	0.68	0.64	0.37	0.33	0.32	0.29
	ChrF	62.4	68.12	37.57	63.19	34.13	39.64	13.27	36.84
	ROUGE-1	0.81	0.73	0.2	0.66	0.54	0.34	0.11	0.28
	ROUGE-2	0.64	0.56	0.18	0.44	0.31	0.19	0.09	0.13
	ROUGE-L	0.74	0.7	0.19	0.63	0.42	0.31	0.1	0.25
Llama-3.1-8B - 32 Shot	BLEU	48.27	46.92	30.56	38.27	15.21	15.29	13.97	12.82
	BERTScore	0.95	0.93	0.87	0.92	0.77	0.76	0.69	0.74
	TER	41.04	42.64	59.25	50.85	102.67	78.6	81.42	84.12
	METEOR	0.72	0.7	0.55	0.64	0.37	0.33	0.31	0.3
	ChrF	62.63	68.33	26.38	62.97	33.87	38.04	13.28	36.11
	ROUGE-1	0.81	0.73	0.17	0.65	0.53	0.39	0.1	0.35
	ROUGE-2	0.64	0.56	0.16	0.44	0.3	0.22	0.08	0.16
	ROUGE-L	0.74	0.7	0.17	0.62	0.41	0.35	0.1	0.31
Qwen-2.5-7B - Ft.	BLEU	54.5	49.63	28.61	38.75	43.37	37.34	23.46	28.5
	BERTScore	0.9	0.9	0.81	0.87	0.77	0.76	0.74	0.74
	TER	38.21	42.42	59.1	51.55	53.52	57.76	64.75	66.7
	METEOR	0.77	0.72	0.5	0.64	0.63	0.58	0.44	0.51
	ChrF	67.82	70.5	27.81	63.39	57.34	60.21	23.19	54.19
	ROUGE-1	0.83	0.74	0.14	0.65	0.76	0.63	0.12	0.55
	ROUGE-2	0.69	0.57	0.13	0.43	0.59	0.47	0.1	0.34
	ROUGE-L	0.77	0.71	0.14	0.62	0.67	0.6	0.12	0.51
Qwen-2.5-7B - 0 Shot	BLEU	38.32	39.1	36.16	30.0	13.95	13.43	17.4	10.05
	BERTScore	0.93	0.91	0.92	0.9	0.73	0.73	0.7	0.72
	TER	53.71	52.77	63.44	62.2	82.95	84.55	95.85	89.8
	METEOR	0.61	0.61	0.66	0.55	0.3	0.3	0.36	0.26
	ChrF	53.19	60.89	34.98	55.39	28.72	35.86	16.47	33.06
	ROUGE-1	0.72	0.64	0.17	0.56	0.51	0.36	0.1	0.31
	ROUGE-2	0.54	0.47	0.15	0.34	0.28	0.19	0.08	0.13
	ROUGE-L	0.64	0.61	0.17	0.53	0.4	0.32	0.1	0.27
Qwen-2.5-7B - 1 Shot	BLEU	43.81	44.22	26.17	28.21	16.09	16.18	15.76	11.56
	BERTScore	0.93	0.93	0.87	0.91	0.75	0.75	0.68	0.68
	TER	45.94	44.7	62.27	62.57	79.98	81.36	85.38	84.33
	METEOR	0.67	0.68	0.48	0.5	0.34	0.35	0.33	0.3
	ChrF	58.22	66.19	25.15	51.07	31.66	39.96	14.43	35.66
	ROUGE-1	0.78	0.71	0.15	0.52	0.56	0.41	0.1	0.35
	ROUGE-2	0.6	0.53	0.13	0.32	0.31	0.22	0.08	0.15
	ROUGE-L	0.7	0.68	0.15	0.49	0.43	0.37	0.1	0.31
Qwen-2.5-7B - 8 Shot	BLEU	42.62	42.14	26.13	31.26	15.15	13.73	9.85	10.12
	BERTScore	0.93	0.92	0.87	0.9	0.73	0.72	0.64	0.7
	TER	50.59	48.99	62.56	62.16	84.56	82.99	86.47	92.32
	METEOR	0.65	0.65	0.48	0.56	0.31	0.3	0.23	0.26
	ChrF	56.84	63.67	24.91	55.45	29.24	34.64	10.07	31.7
	ROUGE-1	0.76	0.68	0.15	0.57	0.5	0.36	0.08	0.29
	ROUGE-2	0.58	0.5	0.14	0.36	0.28	0.19	0.06	0.13
	ROUGE-L	0.68	0.65	0.15	0.53	0.39	0.32	0.08	0.26
	BLEU	43.81	44.22	26.17	28.21	16.03	14.52	12.21	7.8
	BERTScore	0.94	0.93	0.87	0.89	0.75	0.74	0.68	0.68

Table 13 continued from previous page

Model	Metrics	Ground-truth				ASR			
		en-vi	en-fr	en-zh	en-de	en-vi	en-fr	en-zh	en-de
Qwen-2.5-7B - 16 Shot	TER	45.94	44.7	62.27	62.57	82.91	84.47	86.09	96.92
	METEOR	0.67	0.68	0.48	0.5	0.33	0.32	0.27	0.21
	ChrF	58.22	66.19	25.15	51.07	31.11	36.74	11.81	27.56
	ROUGE-1	0.78	0.71	0.15	0.52	0.54	0.38	0.09	0.24
	ROUGE-2	0.6	0.53	0.13	0.32	0.3	0.2	0.07	0.09
	ROUGE-L	0.7	0.68	0.15	0.49	0.42	0.34	0.09	0.21
Qwen-2.5-7B - 32 Shot	BLEU	43.51	44.95	25.72	33.24	16.96	16.86	14.63	9.33
	BERTScore	0.94	0.93	0.87	0.91	0.75	0.76	0.7	0.71
	TER	45.74	44.76	62.47	56.02	82.13	80.46	83.36	97.53
	METEOR	0.66	0.69	0.47	0.59	0.35	0.36	0.31	0.24
	ChrF	58.02	66.94	24.8	58.63	32.33	40.69	13.9	31.57
	ROUGE-1	0.77	0.71	0.15	0.61	0.56	0.42	0.1	0.27
	ROUGE-2	0.6	0.54	0.13	0.38	0.31	0.23	0.08	0.11
	ROUGE-L	0.7	0.68	0.15	0.57	0.44	0.38	0.1	0.24
Mistral-v0.3-7B - Ft.	BLEU	24.77	51.71	26.38	43.99	17.72	36.58	20.27	29.9
	BERTScore	0.82	0.89	0.81	0.88	0.68	0.74	0.69	0.72
	TER	63.33	43.0	60.43	48.54	71.27	59.78	70.94	65.81
	METEOR	0.45	0.73	0.48	0.67	0.35	0.55	0.38	0.5
	ChrF	41.27	71.06	27.29	66.27	33.42	57.32	20.29	52.89
	ROUGE-1	0.66	0.74	0.12	0.68	0.57	0.6	0.09	0.54
	ROUGE-2	0.49	0.6	0.11	0.49	0.39	0.45	0.08	0.36
	ROUGE-L	0.57	0.72	0.12	0.65	0.49	0.58	0.09	0.51
Mistral-v0.3-7B - 0 Shot	BLEU	8.99	37.86	34.21	29.25	1.59	10.56	12.83	6.76
	BERTScore	0.89	0.92	0.92	0.91	0.62	0.73	0.66	0.68
	TER	80.02	51.46	57.09	59.85	94.82	85.99	93.26	95.29
	METEOR	0.25	0.62	0.58	0.55	0.08	0.28	0.28	0.21
	ChrF	24.94	61.82	29.28	56.03	11.44	34.01	11.71	28.79
	ROUGE-1	0.53	0.66	0.18	0.58	0.31	0.35	0.1	0.25
	ROUGE-2	0.26	0.47	0.16	0.35	0.09	0.17	0.08	0.09
	ROUGE-L	0.4	0.63	0.18	0.54	0.23	0.31	0.09	0.22
Mistral-v0.3-7B - 1 Shot	BLEU	12.38	41.24	29.67	35.08	2.37	11.39	13.14	7.52
	BERTScore	0.91	0.93	0.89	0.92	0.64	0.77	0.69	0.75
	TER	75.66	47.52	59.61	52.67	92.12	84.13	81.89	94.24
	METEOR	0.31	0.66	0.52	0.61	0.1	0.29	0.3	0.22
	ChrF	29.17	64.35	26.76	60.68	12.84	34.29	12.62	29.71
	ROUGE-1	0.58	0.69	0.15	0.64	0.35	0.36	0.1	0.26
	ROUGE-2	0.32	0.51	0.14	0.41	0.12	0.18	0.08	0.1
	ROUGE-L	0.45	0.66	0.15	0.6	0.27	0.32	0.1	0.24
Mistral-v0.3-7B - 8 Shot	BLEU	13.64	41.86	21.4	34.17	4.76	15.02	8.54	11.69
	BERTScore	0.91	0.93	0.87	0.92	0.7	0.76	0.67	0.74
	TER	74.2	46.61	64.85	53.98	86.73	76.58	80.51	83.67
	METEOR	0.33	0.66	0.44	0.6	0.17	0.34	0.25	0.3
	ChrF	30.12	64.65	21.72	59.53	18.25	38.97	10.1	35.59
	ROUGE-1	0.59	0.7	0.13	0.63	0.45	0.42	0.08	0.35
	ROUGE-2	0.34	0.51	0.12	0.4	0.19	0.22	0.06	0.15
	ROUGE-L	0.46	0.67	0.13	0.59	0.33	0.38	0.07	0.31
	BLEU	14.22	42.83	21.35	35.08	4.9	15.64	9.83	12.0
	BERTScore	0.91	0.93	0.87	0.92	0.71	0.77	0.69	0.75

Table 13 continued from previous page

Model	Metrics	Ground-truth				ASR			
		en-vi	en-fr	en-zh	en-de	en-vi	en-fr	en-zh	en-de
Mistral-v0.3-7B - 16 Shot	TER	73.33	45.48	64.69	52.67	86.31	75.73	78.37	82.11
	METEOR	0.34	0.67	0.44	0.61	0.18	0.35	0.27	0.3
	ChrF	30.98	65.76	21.77	60.68	18.45	39.6	11.42	36.29
	ROUGE-1	0.6	0.71	0.13	0.64	0.46	0.43	0.09	0.36
	ROUGE-2	0.35	0.52	0.12	0.41	0.19	0.23	0.07	0.16
	ROUGE-L	0.47	0.68	0.13	0.6	0.33	0.38	0.09	0.32
Mistral-v0.3-7B - 32 Shot	BLEU	14.58	43.36	21.35	35.05	5.8	15.81	9.81	12.3
	BERTScore	0.91	0.93	0.87	0.92	0.73	0.77	0.69	0.75
	TER	73.14	45.49	64.51	53.02	84.79	75.53	78.33	82.1
	METEOR	0.34	0.68	0.44	0.61	0.2	0.35	0.27	0.31
	ChrF	31.29	66.06	21.77	60.7	20.24	39.73	11.41	36.63
	ROUGE-1	0.6	0.71	0.13	0.64	0.49	0.43	0.09	0.36
	ROUGE-2	0.36	0.53	0.11	0.41	0.22	0.23	0.07	0.16
	ROUGE-L	0.48	0.68	0.13	0.6	0.35	0.39	0.08	0.32

Table 14: **In-context Learning Results.** Full fine-tuning, zero-shot, few-shot on both ground-truth transcript and ASR transcript in the cascaded setting. **English to X** results are reported in this table. All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

Model	Metrics	Ground-truth				ASR			
		vi-en	vi-fr	vi-zh	vi-de	vi-en	vi-fr	vi-zh	vi-de
Llama-3.1-8B - Ft.	BLEU	23.16	15.57	16.09	11.61	14.55	10.29	11.56	7.71
	BERTScore	0.92	0.79	0.74	0.77	0.78	0.75	0.73	0.73
	TER	83.07	100.23	120.58	112.85	98.39	112.97	131.0	122.33
	METEOR	0.57	0.45	0.5	0.39	0.43	0.35	0.4	0.31
	ChrF	52.29	47.18	21.03	44.25	41.46	39.22	15.65	37.02
	ROUGE-1	0.56	0.45	0.03	0.4	0.44	0.36	0.02	0.32
	ROUGE-2	0.33	0.25	0.01	0.19	0.22	0.18	0.01	0.12
	ROUGE-L	0.51	0.4	0.03	0.36	0.39	0.31	0.02	0.28
Llama-3.1-8B - 0 Shot	BLEU	17.01	11.95	4.23	9.93	11.3	08.06	3.11	6.75
	BERTScore	0.79	0.76	0.61	0.76	0.88	0.73	0.54	0.72
	TER	81.7	89.32	180.31	92.37	95.71	98.96	182.42	102.77
	METEOR	0.42	0.33	0.19	0.3	0.34	0.26	0.15	0.24
	ChrF	43.0	38.83	6.21	37.28	35.53	33.2	4.79	32.0
	ROUGE-1	0.47	0.39	0.02	0.35	0.38	0.32	0.01	0.29
	ROUGE-2	0.24	0.19	0.01	0.14	0.17	0.14	0.0	0.1
	ROUGE-L	0.43	0.34	0.02	0.31	0.34	0.27	0.01	0.25
Llama-3.1-8B - 1 Shot	BLEU	24.75	16.02	18.84	11.93	14.08	9.94	12.24	7.62
	BERTScore	0.83	0.79	0.76	0.77	0.89	0.75	0.7	0.73
	TER	65.83	80.97	75.82	85.74	86.83	95.22	89.86	97.14
	METEOR	0.53	0.38	0.43	0.32	0.37	0.3	0.34	0.25
	ChrF	50.29	43.02	17.01	39.7	39.04	36.81	12.43	33.4
	ROUGE-1	0.56	0.43	0.03	0.38	0.42	0.36	0.02	0.3
	ROUGE-2	0.32	0.23	0.01	0.16	0.2	0.16	0.01	0.1
	ROUGE-L	0.51	0.38	0.03	0.33	0.38	0.31	0.02	0.26
Llama-3.1-8B - 8 Shot	BLEU	24.24	15.42	15.03	11.25	15.8	10.29	11.89	8.84
	BERTScore	0.82	0.78	0.77	0.77	0.9	0.75	0.7	0.74
	TER	66.93	84.03	89.52	86.95	84.82	94.29	93.2	96.51
	METEOR	0.52	0.37	0.41	0.31	0.41	0.29	0.34	0.27
	ChrF	49.83	42.08	15.47	39.18	40.61	35.1	12.52	35.19
	ROUGE-1	0.55	0.41	0.03	0.37	0.44	0.33	0.02	0.32
	ROUGE-2	0.31	0.22	0.01	0.15	0.21	0.15	0.01	0.12
	ROUGE-L	0.51	0.36	0.03	0.32	0.39	0.29	0.02	0.28
Llama-3.1-8B - 16 Shot	BLEU	24.75	16.02	18.84	11.93	16.94	10.28	11.24	8.4
	BERTScore	0.83	0.79	0.78	0.77	0.9	0.75	0.69	0.74
	TER	65.83	80.97	75.82	85.74	81.87	98.72	101.17	98.04
	METEOR	0.53	0.38	0.43	0.32	0.41	0.31	0.34	0.28
	ChrF	50.29	43.02	17.01	39.7	40.78	37.23	12.52	35.24
	ROUGE-1	0.56	0.43	0.03	0.38	0.44	0.35	0.02	0.31
	ROUGE-2	0.32	0.23	0.01	0.16	0.22	0.16	0.01	0.12
	ROUGE-L	0.51	0.38	0.03	0.33	0.4	0.3	0.02	0.27
Llama-3.1-8B - 32 Shot	BLEU	24.53	15.58	17.45	12.44	15.11	10.21	12.28	7.76
	BERTScore	0.82	0.79	0.78	0.77	0.9	0.75	0.7	0.74
	TER	68.49	84.25	78.04	86.52	85.42	97.14	88.95	99.26
	METEOR	0.52	0.39	0.42	0.34	0.39	0.31	0.33	0.27
	ChrF	49.62	43.73	16.15	40.81	39.4	36.88	12.3	34.24
	ROUGE-1	0.54	0.43	0.03	0.39	0.42	0.34	0.02	0.31
	ROUGE-2	0.31	0.23	0.01	0.17	0.2	0.16	0.01	0.11
	ROUGE-L	0.5	0.38	0.03	0.34	0.37	0.3	0.02	0.26
	BLEU	26.21	19.25	29.06	14.44	13.97	11.66	20.27	8.75

Table 14 continued from previous page

Model	Metrics	Ground-truth				ASR			
		vi-en	vi-fr	vi-zh	vi-de	vi-en	vi-fr	vi-zh	vi-de
Qwen-2.5-7B - Ft.	BERTScore	0.93	0.81	0.81	0.79	0.78	0.76	0.78	0.75
	TER	71.1	80.86	60.93	83.75	101.68	101.51	74.57	101.85
	METEOR	0.57	0.47	0.55	0.39	0.43	0.36	0.43	0.3
	ChrF	53.79	48.94	25.1	43.34	41.76	40.73	18.0	36.74
	ROUGE-1	0.58	0.49	0.04	0.42	0.43	0.39	0.03	0.34
	ROUGE-2	0.35	0.28	0.02	0.19	0.21	0.19	0.02	0.13
	ROUGE-L	0.53	0.44	0.04	0.38	0.39	0.33	0.03	0.29
Qwen-2.5-7B - 0 Shot	BLEU	20.89	15.21	26.23	9.73	12.41	9.94	16.57	6.59
	BERTScore	0.79	0.78	0.81	0.75	0.88	0.74	0.73	0.72
	TER	77.98	85.59	73.88	92.37	93.79	97.19	91.72	100.89
	METEOR	0.45	0.38	0.54	0.29	0.34	0.3	0.41	0.24
	ChrF	44.94	42.86	24.17	36.87	35.59	36.05	16.52	32.04
	ROUGE-1	0.47	0.42	0.04	0.33	0.36	0.33	0.02	0.28
	ROUGE-2	0.25	0.22	0.02	0.13	0.16	0.15	0.01	0.09
	ROUGE-L	0.43	0.37	0.04	0.29	0.32	0.29	0.02	0.23
Qwen-2.5-7B - 1 Shot	BLEU	22.31	18.44	27.25	11.33	14.34	10.5	16.78	6.42
	BERTScore	0.82	0.8	0.81	0.77	0.9	0.76	0.73	0.73
	TER	84.56	79.35	62.93	92.8	87.87	96.22	80.56	104.4
	METEOR	0.55	0.44	0.53	0.35	0.4	0.33	0.41	0.26
	ChrF	52.13	48.14	23.53	41.25	40.29	38.73	15.85	34.64
	ROUGE-1	0.54	0.48	0.03	0.38	0.42	0.36	0.02	0.29
	ROUGE-2	0.32	0.27	0.02	0.16	0.2	0.17	0.01	0.1
	ROUGE-L	0.5	0.43	0.03	0.34	0.37	0.31	0.02	0.25
Qwen-2.5-7B - 8 Shot	BLEU	20.36	15.47	22.98	11.41	15.24	11.77	18.31	7.7
	BERTScore	0.81	0.79	0.8	0.77	0.9	0.76	0.74	0.74
	TER	90.02	95.22	80.69	88.14	90.58	93.75	81.27	99.6
	METEOR	0.54	0.43	0.53	0.34	0.42	0.34	0.42	0.28
	ChrF	51.07	47.05	22.71	40.77	42.27	40.16	17.26	35.89
	ROUGE-1	0.52	0.45	0.03	0.38	0.44	0.37	0.02	0.31
	ROUGE-2	0.3	0.25	0.02	0.16	0.21	0.18	0.01	0.11
	ROUGE-L	0.48	0.4	0.03	0.34	0.39	0.33	0.02	0.27
Qwen-2.5-7B - 16 Shot	BLEU	22.31	18.44	27.25	11.33	14.14	11.85	19.11	7.89
	BERTScore	0.82	0.81	0.82	0.77	0.9	0.77	0.75	0.74
	TER	84.56	79.35	62.93	92.8	100.74	96.37	78.32	99.54
	METEOR	0.55	0.44	0.53	0.35	0.43	0.35	0.43	0.28
	ChrF	52.13	48.14	23.53	41.25	42.39	40.74	17.59	35.87
	ROUGE-1	0.54	0.48	0.03	0.38	0.43	0.38	0.02	0.31
	ROUGE-2	0.32	0.27	0.02	0.16	0.21	0.19	0.01	0.11
	ROUGE-L	0.5	0.43	0.03	0.34	0.38	0.33	0.02	0.27
Qwen-2.5-7B - 32 Shot	BLEU	24.52	18.13	26.61	11.97	1.91	3.72	9.15	2.24
	BERTScore	0.83	0.81	0.82	0.78	0.84	0.67	0.65	0.63
	TER	73.31	81.12	65.47	88.73	171.22	116.81	112.45	111.3
	METEOR	0.56	0.45	0.53	0.35	0.16	0.16	0.25	0.11
	ChrF	52.88	48.54	22.98	41.82	18.97	22.33	9.67	15.9
	ROUGE-1	0.56	0.48	0.04	0.39	0.15	0.18	0.03	0.12
	ROUGE-2	0.33	0.27	0.02	0.17	0.04	0.06	0.01	0.03
	ROUGE-L	0.52	0.43	0.04	0.35	0.13	0.16	0.03	0.11
	BLEU	24.56	16.0	25.04	13.38	15.86	10.92	17.92	09.03

Table 14 continued from previous page

Model	Metrics	Ground-truth				ASR			
		vi-en	vi-fr	vi-zh	vi-de	vi-en	vi-fr	vi-zh	vi-de
Mistral-v0.3-7B - Ft.	BERTScore	0.92	0.79	0.78	0.78	0.78	0.75	0.77	0.75
	TER	70.37	87.87	65.43	84.13	85.32	97.69	76.97	94.56
	METEOR	0.53	0.42	0.49	0.37	0.42	0.33	0.39	0.3
	ChrF	48.73	43.93	21.84	40.96	39.64	37.19	16.1	34.82
	ROUGE-1	0.54	0.44	0.03	0.4	0.43	0.36	0.02	0.33
	ROUGE-2	0.31	0.24	0.01	0.18	0.21	0.17	0.01	0.12
	ROUGE-L	0.5	0.39	0.03	0.36	0.39	0.32	0.02	0.29
Mistral-v0.3-7B - 0 Shot	BLEU	17.98	10.65	18.42	8.01	4.84	1.52	1.27	1.15
	BERTScore	0.8	0.76	0.78	0.75	0.85	0.61	0.46	0.6
	TER	77.89	92.55	73.8	94.34	108.45	137.43	136.79	126.4
	METEOR	0.44	0.33	0.42	0.29	0.19	0.1	0.04	0.08
	ChrF	42.7	39.02	16.81	35.76	22.36	17.13	1.45	16.1
	ROUGE-1	0.48	0.38	0.03	0.33	0.24	0.13	0.01	0.1
	ROUGE-2	0.24	0.18	0.02	0.12	0.08	0.03	0.0	0.02
	ROUGE-L	0.43	0.34	0.03	0.29	0.21	0.11	0.01	0.09
Mistral-v0.3-7B - 1 Shot	BLEU	21.04	11.98	17.79	9.32	4.19	1.22	2.22	1.2
	BERTScore	0.8	0.78	0.78	0.75	0.87	0.61	0.54	0.61
	TER	73.49	89.13	71.38	89.96	136.22	147.83	145.54	125.87
	METEOR	0.49	0.35	0.42	0.3	0.22	0.1	0.11	0.09
	ChrF	45.78	39.85	16.51	36.87	24.27	17.7	3.64	17.85
	ROUGE-1	0.51	0.39	0.03	0.35	0.24	0.12	0.01	0.12
	ROUGE-2	0.27	0.19	0.01	0.13	0.08	0.03	0.0	0.02
	ROUGE-L	0.46	0.35	0.03	0.31	0.21	0.11	0.01	0.1
Mistral-v0.3-7B - 8 Shot	BLEU	20.02	11.25	18.5	8.7	6.61	1.79	5.22	1.34
	BERTScore	0.81	0.77	0.79	0.76	0.86	0.62	0.6	0.6
	TER	74.58	93.42	69.95	92.04	106.49	148.88	100.6	151.77
	METEOR	0.47	0.35	0.43	0.29	0.24	0.11	0.17	0.1
	ChrF	44.81	39.49	17.04	36.62	25.31	17.76	5.89	16.9
	ROUGE-1	0.5	0.38	0.03	0.34	0.27	0.13	0.02	0.11
	ROUGE-2	0.26	0.18	0.01	0.12	0.09	0.04	0.01	0.02
	ROUGE-L	0.45	0.33	0.03	0.29	0.23	0.11	0.02	0.09
Mistral-v0.3-7B - 16 Shot	BLEU	21.04	12.42	18.7	9.32	7.45	2.91	6.78	02.07
	BERTScore	0.81	0.78	0.79	0.76	0.87	0.66	0.62	0.64
	TER	73.49	89.54	69.36	89.96	98.15	125.22	88.89	128.43
	METEOR	0.49	0.35	0.44	0.3	0.24	0.14	0.21	0.12
	ChrF	45.78	40.01	17.23	36.87	25.2	21.37	7.58	19.61
	ROUGE-1	0.51	0.39	0.03	0.35	0.27	0.17	0.02	0.14
	ROUGE-2	0.27	0.19	0.01	0.13	0.1	0.05	0.01	0.03
	ROUGE-L	0.46	0.35	0.03	0.31	0.24	0.15	0.02	0.13
Mistral-v0.3-7B - 32 Shot	BLEU	19.89	11.74	18.2	8.82	7.11	08.03	13.63	6.27
	BERTScore	0.81	0.77	0.79	0.76	0.87	0.75	0.72	0.73
	TER	76.59	92.94	69.24	92.13	99.96	103.54	77.74	100.09
	METEOR	0.48	0.35	0.43	0.29	0.24	0.29	0.35	0.24
	ChrF	45.47	39.8	16.86	36.31	25.33	34.67	13.22	32.19
	ROUGE-1	0.5	0.39	0.03	0.34	0.27	0.32	0.02	0.28
	ROUGE-2	0.26	0.18	0.01	0.12	0.1	0.13	0.01	0.09
	ROUGE-L	0.45	0.34	0.03	0.29	0.24	0.27	0.02	0.24

Table 14 continued from previous page

Model	Metrics	Ground-truth				ASR			
		vi-en	vi-fr	vi-zh	vi-de	vi-en	vi-fr	vi-zh	vi-de

Table 15: **In-context Learning Results.** Full fine-tuning, zero-shot, few-shot on both ground-truth transcript and ASR transcript in the cascaded setting. **Vietnamese to X** results are reported in this table.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

Model	Metrics	Ground-truth				ASR			
		fr-en	fr-vi	fr-zh	fr-de	fr-en	fr-vi	fr-zh	fr-de
Llama-3.1-8B - Ft.	BLEU	50.18	39.63	29.25	27.46	30.15	25.36	20.28	16.38
	BERTScore	0.95	0.86	0.79	0.81	0.82	0.8	0.75	0.74
	TER	42.23	54.71	68.66	82.2	65.8	71.69	80.06	99.6
	METEOR	0.76	0.65	0.52	0.56	0.52	0.47	0.4	0.4
	ChrF	69.44	55.16	25.3	56.35	49.71	41.24	17.84	44.43
	ROUGE-1	0.75	0.77	0.11	0.54	0.58	0.67	0.08	0.41
	ROUGE-2	0.58	0.56	0.07	0.34	0.39	0.44	0.05	0.22
	ROUGE-L	0.73	0.67	0.11	0.5	0.54	0.55	0.08	0.37
Llama-3.1-8B - 0 Shot	BLEU	38.9	26.76	6.82	26.87	22.72	15.96	07.05	15.73
	BERTScore	0.86	0.82	0.62	0.82	0.91	0.78	0.59	0.77
	TER	49.46	62.96	159.39	61.65	68.02	75.66	133.79	75.69
	METEOR	0.66	0.52	0.22	0.53	0.46	0.37	0.2	0.36
	ChrF	61.25	44.58	8.95	51.9	45.22	33.7	7.82	39.92
	ROUGE-1	0.69	0.69	0.07	0.57	0.54	0.61	0.07	0.44
	ROUGE-2	0.49	0.47	0.05	0.33	0.35	0.37	0.05	0.23
	ROUGE-L	0.66	0.59	0.07	0.52	0.5	0.48	0.06	0.4
Llama-3.1-8B - 1 Shot	BLEU	48.24	34.22	21.89	31.84	25.39	19.62	14.37	17.22
	BERTScore	0.89	0.85	0.79	0.84	0.92	0.79	0.73	0.78
	TER	41.44	55.64	68.08	57.83	65.29	71.64	76.66	73.8
	METEOR	0.73	0.6	0.46	0.57	0.48	0.42	0.34	0.38
	ChrF	67.71	51.17	19.56	55.77	47.3	37.8	14.16	41.72
	ROUGE-1	0.75	0.75	0.13	0.6	0.56	0.66	0.09	0.46
	ROUGE-2	0.56	0.53	0.1	0.37	0.36	0.41	0.07	0.24
	ROUGE-L	0.72	0.64	0.13	0.56	0.52	0.52	0.09	0.42
Llama-3.1-8B - 8 Shot	BLEU	48.25	34.08	20.97	32.13	28.69	20.92	14.97	19.07
	BERTScore	0.89	0.86	0.79	0.85	0.91	0.8	0.73	0.78
	TER	41.9	55.64	69.83	58.46	62.69	71.54	75.89	73.3
	METEOR	0.73	0.6	0.44	0.57	0.51	0.42	0.34	0.4
	ChrF	67.69	51.65	18.89	55.53	49.39	38.34	14.81	43.09
	ROUGE-1	0.74	0.76	0.12	0.6	0.58	0.65	0.08	0.47
	ROUGE-2	0.55	0.53	0.09	0.37	0.39	0.41	0.06	0.26
	ROUGE-L	0.72	0.65	0.12	0.56	0.54	0.52	0.08	0.43
Llama-3.1-8B - 16 Shot	BLEU	48.24	34.22	21.89	31.84	29.17	22.03	16.74	20.64
	BERTScore	0.89	0.85	0.8	0.84	0.92	0.79	0.73	0.79
	TER	41.44	55.64	68.08	57.83	62.62	70.32	73.95	73.17
	METEOR	0.73	0.6	0.46	0.57	0.51	0.43	0.36	0.4
	ChrF	67.71	51.17	19.56	55.77	49.83	38.95	16.22	43.47
	ROUGE-1	0.75	0.75	0.13	0.6	0.58	0.64	0.1	0.47
	ROUGE-2	0.56	0.53	0.1	0.37	0.4	0.41	0.07	0.27
	ROUGE-L	0.72	0.64	0.13	0.56	0.55	0.52	0.1	0.43
Llama-3.1-8B - 32 Shot	BLEU	49.7	34.83	22.38	31.31	28.66	21.09	14.24	18.48
	BERTScore	0.89	0.85	0.8	0.84	0.91	0.79	0.73	0.78
	TER	40.74	56.07	68.51	58.73	63.23	71.72	75.78	74.11
	METEOR	0.74	0.6	0.46	0.57	0.51	0.42	0.34	0.38
	ChrF	68.13	51.36	19.94	55.15	49.56	38.04	14.09	42.15
	ROUGE-1	0.75	0.75	0.13	0.6	0.58	0.64	0.09	0.46
	ROUGE-2	0.57	0.53	0.1	0.36	0.4	0.4	0.07	0.25
	ROUGE-L	0.72	0.64	0.13	0.55	0.55	0.51	0.09	0.42
	BLEU	49.69	40.67	20.97	33.91	30.35	25.59	15.33	20.38

Table 15 continued from previous page

Model	Metrics	Ground-truth				ASR			
		fr-en	fr-vi	fr-zh	fr-de	fr-en	fr-vi	fr-zh	fr-de
Qwen-2.5-7B - Ft.	BERTScore	0.95	0.86	0.78	0.84	0.81	0.8	0.76	0.78
	TER	42.17	52.57	67.39	59.28	69.6	71.42	73.88	76.51
	METEOR	0.76	0.66	0.43	0.58	0.52	0.47	0.36	0.4
	ChrF	70.03	56.12	20.86	56.82	49.88	41.85	15.7	43.63
	ROUGE-1	0.76	0.78	0.09	0.6	0.57	0.67	0.07	0.47
	ROUGE-2	0.58	0.58	0.06	0.37	0.39	0.45	0.05	0.26
	ROUGE-L	0.73	0.68	0.08	0.56	0.54	0.55	0.07	0.43
Qwen-2.5-7B - 0 Shot	BLEU	39.14	27.77	26.36	24.27	22.81	16.46	21.05	13.9
	BERTScore	0.83	0.81	0.8	0.8	0.89	0.75	0.73	0.74
	TER	59.08	65.38	80.39	69.54	76.68	80.86	84.82	81.83
	METEOR	0.61	0.5	0.56	0.47	0.43	0.34	0.4	0.32
	ChrF	57.98	44.54	26.81	48.68	43.3	32.15	18.48	36.85
	ROUGE-1	0.61	0.67	0.13	0.5	0.48	0.56	0.09	0.38
	ROUGE-2	0.43	0.45	0.1	0.27	0.29	0.33	0.07	0.18
	ROUGE-L	0.58	0.56	0.13	0.46	0.44	0.44	0.09	0.34
Qwen-2.5-7B - 1 Shot	BLEU	44.9	31.33	18.7	26.55	27.03	16.24	20.12	16.14
	BERTScore	0.89	0.83	0.79	0.81	0.91	0.78	0.74	0.77
	TER	51.04	60.82	68.76	64.25	65.99	74.57	74.12	76.47
	METEOR	0.72	0.55	0.42	0.5	0.49	0.37	0.4	0.36
	ChrF	66.38	47.26	18.54	49.7	48.73	33.57	18.39	40.39
	ROUGE-1	0.71	0.7	0.09	0.53	0.57	0.59	0.1	0.44
	ROUGE-2	0.53	0.49	0.07	0.31	0.37	0.37	0.08	0.22
	ROUGE-L	0.68	0.6	0.09	0.49	0.53	0.47	0.1	0.39
Qwen-2.5-7B - 8 Shot	BLEU	39.58	32.28	23.72	23.08	28.75	20.7	14.51	13.75
	BERTScore	0.85	0.84	0.79	0.76	0.91	0.78	0.73	0.72
	TER	62.65	58.75	72.23	71.32	76.87	74.53	75.67	83.31
	METEOR	0.7	0.57	0.46	0.42	0.5	0.41	0.34	0.31
	ChrF	64.76	48.86	21.33	43.77	48.97	37.33	14.89	35.02
	ROUGE-1	0.67	0.72	0.1	0.45	0.53	0.62	0.08	0.36
	ROUGE-2	0.49	0.51	0.08	0.25	0.35	0.39	0.06	0.18
	ROUGE-L	0.64	0.62	0.1	0.41	0.5	0.5	0.07	0.33
Qwen-2.5-7B - 16 Shot	BLEU	44.9	31.33	18.7	26.55	28.7	21.72	14.96	18.01
	BERTScore	0.87	0.83	0.79	0.81	0.91	0.78	0.74	0.77
	TER	51.04	60.82	68.76	64.25	76.32	75.03	75.23	76.62
	METEOR	0.72	0.55	0.42	0.5	0.51	0.42	0.35	0.38
	ChrF	66.38	47.26	18.54	49.7	49.28	38.09	15.52	41.47
	ROUGE-1	0.71	0.7	0.09	0.53	0.54	0.63	0.07	0.44
	ROUGE-2	0.53	0.49	0.07	0.31	0.36	0.4	0.05	0.24
	ROUGE-L	0.68	0.6	0.09	0.49	0.51	0.51	0.07	0.4
Qwen-2.5-7B - 32 Shot	BLEU	49.71	34.93	19.4	30.03	29.17	21.13	13.88	17.91
	BERTScore	0.9	0.85	0.79	0.84	0.91	0.79	0.73	0.78
	TER	41.23	55.79	68.27	60.61	66.49	74.12	74.79	76.11
	METEOR	0.74	0.6	0.42	0.55	0.51	0.42	0.34	0.38
	ChrF	68.28	51.18	19.2	54.6	49.55	38.05	14.66	41.88
	ROUGE-1	0.75	0.74	0.09	0.58	0.57	0.64	0.07	0.45
	ROUGE-2	0.56	0.53	0.08	0.34	0.38	0.4	0.06	0.24
	ROUGE-L	0.72	0.64	0.09	0.53	0.54	0.51	0.07	0.41
	BLEU	42.49	14.47	19.92	33.73	29.35	9.2	13.94	18.65

Table 15 continued from previous page

Model	Metrics	Ground-truth				ASR			
		fr-en	fr-vi	fr-zh	fr-de	fr-en	fr-vi	fr-zh	fr-de
Mistral-v0.3-7B - Ft.	BERTScore	0.93	0.79	0.78	0.85	0.79	0.74	0.76	0.78
	TER	56.11	74.72	67.02	57.73	74.28	82.83	74.11	74.55
	METEOR	0.75	0.36	0.42	0.58	0.52	0.26	0.34	0.39
	ChrF	68.36	30.93	20.91	56.06	49.85	23.69	15.16	41.66
	ROUGE-1	0.72	0.6	0.08	0.61	0.55	0.52	0.07	0.46
	ROUGE-2	0.55	0.37	0.05	0.38	0.38	0.27	0.05	0.25
	ROUGE-L	0.7	0.49	0.08	0.57	0.52	0.4	0.07	0.42
Mistral-v0.3-7B - 0 Shot	BLEU	42.47	4.86	22.46	21.79	25.09	2.45	14.33	12.68
	BERTScore	0.87	0.7	0.78	0.81	0.91	0.65	0.7	0.75
	TER	47.18	86.55	72.83	67.21	67.69	92.55	79.4	82.32
	METEOR	0.7	0.2	0.45	0.48	0.48	0.12	0.33	0.32
	ChrF	64.09	19.74	19.69	48.49	47.21	14.7	13.6	36.69
	ROUGE-1	0.71	0.47	0.12	0.52	0.55	0.39	0.08	0.39
	ROUGE-2	0.5	0.21	0.09	0.28	0.35	0.14	0.06	0.19
	ROUGE-L	0.68	0.35	0.12	0.48	0.51	0.29	0.08	0.35
Mistral-v0.3-7B - 1 Shot	BLEU	46.08	10.11	22.5	24.75	27.81	5.32	13.14	15.01
	BERTScore	0.89	0.78	0.8	0.82	0.92	0.73	0.72	0.77
	TER	44.67	78.64	67.06	64.05	63.69	85.9	75.34	76.91
	METEOR	0.73	0.3	0.46	0.51	0.51	0.21	0.33	0.35
	ChrF	66.58	26.79	20.3	50.42	49.02	20.47	13.9	39.43
	ROUGE-1	0.73	0.58	0.13	0.55	0.58	0.51	0.08	0.43
	ROUGE-2	0.53	0.31	0.09	0.3	0.39	0.23	0.06	0.21
	ROUGE-L	0.7	0.44	0.13	0.51	0.54	0.37	0.08	0.39
Mistral-v0.3-7B - 8 Shot	BLEU	48.65	9.48	15.76	27.24	30.24	8.36	12.72	17.61
	BERTScore	0.89	0.77	0.78	0.84	0.92	0.76	0.73	0.78
	TER	42.4	79.29	70.68	60.95	62.14	82.09	75.07	74.34
	METEOR	0.74	0.3	0.39	0.54	0.52	0.26	0.33	0.38
	ChrF	67.56	26.29	16.65	52.79	50.59	24.14	14.24	41.48
	ROUGE-1	0.74	0.57	0.1	0.58	0.58	0.55	0.07	0.46
	ROUGE-2	0.55	0.3	0.07	0.33	0.4	0.28	0.06	0.24
	ROUGE-L	0.71	0.43	0.1	0.54	0.55	0.41	0.07	0.42
Mistral-v0.3-7B - 16 Shot	BLEU	48.33	10.11	16.15	28.19	30.89	09.06	13.81	18.76
	BERTScore	0.89	0.78	0.78	0.84	0.92	0.76	0.73	0.78
	TER	41.91	78.64	70.06	60.43	61.73	81.41	74.05	72.97
	METEOR	0.74	0.3	0.39	0.55	0.53	0.27	0.34	0.39
	ChrF	67.39	26.79	17.05	53.07	50.88	24.54	15.33	42.08
	ROUGE-1	0.75	0.58	0.1	0.59	0.59	0.55	0.08	0.47
	ROUGE-2	0.55	0.31	0.07	0.34	0.41	0.28	0.05	0.26
	ROUGE-L	0.72	0.44	0.1	0.54	0.56	0.41	0.08	0.43
Mistral-v0.3-7B - 32 Shot	BLEU	49.35	10.57	16.67	28.83	29.19	8.12	12.39	17.13
	BERTScore	0.89	0.78	0.79	0.84	0.92	0.76	0.73	0.78
	TER	41.31	78.39	69.57	60.23	63.3	82.07	75.49	73.87
	METEOR	0.74	0.31	0.4	0.55	0.51	0.26	0.33	0.38
	ChrF	67.61	27.19	17.48	53.54	49.69	23.99	13.84	41.4
	ROUGE-1	0.75	0.58	0.1	0.59	0.58	0.54	0.07	0.46
	ROUGE-2	0.56	0.32	0.07	0.34	0.39	0.28	0.05	0.24
	ROUGE-L	0.72	0.45	0.1	0.55	0.55	0.4	0.07	0.42

Table 15 continued from previous page

Model	Metrics	Ground-truth				ASR			
		fr-en	fr-vi	fr-zh	fr-de	fr-en	fr-vi	fr-zh	fr-de

Table 16: **In-context Learning Results.** Full fine-tuning, zero-shot, few-shot on both ground-truth transcript and ASR transcript in the cascaded setting. **French to X** results are reported in this table.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

Model	Metrics	Ground-truth				ASR			
		de-en	de-vi	de-fr	de-zh	de-en	de-vi	de-fr	de-zh
Llama-3.1-8B - Ft.	BLEU	49.44	40.01	33.45	31.16	40.63	33.63	26.97	26.31
	BERTScore	0.95	0.87	0.84	0.81	0.86	0.84	0.8	0.79
	TER	44.99	54.45	72.36	62.08	56.53	63.07	79.21	68.19
	METEOR	0.76	0.67	0.62	0.54	0.63	0.56	0.52	0.47
	ChrF	69.74	57.57	61.36	27.18	58.95	49.74	54.06	22.87
	ROUGE-1	0.75	0.78	0.61	0.12	0.65	0.73	0.54	0.1
	ROUGE-2	0.57	0.58	0.42	0.09	0.48	0.51	0.35	0.07
	ROUGE-L	0.72	0.68	0.57	0.12	0.62	0.61	0.49	0.1
Llama-3.1-8B - 0 Shot	BLEU	40.09	25.98	29.35	10.77	30.35	20.74	22.36	10.55
	BERTScore	0.86	0.82	0.83	0.67	0.92	0.8	0.8	0.64
	TER	51.7	65.04	63.64	136.05	61.48	70.86	70.74	124.82
	METEOR	0.66	0.5	0.54	0.32	0.54	0.43	0.44	0.28
	ChrF	61.07	43.55	53.42	13.06	51.9	38.24	46.34	11.47
	ROUGE-1	0.68	0.66	0.58	0.09	0.59	0.63	0.51	0.08
	ROUGE-2	0.48	0.44	0.38	0.07	0.4	0.4	0.32	0.06
	ROUGE-L	0.64	0.54	0.54	0.09	0.56	0.5	0.47	0.08
Llama-3.1-8B - 1 Shot	BLEU	49.78	34.38	35.13	22.03	34.38	24.15	25.43	20.26
	BERTScore	0.9	0.86	0.86	0.79	0.93	0.82	0.82	0.76
	TER	41.64	56.94	58.72	68.59	56.38	64.68	67.46	70.79
	METEOR	0.76	0.61	0.6	0.46	0.59	0.48	0.49	0.42
	ChrF	68.35	52.11	58.58	19.58	55.43	42.51	49.63	18.39
	ROUGE-1	0.75	0.76	0.63	0.11	0.64	0.69	0.55	0.1
	ROUGE-2	0.56	0.53	0.42	0.08	0.44	0.45	0.35	0.08
	ROUGE-L	0.72	0.64	0.58	0.11	0.6	0.56	0.5	0.1
Llama-3.1-8B - 8 Shot	BLEU	48.83	32.76	35.32	22.4	36.04	26.67	27.05	18.32
	BERTScore	0.9	0.86	0.86	0.8	0.93	0.83	0.82	0.75
	TER	42.03	56.98	57.29	65.82	55.59	63.7	66.22	73.43
	METEOR	0.75	0.6	0.61	0.47	0.61	0.51	0.5	0.4
	ChrF	67.67	50.99	58.99	20.0	56.8	44.67	51.05	16.95
	ROUGE-1	0.75	0.76	0.63	0.12	0.64	0.7	0.56	0.1
	ROUGE-2	0.55	0.52	0.43	0.09	0.45	0.46	0.36	0.07
	ROUGE-L	0.71	0.63	0.59	0.12	0.61	0.58	0.51	0.1
Llama-3.1-8B - 16 Shot	BLEU	49.78	34.38	35.13	22.03	37.7	27.72	27.25	19.7
	BERTScore	0.9	0.86	0.86	0.79	0.94	0.83	0.83	0.75
	TER	41.64	56.94	58.72	68.59	53.54	64.49	65.72	71.5
	METEOR	0.76	0.61	0.6	0.46	0.62	0.51	0.5	0.41
	ChrF	68.35	52.11	58.58	19.58	57.92	45.3	50.88	18.11
	ROUGE-1	0.75	0.76	0.63	0.11	0.66	0.7	0.56	0.1
	ROUGE-2	0.56	0.53	0.42	0.08	0.46	0.47	0.36	0.07
	ROUGE-L	0.72	0.64	0.58	0.11	0.62	0.57	0.51	0.1
Llama-3.1-8B - 32 Shot	BLEU	49.07	35.32	34.95	23.09	36.84	27.09	26.84	19.25
	BERTScore	0.9	0.86	0.86	0.8	0.93	0.83	0.83	0.76
	TER	42.3	55.79	57.43	65.91	54.34	64.29	66.25	70.33
	METEOR	0.75	0.62	0.6	0.47	0.61	0.51	0.5	0.41
	ChrF	68.06	52.79	58.84	20.55	57.55	45.02	50.56	17.79
	ROUGE-1	0.75	0.77	0.64	0.12	0.66	0.7	0.56	0.11
	ROUGE-2	0.55	0.53	0.43	0.1	0.46	0.46	0.36	0.08
	ROUGE-L	0.72	0.65	0.59	0.12	0.62	0.57	0.51	0.11
	BLEU	52.1	43.73	40.72	23.26	40.52	34.24	31.45	19.87

Table 16 continued from previous page

Model	Metrics	Ground-truth				ASR			
		de-en	de-vi	de-fr	de-zh	de-en	de-vi	de-fr	de-zh
Qwen-2.5-7B - Ft.	BERTScore	0.96	0.88	0.88	0.79	0.86	0.84	0.84	0.79
	TER	39.94	48.63	51.44	63.95	55.76	59.16	61.65	68.06
	METEOR	0.77	0.69	0.65	0.44	0.63	0.57	0.53	0.4
	ChrF	70.13	59.36	62.75	23.28	59.05	50.58	54.36	20.15
	ROUGE-1	0.76	0.8	0.67	0.11	0.66	0.74	0.59	0.1
	ROUGE-2	0.58	0.6	0.47	0.09	0.47	0.53	0.4	0.08
	ROUGE-L	0.73	0.7	0.63	0.11	0.62	0.62	0.54	0.1
Qwen-2.5-7B - 0 Shot	BLEU	42.12	27.26	29.93	33.09	32.03	20.9	22.74	29.78
	BERTScore	0.86	0.81	0.83	0.83	0.92	0.79	0.79	0.79
	TER	54.87	65.49	63.02	63.12	65.25	71.74	70.78	67.85
	METEOR	0.66	0.5	0.53	0.62	0.54	0.42	0.44	0.52
	ChrF	61.0	43.92	53.64	31.26	51.78	37.93	46.53	25.63
	ROUGE-1	0.65	0.66	0.56	0.12	0.57	0.62	0.49	0.11
	ROUGE-2	0.46	0.44	0.36	0.1	0.38	0.39	0.3	0.09
	ROUGE-L	0.62	0.55	0.52	0.12	0.53	0.5	0.45	0.11
Qwen-2.5-7B - 1 Shot	BLEU	40.55	28.29	32.3	19.06	35.85	23.42	25.1	19.58
	BERTScore	0.89	0.83	0.85	0.79	0.93	0.81	0.82	0.76
	TER	57.2	65.45	62.21	68.52	57.82	66.64	67.69	72.04
	METEOR	0.6	0.5	0.56	0.4	0.6	0.47	0.47	0.4
	ChrF	58.35	44.68	56.84	19.01	56.18	41.7	49.54	18.35
	ROUGE-1	0.6	0.66	0.59	0.09	0.63	0.67	0.53	0.09
	ROUGE-2	0.42	0.44	0.38	0.07	0.43	0.44	0.33	0.07
	ROUGE-L	0.57	0.55	0.55	0.09	0.6	0.55	0.49	0.09
Qwen-2.5-7B - 8 Shot	BLEU	24.79	31.91	29.74	18.83	27.78	24.65	26.31	18.09
	BERTScore	0.72	0.84	0.81	0.73	0.89	0.81	0.82	0.76
	TER	75.48	61.27	63.77	94.17	75.66	67.64	65.23	71.02
	METEOR	0.39	0.57	0.51	0.35	0.46	0.46	0.49	0.38
	ChrF	38.36	49.02	52.61	16.91	45.07	41.71	50.37	17.82
	ROUGE-1	0.41	0.72	0.54	0.07	0.48	0.65	0.54	0.09
	ROUGE-2	0.27	0.49	0.35	0.06	0.31	0.43	0.35	0.07
	ROUGE-L	0.38	0.6	0.5	0.07	0.45	0.53	0.5	0.09
Qwen-2.5-7B - 16 Shot	BLEU	40.55	28.29	32.3	19.06	36.96	25.86	26.52	16.81
	BERTScore	0.82	0.81	0.84	0.79	0.93	0.81	0.82	0.76
	TER	57.2	65.45	62.21	68.52	60.11	68.28	65.34	70.93
	METEOR	0.6	0.5	0.56	0.4	0.6	0.48	0.5	0.37
	ChrF	58.35	44.68	56.84	19.01	56.41	42.7	51.05	17.27
	ROUGE-1	0.6	0.66	0.59	0.09	0.62	0.66	0.55	0.09
	ROUGE-2	0.42	0.44	0.38	0.07	0.43	0.43	0.35	0.07
	ROUGE-L	0.57	0.55	0.55	0.09	0.59	0.54	0.51	0.09
Qwen-2.5-7B - 32 Shot	BLEU	48.91	33.61	35.22	20.52	37.22	26.41	26.92	17.3
	BERTScore	0.9	0.84	0.86	0.8	0.93	0.81	0.83	0.76
	TER	42.63	59.35	55.55	66.94	55.91	67.06	64.63	70.33
	METEOR	0.75	0.59	0.6	0.42	0.61	0.49	0.5	0.38
	ChrF	67.98	50.16	59.21	20.28	57.36	43.45	51.4	17.6
	ROUGE-1	0.75	0.73	0.63	0.1	0.65	0.68	0.56	0.1
	ROUGE-2	0.55	0.51	0.42	0.08	0.45	0.45	0.36	0.07
	ROUGE-L	0.71	0.61	0.59	0.1	0.61	0.55	0.51	0.09
	BLEU	36.39	15.68	40.77	21.28	28.33	12.38	31.15	17.82

Table 16 continued from previous page

Model	Metrics	Ground-truth				ASR			
		de-en	de-vi	de-fr	de-zh	de-en	de-vi	de-fr	de-zh
Mistral-v0.3-7B - Ft.	BERTScore	0.92	0.79	0.86	0.78	0.78	0.77	0.83	0.78
	TER	80.09	73.73	53.76	66.07	91.13	77.89	63.4	70.11
	METEOR	0.72	0.35	0.63	0.42	0.6	0.3	0.52	0.37
	ChrF	66.77	32.48	61.34	22.37	57.3	28.44	52.73	18.9
	ROUGE-1	0.64	0.61	0.66	0.09	0.55	0.57	0.57	0.09
	ROUGE-2	0.49	0.38	0.47	0.07	0.4	0.33	0.39	0.07
	ROUGE-L	0.61	0.48	0.62	0.09	0.52	0.44	0.53	0.09
Mistral-v0.3-7B - 0 Shot	BLEU	44.26	6.71	28.82	27.14	34.54	4.98	21.44	21.22
	BERTScore	0.89	0.73	0.84	0.81	0.93	0.71	0.8	0.75
	TER	48.08	84.37	63.72	63.83	59.16	87.37	71.68	72.01
	METEOR	0.72	0.22	0.54	0.52	0.59	0.18	0.44	0.43
	ChrF	64.65	22.36	54.28	23.4	55.07	19.2	46.78	18.85
	ROUGE-1	0.71	0.52	0.59	0.12	0.62	0.48	0.51	0.1
	ROUGE-2	0.5	0.24	0.36	0.1	0.42	0.2	0.3	0.08
	ROUGE-L	0.67	0.37	0.54	0.12	0.58	0.34	0.46	0.1
Mistral-v0.3-7B - 1 Shot	BLEU	48.04	10.15	30.75	24.33	36.31	6.99	22.89	16.01
	BERTScore	0.9	0.76	0.85	0.81	0.93	0.77	0.82	0.74
	TER	44.93	79.06	60.96	64.11	56.18	83.25	68.21	72.27
	METEOR	0.75	0.29	0.57	0.48	0.6	0.24	0.46	0.37
	ChrF	67.21	26.69	55.61	21.94	56.33	23.1	48.16	16.11
	ROUGE-1	0.74	0.57	0.6	0.11	0.64	0.54	0.53	0.09
	ROUGE-2	0.54	0.3	0.38	0.09	0.44	0.26	0.32	0.07
	ROUGE-L	0.7	0.43	0.56	0.11	0.6	0.39	0.48	0.09
Mistral-v0.3-7B - 8 Shot	BLEU	47.79	9.31	31.72	15.2	37.63	8.39	24.35	13.54
	BERTScore	0.9	0.77	0.85	0.78	0.93	0.76	0.82	0.75
	TER	44.84	79.09	59.29	70.09	54.56	80.9	66.67	72.64
	METEOR	0.75	0.29	0.57	0.37	0.62	0.27	0.47	0.34
	ChrF	67.13	26.2	56.37	16.5	57.28	24.75	49.07	15.11
	ROUGE-1	0.74	0.57	0.61	0.09	0.65	0.56	0.54	0.07
	ROUGE-2	0.54	0.3	0.39	0.07	0.46	0.28	0.34	0.06
	ROUGE-L	0.7	0.43	0.57	0.09	0.62	0.41	0.5	0.07
Mistral-v0.3-7B - 16 Shot	BLEU	48.04	10.15	32.05	15.2	38.29	9.12	25.27	14.0
	BERTScore	0.9	0.77	0.85	0.78	0.93	0.77	0.82	0.74
	TER	44.93	79.06	59.36	70.52	55.25	80.37	66.71	72.83
	METEOR	0.75	0.29	0.58	0.37	0.62	0.27	0.48	0.34
	ChrF	67.21	26.69	56.27	16.48	57.55	25.3	49.37	15.38
	ROUGE-1	0.74	0.57	0.61	0.09	0.65	0.56	0.54	0.09
	ROUGE-2	0.54	0.3	0.4	0.07	0.46	0.29	0.34	0.07
	ROUGE-L	0.7	0.43	0.57	0.09	0.61	0.41	0.5	0.09
Mistral-v0.3-7B - 32 Shot	BLEU	48.21	10.66	31.71	16.08	37.85	8.81	24.2	13.87
	BERTScore	0.9	0.78	0.85	0.78	0.93	0.76	0.82	0.75
	TER	44.1	78.22	58.9	70.11	55.57	80.55	66.83	72.73
	METEOR	0.75	0.3	0.57	0.38	0.62	0.27	0.47	0.34
	ChrF	67.3	27.12	56.11	17.19	57.32	25.25	48.92	15.29
	ROUGE-1	0.74	0.57	0.61	0.08	0.65	0.56	0.54	0.07
	ROUGE-2	0.54	0.31	0.39	0.06	0.45	0.29	0.33	0.06
	ROUGE-L	0.71	0.43	0.56	0.08	0.62	0.41	0.5	0.07

Table 16 continued from previous page

Model	Metrics	Ground-truth				ASR			
		de-en	de-vi	de-fr	de-zh	de-en	de-vi	de-fr	de-zh

Table 17: **In-context Learning Results.** Full fine-tuning, zero-shot, few-shot on both ground-truth transcript and ASR transcript in the cascaded setting. **German to X** results are reported in this table.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

Model	Metrics	Ground-truth				ASR			
		zh-en	zh-vi	zh-fr	zh-de	zh-en	zh-vi	zh-fr	zh-de
Llama-3.1-8B - Ft.	BLEU	28.21	23.49	18.87	13.07	19.01	17.65	13.84	11.13
	BERTScore	0.91	0.79	0.77	0.74	0.78	0.75	0.75	0.74
	TER	84.4	96.2	99.18	107.48	101.13	104.68	106.77	109.71
	METEOR	0.59	0.55	0.46	0.4	0.48	0.45	0.39	0.35
	ChrF	47.42	41.33	41.69	32.09	40.96	34.68	36.97	32.37
	ROUGE-1	0.57	0.63	0.46	0.37	0.47	0.58	0.39	0.34
	ROUGE-2	0.37	0.44	0.28	0.19	0.27	0.36	0.21	0.16
	ROUGE-L	0.53	0.53	0.41	0.33	0.43	0.46	0.34	0.3
Llama-3.1-8B - 0 Shot	BLEU	16.21	16.01	11.95	7.41	12.29	12.17	8.78	06.02
	BERTScore	0.79	0.86	0.76	0.75	0.89	0.77	0.74	0.73
	TER	80.87	79.51	85.88	91.67	89.32	87.12	93.56	95.73
	METEOR	0.46	0.44	0.35	0.31	0.37	0.35	0.29	0.26
	ChrF	36.29	31.91	33.47	26.79	32.05	27.57	28.97	23.41
	ROUGE-1	0.47	0.61	0.4	0.35	0.39	0.55	0.33	0.28
	ROUGE-2	0.25	0.38	0.2	0.15	0.18	0.31	0.15	0.12
	ROUGE-L	0.43	0.49	0.35	0.31	0.35	0.43	0.28	0.25
Llama-3.1-8B - 1 Shot	BLEU	27.32	24.22	16.23	12.69	15.79	13.94	9.28	08.02
	BERTScore	0.85	0.91	0.81	0.78	0.92	0.8	0.78	0.74
	TER	65.78	70.48	79.85	83.63	84.27	84.98	90.0	92.65
	METEOR	0.59	0.55	0.44	0.41	0.43	0.39	0.31	0.3
	ChrF	48.7	41.34	40.15	35.9	36.21	29.88	30.57	25.37
	ROUGE-1	0.59	0.7	0.47	0.44	0.44	0.6	0.35	0.32
	ROUGE-2	0.36	0.47	0.27	0.21	0.22	0.34	0.17	0.15
	ROUGE-L	0.55	0.58	0.42	0.4	0.39	0.46	0.3	0.29
Llama-3.1-8B - 8 Shot	BLEU	27.76	24.74	15.35	11.82	21.16	19.56	14.14	12.08
	BERTScore	0.85	0.91	0.8	0.8	0.92	0.8	0.78	0.79
	TER	66.94	69.48	79.05	82.4	80.6	81.08	89.56	87.74
	METEOR	0.59	0.56	0.43	0.4	0.48	0.44	0.36	0.36
	ChrF	48.41	41.96	39.48	34.26	41.81	35.35	36.66	33.07
	ROUGE-1	0.59	0.71	0.46	0.44	0.49	0.64	0.4	0.4
	ROUGE-2	0.36	0.48	0.25	0.2	0.28	0.38	0.21	0.21
	ROUGE-L	0.55	0.59	0.42	0.39	0.44	0.5	0.34	0.36
Llama-3.1-8B - 16 Shot	BLEU	27.32	24.22	16.23	12.69	24.3	21.86	16.72	17.19
	BERTScore	0.85	0.91	0.81	0.81	0.92	0.8	0.78	0.79
	TER	65.78	70.48	79.85	83.63	78.01	79.45	86.03	85.58
	METEOR	0.59	0.55	0.44	0.41	0.49	0.46	0.38	0.39
	ChrF	48.7	41.34	40.15	35.9	44.39	37.37	38.22	37.05
	ROUGE-1	0.59	0.7	0.47	0.44	0.5	0.65	0.41	0.42
	ROUGE-2	0.36	0.47	0.27	0.21	0.29	0.4	0.22	0.23
	ROUGE-L	0.55	0.58	0.42	0.4	0.45	0.51	0.36	0.38
Llama-3.1-8B - 32 Shot	BLEU	28.69	24.39	16.43	12.34	19.18	19.54	12.37	10.51
	BERTScore	0.85	0.91	0.81	0.81	0.92	0.81	0.78	0.78
	TER	67.3	69.1	81.45	82.31	81.2	79.47	88.9	89.66
	METEOR	0.6	0.55	0.45	0.42	0.47	0.46	0.36	0.37
	ChrF	49.42	41.82	40.47	36.57	41.43	36.04	36.19	33.27
	ROUGE-1	0.6	0.71	0.47	0.45	0.48	0.65	0.4	0.4
	ROUGE-2	0.37	0.47	0.27	0.21	0.26	0.39	0.2	0.18
	ROUGE-L	0.55	0.59	0.42	0.41	0.43	0.51	0.35	0.34
	BLEU	35.63	32.95	24.05	16.95	25.36	26.31	17.84	12.61

Table 17 continued from previous page

Model	Metrics	Ground-truth				ASR			
		zh-en	zh-vi	zh-fr	zh-de	zh-en	zh-vi	zh-fr	zh-de
Qwen-2.5-7B - Ft.	BERTScore	0.95	0.85	0.84	0.83	0.82	0.81	0.79	0.78
	TER	60.98	67.86	73.13	77.31	79.48	80.4	84.11	91.04
	METEOR	0.67	0.63	0.53	0.5	0.53	0.53	0.43	0.4
	ChrF	55.3	47.73	47.09	39.65	45.87	41.37	40.88	34.86
	ROUGE-1	0.66	0.74	0.55	0.52	0.53	0.67	0.45	0.42
	ROUGE-2	0.45	0.53	0.35	0.28	0.32	0.45	0.26	0.2
	ROUGE-L	0.62	0.63	0.5	0.48	0.48	0.55	0.4	0.37
Qwen-2.5-7B - 0 Shot	BLEU	22.51	15.93	17.86	7.58	16.9	14.42	12.07	7.78
	BERTScore	0.83	0.82	0.8	0.77	0.91	0.77	0.77	0.74
	TER	74.01	76.17	78.49	87.24	87.43	85.97	88.63	95.0
	METEOR	0.53	0.41	0.44	0.34	0.43	0.37	0.35	0.31
	ChrF	44.85	32.32	40.84	29.61	37.97	29.62	33.97	28.17
	ROUGE-1	0.56	0.61	0.47	0.39	0.44	0.57	0.39	0.34
	ROUGE-2	0.32	0.39	0.27	0.16	0.23	0.34	0.2	0.14
	ROUGE-L	0.51	0.5	0.42	0.34	0.4	0.45	0.34	0.3
Qwen-2.5-7B - 1 Shot	BLEU	28.38	22.63	20.23	11.97	20.61	16.91	13.16	9.4
	BERTScore	0.86	0.85	0.81	0.8	0.91	0.8	0.79	0.77
	TER	67.96	71.83	76.08	82.62	83.98	79.69	87.86	93.12
	METEOR	0.59	0.5	0.48	0.42	0.49	0.41	0.35	0.33
	ChrF	47.86	37.41	43.55	34.08	42.1	33.0	35.74	30.35
	ROUGE-1	0.59	0.64	0.5	0.46	0.49	0.61	0.39	0.36
	ROUGE-2	0.38	0.44	0.31	0.22	0.27	0.38	0.2	0.16
	ROUGE-L	0.55	0.54	0.46	0.42	0.44	0.48	0.34	0.32
Qwen-2.5-7B - 8 Shot	BLEU	32.37	22.67	20.25	12.79	21.73	18.36	16.13	12.29
	BERTScore	0.85	0.86	0.81	0.8	0.9	0.79	0.79	0.78
	TER	64.09	70.76	75.17	81.02	80.02	79.28	85.2	89.03
	METEOR	0.62	0.5	0.48	0.43	0.46	0.43	0.39	0.36
	ChrF	51.54	38.91	43.96	35.4	40.54	34.22	38.5	33.67
	ROUGE-1	0.61	0.66	0.5	0.45	0.47	0.62	0.43	0.41
	ROUGE-2	0.4	0.45	0.3	0.22	0.28	0.39	0.24	0.2
	ROUGE-L	0.58	0.56	0.46	0.41	0.42	0.49	0.38	0.37
Qwen-2.5-7B - 16 Shot	BLEU	28.38	22.63	20.23	11.29	22.03	20.66	17.3	14.75
	BERTScore	0.84	0.85	0.81	0.79	0.91	0.81	0.79	0.77
	TER	67.96	71.83	76.08	83.85	81.52	76.13	82.84	87.24
	METEOR	0.59	0.5	0.48	0.42	0.48	0.45	0.41	0.38
	ChrF	47.86	37.41	43.55	33.81	40.79	36.53	39.15	33.33
	ROUGE-1	0.59	0.64	0.5	0.45	0.49	0.63	0.43	0.4
	ROUGE-2	0.38	0.44	0.31	0.22	0.29	0.41	0.25	0.22
	ROUGE-L	0.55	0.54	0.46	0.41	0.44	0.51	0.38	0.37
Qwen-2.5-7B - 32 Shot	BLEU	33.23	25.13	20.92	11.48	23.3	18.88	15.24	10.5
	BERTScore	0.86	0.87	0.82	0.8	0.92	0.79	0.79	0.78
	TER	64.02	69.67	75.67	84.04	77.07	83.31	85.2	90.95
	METEOR	0.64	0.53	0.5	0.41	0.52	0.43	0.4	0.37
	ChrF	53.52	41.0	44.95	34.77	44.66	33.94	38.97	33.36
	ROUGE-1	0.64	0.68	0.52	0.45	0.52	0.61	0.43	0.4
	ROUGE-2	0.42	0.47	0.31	0.21	0.3	0.38	0.24	0.19
	ROUGE-L	0.59	0.57	0.47	0.4	0.47	0.49	0.38	0.35
	BLEU	27.68	10.67	18.46	11.4	20.17	08.01	12.58	7.14

Table 17 continued from previous page

Model	Metrics	Ground-truth				ASR			
		zh-en	zh-vi	zh-fr	zh-de	zh-en	zh-vi	zh-fr	zh-de
Mistral-v0.3-7B - Ft.	BERTScore	0.93	0.75	0.8	0.76	0.8	0.73	0.76	0.72
	TER	72.65	82.73	81.15	90.07	83.87	85.69	87.85	98.87
	METEOR	0.59	0.33	0.45	0.39	0.48	0.27	0.37	0.29
	ChrF	47.23	24.79	39.83	31.55	40.58	21.71	34.68	25.42
	ROUGE-1	0.57	0.56	0.47	0.4	0.48	0.53	0.39	0.3
	ROUGE-2	0.36	0.32	0.28	0.2	0.26	0.27	0.2	0.13
	ROUGE-L	0.53	0.43	0.42	0.37	0.43	0.4	0.34	0.27
Mistral-v0.3-7B - 0 Shot	BLEU	22.94	16.76	17.51	8.66	12.4	1.22	3.65	01.05
	BERTScore	0.83	0.83	0.8	0.78	0.89	0.6	0.64	0.59
	TER	72.61	75.49	78.0	88.21	93.41	98.06	113.15	114.58
	METEOR	0.54	0.43	0.44	0.36	0.37	0.07	0.16	0.09
	ChrF	44.74	33.32	40.21	30.97	32.91	8.79	19.5	12.75
	ROUGE-1	0.55	0.62	0.47	0.4	0.39	0.28	0.18	0.1
	ROUGE-2	0.32	0.4	0.28	0.17	0.17	0.08	0.07	0.02
	ROUGE-L	0.5	0.51	0.42	0.35	0.34	0.21	0.16	0.09
Mistral-v0.3-7B - 1 Shot	BLEU	30.82	17.97	19.57	11.3	16.52	2.42	6.53	2.85
	BERTScore	0.85	0.83	0.8	0.81	0.92	0.74	0.67	0.62
	TER	65.66	73.89	76.46	83.38	90.42	95.99	105.51	112.79
	METEOR	0.62	0.44	0.47	0.4	0.43	0.12	0.22	0.14
	ChrF	51.57	34.53	43.3	33.91	37.51	12.0	24.06	16.69
	ROUGE-1	0.62	0.61	0.51	0.44	0.43	0.37	0.24	0.16
	ROUGE-2	0.4	0.41	0.29	0.21	0.21	0.13	0.11	0.05
	ROUGE-L	0.58	0.51	0.45	0.39	0.39	0.27	0.21	0.14
Mistral-v0.3-7B - 8 Shot	BLEU	32.47	22.55	20.32	12.8	19.65	6.8	11.95	9.42
	BERTScore	0.85	0.85	0.81	0.8	0.92	0.73	0.76	0.75
	TER	64.66	71.4	74.39	83.66	83.2	87.81	90.14	95.41
	METEOR	0.62	0.48	0.49	0.43	0.47	0.24	0.34	0.32
	ChrF	51.57	38.16	43.95	34.8	40.99	20.18	33.31	29.01
	ROUGE-1	0.61	0.64	0.5	0.45	0.48	0.52	0.37	0.34
	ROUGE-2	0.4	0.43	0.3	0.22	0.26	0.25	0.19	0.16
	ROUGE-L	0.58	0.54	0.46	0.41	0.43	0.38	0.32	0.3
Mistral-v0.3-7B - 16 Shot	BLEU	27.42	24.62	16.8	12.0	22.51	6.97	14.27	12.93
	BERTScore	0.85	0.91	0.8	0.81	0.92	0.74	0.77	0.75
	TER	66.91	70.38	79.92	82.87	81.31	88.06	88.7	93.21
	METEOR	0.59	0.55	0.43	0.41	0.49	0.25	0.36	0.35
	ChrF	49.02	41.34	39.91	35.45	42.73	20.12	34.96	29.34
	ROUGE-1	0.6	0.7	0.46	0.44	0.49	0.53	0.4	0.37
	ROUGE-2	0.36	0.47	0.26	0.21	0.28	0.25	0.22	0.21
	ROUGE-L	0.55	0.58	0.41	0.4	0.44	0.38	0.35	0.34
Mistral-v0.3-7B - 32 Shot	BLEU	27.67	8.27	15.7	9.56	19.19	6.72	12.05	7.59
	BERTScore	0.84	0.88	0.8	0.78	0.92	0.75	0.77	0.76
	TER	69.14	83.79	82.81	87.74	82.35	86.05	89.62	93.68
	METEOR	0.6	0.31	0.43	0.38	0.48	0.27	0.35	0.31
	ChrF	48.47	23.3	39.44	32.17	40.86	21.51	34.87	28.86
	ROUGE-1	0.59	0.58	0.44	0.41	0.48	0.55	0.38	0.35
	ROUGE-2	0.36	0.3	0.24	0.17	0.25	0.26	0.19	0.13
	ROUGE-L	0.54	0.43	0.4	0.36	0.43	0.4	0.32	0.3

Table 17 continued from previous page

Model	Metrics	Ground-truth				ASR			
		zh-en	zh-vi	zh-fr	zh-de	zh-en	zh-vi	zh-fr	zh-de

Table 18: **In-context Learning Results.** Full fine-tuning, zero-shot, few-shot on both ground-truth transcript and ASR transcript in the cascaded setting. **Chinese to X** results are reported in this table.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

F.2 Full Results: Ground-truth Translation Baselines

This section presents full results of ground-truth MT baselines for all evaluation metrics, which is an extension of Table 3 in the main paper. Full results are shown in Table 19 (English to X), Table 20 (Vietnamese to X), Table 21 (French to X), Table 22 (German to X), and Table 23 (Chinese to X) below.

MT	Metrics	en-vi	en-fr	en-zh	en-de
Decoder					
Llama -3.1-8B	BLEU	53.44	48.24	37.50	40.49
	BERTSc	0.90	0.89	0.83	0.87
	TER	39.42	46.9	58.37	53.12
	METEOR	0.77	0.72	0.6	0.66
	ChrF	67.74	70.97	32.39	66.08
	ROUGE-1	0.83	0.73	0.15	0.67
	ROUGE-2	0.68	0.57	0.13	0.47
	ROUGE-L	0.76	0.7	0.15	0.63
Qwen -2.5-7B	BLEU	54.5	49.63	28.61	38.75
	BERTSc	0.9	0.9	0.81	0.87
	TER	38.21	42.42	59.1	51.55
	METEOR	0.77	0.72	0.5	0.64
	ChrF	67.82	70.5	27.81	63.39
	ROUGE-1	0.83	0.74	0.14	0.65
	ROUGE-2	0.69	0.57	0.13	0.43
	ROUGE-L	0.77	0.71	0.14	0.62
Mistral -v0.3-7B	BLEU	24.77	51.71	26.38	43.99
	BERTSc	0.82	0.89	0.81	0.88
	TER	63.33	43.0	60.43	48.54
	METEOR	0.45	0.73	0.48	0.67
	ChrF	41.27	71.06	27.29	66.27
	ROUGE-1	0.66	0.74	0.12	0.68
	ROUGE-2	0.49	0.6	0.11	0.49
	ROUGE-L	0.57	0.72	0.12	0.65
Encoder-decoder					
mBart -large-50	BLEU	62.08	57.04	44.77	47.28
	BERTSc	0.92	0.92	0.86	0.89
	TER	29.87	34.26	41.08	42.76
	METEOR	0.81	0.77	0.68	0.7
	ChrF	72.75	75.3	38.7	68.74
	ROUGE-1	0.86	0.78	0.19	0.71
	ROUGE-2	0.74	0.64	0.17	0.52
	ROUGE-L	0.81	0.76	0.19	0.68
M2M100 -418M	BLEU	62.31	57.49	46.38	49.36
	BERTSc	0.97	0.95	0.93	0.94
	TER	29.4	33.52	39.38	40.6
	METEOR	0.81	0.77	0.7	0.72
	ChrF	73.15	75.72	40.11	71.04
	ROUGE-1	0.86	0.79	0.2	0.73
	ROUGE-2	0.74	0.64	0.19	0.54
	ROUGE-L	0.81	0.76	0.2	0.7
Marian	BLEU	58.22	53.84	38.67	45.81
	BERTSc	0.91	0.91	0.85	0.89
	TER	32.56	36.53	45.9	43.77
	METEOR	0.79	0.75	0.64	0.69
	ChrF	70.19	73.68	33.27	68.72
	ROUGE-1	0.85	0.77	0.19	0.71
	ROUGE-2	0.71	0.61	0.18	0.5
	ROUGE-L	0.79	0.74	0.19	0.67

Table 18 continued from previous page

MT	Metrics	en-vi	en-fr	en-zh	en-de
Commercial tool					
Google Translate	BLEU	46.21	44.77	44.74	36.29
	BERTSc	0.91	0.91	0.9	0.89
	TER	48.02	50.96	52.47	59.78
	METEOR	0.67	0.64	0.63	0.58
	ChrF	59.7	64.44	39.17	59.39
	ROUGE-1	0.78	0.7	0.16	0.63
	ROUGE-2	0.64	0.56	0.14	0.44
	ROUGE-L	0.72	0.67	0.16	0.6

Table 19: **Full Results: Ground-truth Translation Baselines. English to X** results are reported in this table. Extension of Table 3 in the main paper.

MT	Metrics	vi-en	vi-fr	vi-zh	vi-de
Decoder					
Llama -3.1-8B	BLEU	23.16	15.57	16.09	11.61
	BERTSc	0.92	0.79	0.74	0.77
	TER	83.07	100.23	120.58	112.85
	METEOR	0.57	0.45	0.5	0.39
	ChrF	52.29	47.18	21.03	44.25
	ROUGE-1	0.56	0.45	0.03	0.4
	ROUGE-2	0.33	0.25	0.01	0.19
	ROUGE-L	0.51	0.4	0.03	0.36
Qwen -2.5-7B	BLEU	26.21	19.25	29.06	14.44
	BERTSc	0.93	0.81	0.81	0.79
	TER	71.1	80.86	60.93	83.75
	METEOR	0.57	0.47	0.55	0.39
	ChrF	53.79	48.94	25.1	43.34
	ROUGE-1	0.58	0.49	0.04	0.42
	ROUGE-2	0.35	0.28	0.02	0.19
	ROUGE-L	0.53	0.44	0.04	0.38
Mistral -v0.3-7B	BLEU	24.56	16.0	25.04	13.38
	BERTSc	0.92	0.79	0.78	0.78
	TER	70.37	87.87	65.43	84.13
	METEOR	0.53	0.42	0.49	0.37
	ChrF	48.73	43.93	21.84	40.96
	ROUGE-1	0.54	0.44	0.03	0.4
	ROUGE-2	0.31	0.24	0.01	0.18
	ROUGE-L	0.5	0.39	0.03	0.36
Encoder-decoder					
mBart -large-50	BLEU	13.34	17.9	22.97	9.85
	BERTSc	0.89	0.8	0.78	0.75
	TER	79.27	76.57	62.13	95.17
	METEOR	0.38	0.41	0.48	0.29
	ChrF	36.03	43.81	20.38	34.93
	ROUGE-1	0.42	0.45	0.03	0.32
	ROUGE-2	0.18	0.25	0.01	0.13
	ROUGE-L	0.38	0.41	0.03	0.29
M2M100 -418M	BLEU	23.01	21.01	24.95	16.72
	BERTSc	0.82	0.81	0.8	0.79
	TER	68.26	73.61	60.67	77.59
	METEOR	0.51	0.45	0.51	0.4
	ChrF	48.0	48.05	21.95	44.52
	ROUGE-1	0.54	0.49	0.04	0.45
	ROUGE-2	0.29	0.29	0.02	0.21
	ROUGE-L	0.49	0.45	0.04	0.4
Marian	BLEU	17.63	15.97	15.56	12.84
	BERTSc	0.8	0.79	0.78	0.77
	TER	75.09	79.31	71.43	82.01
	METEOR	0.45	0.39	0.4	0.35
	ChrF	42.61	42.83	14.79	39.82
	ROUGE-1	0.48	0.44	0.03	0.4
	ROUGE-2	0.23	0.23	0.01	0.17
	ROUGE-L	0.44	0.4	0.03	0.35

Table 19 continued from previous page

MT	Metrics	vi-en	vi-fr	vi-zh	vi-de
Commercial tool					
Google Translate	BLEU	18.79	16.42	21.63	12.54
	BERTSc	0.84	0.83	0.83	0.81
	TER	75.72	82.76	71.96	87.12
	METEOR	0.43	0.37	0.44	0.32
	ChrF	43.91	42.82	19.03	40.21
	ROUGE-1	0.49	0.43	0.02	0.38
	ROUGE-2	0.26	0.24	0.01	0.17
	ROUGE-L	0.45	0.38	0.02	0.33

Table 20: **Full Results: Ground-truth Translation Baselines. Vietnamese to X** results are reported in this table. Extension of Table 3 in the main paper.

MT	Metrics	fr-en	fr-vi	fr-zh	fr-de
Decoder					
Llama -3.1-8B	BLEU	50.18	39.63	29.25	27.46
	BERTSc	0.95	0.86	0.79	0.81
	TER	42.23	54.71	68.66	82.2
	METEOR	0.76	0.65	0.52	0.56
	ChrF	69.44	55.16	25.3	56.35
	ROUGE-1	0.75	0.77	0.11	0.54
	ROUGE-2	0.58	0.56	0.07	0.34
	ROUGE-L	0.73	0.67	0.11	0.5
Qwen -2.5-7B	BLEU	49.69	40.67	20.97	33.91
	BERTSc	0.95	0.86	0.78	0.84
	TER	42.17	52.57	67.39	59.28
	METEOR	0.76	0.66	0.43	0.58
	ChrF	70.03	56.12	20.86	56.82
	ROUGE-1	0.76	0.78	0.09	0.6
	ROUGE-2	0.58	0.58	0.06	0.37
	ROUGE-L	0.73	0.68	0.08	0.56
Mistral -v0.3-7B	BLEU	42.49	14.47	19.92	33.73
	BERTSc	0.93	0.79	0.78	0.85
	TER	56.11	74.72	67.02	57.73
	METEOR	0.75	0.36	0.42	0.58
	ChrF	68.36	30.93	20.91	56.06
	ROUGE-1	0.72	0.6	0.08	0.61
	ROUGE-2	0.55	0.37	0.05	0.38
	ROUGE-L	0.7	0.49	0.08	0.57
Encoder-decoder					
mBart -large-50	BLEU	39.79	37.26	24.63	29.03
	BERTSc	0.93	0.86	0.77	0.83
	TER	51.84	54.81	63.66	64.46
	METEOR	0.67	0.62	0.49	0.53
	ChrF	59.96	53.25	21.41	51.13
	ROUGE-1	0.67	0.76	0.13	0.55
	ROUGE-2	0.45	0.53	0.1	0.32
	ROUGE-L	0.63	0.65	0.13	0.5
M2M100 -418M	BLEU	43.73	35.04	29.41	34.72
	BERTSc	0.88	0.82	0.82	0.83
	TER	45.48	57.02	53.94	55.57
	METEOR	0.71	0.57	0.56	0.59
	ChrF	63.79	49.8	25.65	57.69
	ROUGE-1	0.71	0.67	0.15	0.6
	ROUGE-2	0.49	0.49	0.11	0.37
	ROUGE-L	0.68	0.59	0.15	0.56
Marian	BLEU	39.97	33.41	17.13	32.62
	BERTSc	0.87	0.86	0.78	0.85
	TER	48.76	57.15	67.55	56.39
	METEOR	0.68	0.62	0.41	0.59
	ChrF	60.97	52.06	16.05	56.73
	ROUGE-1	0.69	0.76	0.14	0.61
	ROUGE-2	0.46	0.53	0.1	0.36
	ROUGE-L	0.65	0.65	0.14	0.57

Table 20 continued from previous page

MT	Metrics	fr-en	fr-vi	fr-zh	fr-de
Commercial tool					
Google Translate	BLEU	27.82	24.18	24.49	22.38
	BERTSc	0.88	0.86	0.85	0.86
	TER	63.0	67.34	66.27	70.46
	METEOR	0.51	0.46	0.44	0.43
	ChrF	50.04	42.11	22.47	46.46
	ROUGE-1	0.59	0.68	0.1	0.5
	ROUGE-2	0.41	0.46	0.07	0.3
	ROUGE-L	0.56	0.56	0.1	0.47

Table 21: **Full Results: Ground-truth Translation Baselines. French to X** results are reported in this table. Extension of Table 3 in the main paper.

MT	Metrics	de-en	de-vi	de-fr	de-zh
Decoder					
Llama -3.1-8B	BLEU	49.44	40.01	33.45	31.16
	BERTSc	0.95	0.87	0.84	0.81
	TER	44.99	54.45	72.36	62.08
	METEOR	0.76	0.67	0.62	0.54
	ChrF	69.74	57.57	61.36	27.18
	ROUGE-1	0.75	0.78	0.61	0.12
	ROUGE-2	0.57	0.58	0.42	0.09
	ROUGE-L	0.72	0.68	0.57	0.12
Qwen -2.5-7B	BLEU	52.1	43.73	40.72	23.26
	BERTSc	0.96	0.88	0.88	0.79
	TER	39.94	48.63	51.44	63.95
	METEOR	0.77	0.69	0.65	0.44
	ChrF	70.13	59.36	62.75	23.28
	ROUGE-1	0.76	0.8	0.67	0.11
	ROUGE-2	0.58	0.6	0.47	0.09
	ROUGE-L	0.73	0.7	0.63	0.11
Mistral -v0.3-7B	BLEU	36.39	15.68	40.77	21.28
	BERTSc	0.92	0.79	0.86	0.78
	TER	80.09	73.73	53.76	66.07
	METEOR	0.72	0.35	0.63	0.42
	ChrF	66.77	32.48	61.34	22.37
	ROUGE-1	0.64	0.61	0.66	0.09
	ROUGE-2	0.49	0.38	0.47	0.07
	ROUGE-L	0.61	0.48	0.62	0.09
Encoder-decoder					
mBart -large-50	BLEU	41.45	41.12	40.48	30.43
	BERTSc	0.94	0.87	0.87	0.8
	TER	48.34	51.45	50.82	56.79
	METEOR	0.68	0.65	0.63	0.54
	ChrF	60.02	55.8	60.48	26.35
	ROUGE-1	0.68	0.77	0.65	0.13
	ROUGE-2	0.47	0.56	0.46	0.11
	ROUGE-L	0.65	0.67	0.61	0.13
M2M100 -418M	BLEU	44.76	43.83	43.53	30.42
	BERTSc	0.83	0.85	0.88	0.75
	TER	46.56	48.72	47.31	54.14
	METEOR	0.68	0.67	0.67	0.53
	ChrF	62.33	58.19	64.91	27.56
	ROUGE-1	0.67	0.77	0.69	0.13
	ROUGE-2	0.48	0.58	0.5	0.11
	ROUGE-L	0.64	0.67	0.65	0.13
Marian	BLEU	42.74	38.26	39.59	18.11
	BERTSc	0.88	0.87	0.87	0.78
	TER	47.25	52.21	50.87	67.04
	METEOR	0.71	0.64	0.64	0.42
	ChrF	62.6	54.26	61.55	16.79
	ROUGE-1	0.7	0.78	0.66	0.13
	ROUGE-2	0.49	0.55	0.46	0.12
	ROUGE-L	0.67	0.66	0.62	0.13

Table 21 continued from previous page

MT	Metrics	de-en	de-vi	de-fr	de-zh
Commercial tool					
Google Translate	BLEU	40.74	32.69	33.15	31.89
	BERTSc	0.9	0.88	0.88	0.86
	TER	50.5	57.38	58.0	57.64
	METEOR	0.63	0.56	0.55	0.54
	ChrF	60.31	50.13	56.07	28.01
	ROUGE-1	0.68	0.74	0.61	0.12
	ROUGE-2	0.51	0.53	0.43	0.09
	ROUGE-L	0.65	0.63	0.57	0.12

Table 22: **Full Results: Ground-truth Translation Baselines. German to X** results are reported in this table. Extension of Table 3 in the main paper.

MT	Metrics	zh-en	zh-vi	zh-fr	zh-de
Decoder					
Llama -3.1-8B	BLEU	28.21	23.49	18.87	13.07
	BERTSc	0.91	0.79	0.77	0.74
	TER	84.4	96.2	99.18	107.48
	METEOR	0.59	0.55	0.46	0.4
	ChrF	47.42	41.33	41.69	32.09
	ROUGE-1	0.57	0.63	0.46	0.37
	ROUGE-2	0.37	0.44	0.28	0.19
	ROUGE-L	0.53	0.53	0.41	0.33
Qwen -2.5-7B	BLEU	35.63	32.95	24.05	16.95
	BERTSc	0.95	0.85	0.84	0.83
	TER	60.98	67.86	73.13	77.31
	METEOR	0.67	0.63	0.53	0.5
	ChrF	55.3	47.73	47.09	39.65
	ROUGE-1	0.66	0.74	0.55	0.52
	ROUGE-2	0.45	0.53	0.35	0.28
	ROUGE-L	0.62	0.63	0.5	0.48
Mistral -v0.3-7B	BLEU	27.68	10.67	18.46	11.4
	BERTSc	0.93	0.75	0.8	0.76
	TER	72.65	82.73	81.15	90.07
	METEOR	0.59	0.33	0.45	0.39
	ChrF	47.23	24.79	39.83	31.55
	ROUGE-1	0.57	0.56	0.47	0.4
	ROUGE-2	0.36	0.32	0.28	0.2
	ROUGE-L	0.53	0.43	0.42	0.37
Encoder-decoder					
mBart -large-50	BLEU	15.03	22.28	15.7	10.67
	BERTSc	0.9	0.82	0.79	0.77
	TER	76.4	71.86	79.51	85.8
	METEOR	0.42	0.49	0.39	0.36
	ChrF	36.18	39.38	38.64	33.11
	ROUGE-1	0.46	0.68	0.44	0.4
	ROUGE-2	0.23	0.44	0.23	0.18
	ROUGE-L	0.42	0.55	0.39	0.34
M2M100 -418M	BLEU	21.65	27.69	21.88	15.17
	BERTSc	0.76	0.85	0.82	0.82
	TER	72.58	66.12	72.93	78.47
	METEOR	0.52	0.55	0.5	0.46
	ChrF	43.67	43.39	46.57	39.77
	ROUGE-1	0.51	0.67	0.52	0.49
	ROUGE-2	0.28	0.47	0.32	0.25
	ROUGE-L	0.47	0.57	0.47	0.44
Marian	BLEU	11.44	16.14	11.33	6.24
	BERTSc	0.78	0.79	0.77	0.75
	TER	86.03	83.81	89.53	104.78
	METEOR	0.38	0.42	0.33	0.3
	ChrF	32.3	33.04	33.64	29.48
	ROUGE-1	0.39	0.63	0.37	0.33
	ROUGE-2	0.16	0.35	0.17	0.11
	ROUGE-L	0.35	0.48	0.32	0.28

Table 22 continued from previous page

MT	Metrics	zh-en	zh-vi	zh-fr	zh-de
Commercial tool					
Google Translate	BLEU	27.74	30.7	20.71	19.11
	BERTSc	0.88	0.87	0.85	0.77
	TER	74.94	72.9	81.21	83.94
	METEOR	0.52	0.52	0.43	0.45
	ChrF	51.33	49.89	49.91	53.27
	ROUGE-1	0.55	0.69	0.48	0.48
	ROUGE-2	0.34	0.46	0.29	0.26
	ROUGE-L	0.51	0.56	0.43	0.43

Table 23: **Full Results: Ground-truth Translation Baselines. Chinese to X** results are reported in this table. Extension of Table 3 in the main paper.

F.3 Extra Results: Cascaded Speech Translation Baselines

This section presents extra results of cascaded ST baselines for all evaluation metrics, which is a supplement of Table 5 in the main paper. Extra results are shown in Table 24 (English to X), Table 25 (Vietnamese to X), Table 26 (French to X), Table 27 (German to X), and Table 28 (Chinese to X) below.

MT	Metrics	en-vi	en-fr	en-zh	en-de
Decoder					
Llama -3.1-8B	BLEU	43.32	37.92	30.78	31.36
	BERTSc	0.78	0.76	0.73	0.74
	TER	53.71	60.93	66.54	67.05
	METEOR	0.63	0.59	0.51	0.53
	ChrF	57.23	60.63	26.34	56.58
	ROUGE-1	0.76	0.63	0.13	0.57
	ROUGE-2	0.58	0.47	0.11	0.37
	ROUGE-L	0.67	0.59	0.13	0.53
Qwen -2.5-7B	BLEU	43.37	37.34	23.46	28.5
	BERTSc	0.77	0.76	0.74	0.74
	TER	53.52	57.76	64.75	66.7
	METEOR	0.63	0.58	0.44	0.51
	ChrF	57.34	60.21	23.19	54.19
	ROUGE-1	0.76	0.63	0.12	0.55
	ROUGE-2	0.59	0.47	0.1	0.34
	ROUGE-L	0.67	0.6	0.12	0.51
Mistral -v0.3-7B	BLEU	17.72	36.58	20.27	29.9
	BERTSc	0.68	0.74	0.69	0.72
	TER	71.27	59.78	70.94	65.81
	METEOR	0.35	0.55	0.38	0.5
	ChrF	33.42	57.32	20.29	52.89
	ROUGE-1	0.57	0.6	0.09	0.54
	ROUGE-2	0.39	0.45	0.08	0.36
	ROUGE-L	0.49	0.58	0.09	0.51
Encoder-decoder					
mBart -large-50	BLEU	48.0	43.2	35.7	35.07
	BERTSc	0.87	0.86	0.81	0.84
	TER	46.21	51.32	54.44	59.01
	METEOR	0.66	0.62	0.57	0.56
	ChrF	61.1	64.07	31.25	58.51
	ROUGE-1	0.78	0.67	0.15	0.61
	ROUGE-2	0.63	0.52	0.13	0.41
	ROUGE-L	0.71	0.64	0.15	0.57
M2M100 -418M	BLEU	48.21	43.16	36.94	36.55
	BERTSc	0.95	0.92	0.92	0.92
	TER	46.94	51.58	53.79	57.62
	METEOR	0.67	0.63	0.58	0.57
	ChrF	61.29	64.22	32.22	60.13
	ROUGE-1	0.78	0.67	0.16	0.62
	ROUGE-2	0.63	0.52	0.14	0.43
	ROUGE-L	0.71	0.65	0.16	0.58
Marian	BLEU	45.07	40.54	31.17	33.9
	BERTSc	0.87	0.86	0.82	0.84
	TER	49.1	53.41	57.5	59.75
	METEOR	0.65	0.61	0.53	0.56
	ChrF	58.93	62.68	27.33	58.41
	ROUGE-1	0.77	0.66	0.15	0.6
	ROUGE-2	0.61	0.5	0.13	0.4
	ROUGE-L	0.69	0.63	0.15	0.57

Table 23 continued from previous page

MT	Metrics	en-vi	en-fr	en-zh	en-de
----	---------	-------	-------	-------	-------

Table 24: **Extra Results: Cascaded ST Baselines. English to X** results are reported in this table. Supplement of Table 5 in the main paper.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

MT	Metrics	vi-en	vi-fr	vi-zh	vi-de
Decoder					
Llama -3.1-8B	BLEU	14.55	10.29	11.56	7.71
	BERTSc	0.78	0.75	0.73	0.73
	TER	98.39	112.97	131.0	122.33
	METEOR	0.43	0.35	0.4	0.31
	ChrF	41.46	39.22	15.65	37.02
	ROUGE-1	0.44	0.36	0.02	0.32
	ROUGE-2	0.22	0.18	0.01	0.12
	ROUGE-L	0.39	0.31	0.02	0.28
Qwen -2.5-7B	BLEU	13.97	11.66	20.27	8.75
	BERTSc	0.78	0.76	0.78	0.75
	TER	101.68	101.51	74.57	101.85
	METEOR	0.43	0.36	0.43	0.3
	ChrF	41.76	40.73	18.0	36.74
	ROUGE-1	0.43	0.39	0.03	0.34
	ROUGE-2	0.21	0.19	0.02	0.13
	ROUGE-L	0.39	0.33	0.03	0.29
Mistral -v0.3-7B	BLEU	15.86	10.92	17.92	09.03
	BERTSc	0.78	0.75	0.77	0.75
	TER	85.32	97.69	76.97	94.56
	METEOR	0.42	0.33	0.39	0.3
	ChrF	39.64	37.19	16.1	34.82
	ROUGE-1	0.43	0.36	0.02	0.33
	ROUGE-2	0.21	0.17	0.01	0.12
	ROUGE-L	0.39	0.32	0.02	0.29
Encoder-decoder					
mBart -large-50	BLEU	10.17	12.8	16.77	7.23
	BERTSc	0.88	0.76	0.73	0.72
	TER	86.47	87.93	74.4	104.87
	METEOR	0.32	0.33	0.39	0.25
	ChrF	31.78	37.76	15.29	31.98
	ROUGE-1	0.36	0.37	0.02	0.28
	ROUGE-2	0.14	0.18	0.01	0.1
	ROUGE-L	0.32	0.33	0.02	0.24
M2M100 -418M	BLEU	15.64	13.95	16.99	11.1
	BERTSc	0.78	0.77	0.74	0.75
	TER	82.03	86.73	75.18	89.49
	METEOR	0.41	0.36	0.39	0.32
	ChrF	39.68	40.25	15.51	37.54
	ROUGE-1	0.43	0.4	0.03	0.36
	ROUGE-2	0.2	0.2	0.01	0.14
	ROUGE-L	0.39	0.35	0.03	0.31
Marian	BLEU	12.95	11.23	12.09	09.08
	BERTSc	0.77	0.76	0.75	0.74
	TER	85.96	89.57	79.67	91.54
	METEOR	0.37	0.32	0.33	0.28
	ChrF	36.61	37.06	11.93	34.59
	ROUGE-1	0.4	0.36	0.02	0.33
	ROUGE-2	0.17	0.17	0.01	0.12
	ROUGE-L	0.36	0.32	0.02	0.28

Table 24 continued from previous page

MT	Metrics	vi-en	vi-fr	vi-zh	vi-de
----	---------	-------	-------	-------	-------

Table 25: **Extra Results: Cascaded ST Baselines. Vietnamese to X** results are reported in this table. Supplement of Table 5 in the main paper.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

MT	Metrics	fr-en	fr-vi	fr-zh	fr-de
Decoder					
Llama -3.1-8B	BLEU	30.15	25.36	20.28	16.38
	BERTSc	0.82	0.8	0.75	0.74
	TER	65.8	71.69	80.06	99.6
	METEOR	0.52	0.47	0.4	0.4
	ChrF	49.71	41.24	17.84	44.43
	ROUGE-1	0.58	0.67	0.08	0.41
	ROUGE-2	0.39	0.44	0.05	0.22
	ROUGE-L	0.54	0.55	0.08	0.37
Qwen -2.5-7B	BLEU	30.35	25.59	15.33	20.38
	BERTSc	0.81	0.8	0.76	0.78
	TER	69.6	71.42	73.88	76.51
	METEOR	0.52	0.47	0.36	0.4
	ChrF	49.88	41.85	15.7	43.63
	ROUGE-1	0.57	0.67	0.07	0.47
	ROUGE-2	0.39	0.45	0.05	0.26
	ROUGE-L	0.54	0.55	0.07	0.43
Mistral -v0.3-7B	BLEU	29.35	9.2	13.94	18.65
	BERTSc	0.79	0.74	0.76	0.78
	TER	74.28	82.83	74.11	74.55
	METEOR	0.52	0.26	0.34	0.39
	ChrF	49.85	23.69	15.16	41.66
	ROUGE-1	0.55	0.52	0.07	0.46
	ROUGE-2	0.38	0.27	0.05	0.25
	ROUGE-L	0.52	0.4	0.07	0.42
Encoder-decoder					
mBart -large-50	BLEU	23.82	22.86	16.46	17.39
	BERTSc	0.9	0.8	0.72	0.77
	TER	71.43	71.72	74.54	77.56
	METEOR	0.47	0.44	0.36	0.38
	ChrF	44.7	39.14	15.58	40.21
	ROUGE-1	0.52	0.66	0.08	0.44
	ROUGE-2	0.32	0.41	0.05	0.22
	ROUGE-L	0.48	0.52	0.08	0.39
M2M100 -418M	BLEU	25.65	21.88	18.44	19.98
	BERTSc	0.81	0.76	0.75	0.76
	TER	66.39	72.12	69.65	72.77
	METEOR	0.49	0.41	0.39	0.4
	ChrF	47.0	37.81	17.68	43.01
	ROUGE-1	0.55	0.58	0.1	0.47
	ROUGE-2	0.35	0.39	0.07	0.26
	ROUGE-L	0.51	0.48	0.1	0.43
Marian	BLEU	24.03	22.2	11.27	19.14
	BERTSc	0.81	0.8	0.74	0.79
	TER	67.61	73.99	78.35	73.19
	METEOR	0.47	0.44	0.31	0.4
	ChrF	45.53	38.42	11.78	42.62
	ROUGE-1	0.54	0.66	0.09	0.47
	ROUGE-2	0.33	0.41	0.06	0.25
	ROUGE-L	0.5	0.53	0.09	0.43

Table 25 continued from previous page

MT	Metrics	fr-en	fr-vi	fr-zh	fr-de
----	---------	-------	-------	-------	-------

Table 26: **Extra Results: Cascaded ST Baselines. French to X** results are reported in this table. Supplement of Table 5 in the main paper.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

MT	Metrics	de-en	de-vi	de-fr	de-zh
Decoder					
Llama -3.1-8B	BLEU	40.63	33.63	26.97	26.31
	BERTSc	0.86	0.84	0.8	0.79
	TER	56.53	63.07	79.21	68.19
	METEOR	0.63	0.56	0.52	0.47
	ChrF	58.95	49.74	54.06	22.87
	ROUGE-1	0.65	0.73	0.54	0.1
	ROUGE-2	0.48	0.51	0.35	0.07
	ROUGE-L	0.62	0.61	0.49	0.1
Qwen -2.5-7B	BLEU	40.52	34.24	31.45	19.87
	BERTSc	0.86	0.84	0.84	0.79
	TER	55.76	59.16	61.65	68.06
	METEOR	0.63	0.57	0.53	0.4
	ChrF	59.05	50.58	54.36	20.15
	ROUGE-1	0.66	0.74	0.59	0.1
	ROUGE-2	0.47	0.53	0.4	0.08
	ROUGE-L	0.62	0.62	0.54	0.1
Mistral -v0.3-7B	BLEU	28.33	12.38	31.15	17.82
	BERTSc	0.78	0.77	0.83	0.78
	TER	91.13	77.89	63.4	70.11
	METEOR	0.6	0.3	0.52	0.37
	ChrF	57.3	28.44	52.73	18.9
	ROUGE-1	0.55	0.57	0.57	0.09
	ROUGE-2	0.4	0.33	0.39	0.07
	ROUGE-L	0.52	0.44	0.53	0.09
Encoder-decoder					
mBart -large-50	BLEU	31.95	32.62	31.96	25.07
	BERTSc	0.92	0.84	0.84	0.77
	TER	58.6	60.62	60.52	63.61
	METEOR	0.56	0.54	0.52	0.46
	ChrF	51.5	48.21	52.74	22.5
	ROUGE-1	0.6	0.72	0.58	0.11
	ROUGE-2	0.39	0.5	0.39	0.08
	ROUGE-L	0.57	0.6	0.54	0.11
M2M100 -418M	BLEU	33.66	34.7	34.67	24.31
	BERTSc	0.77	0.81	0.85	0.72
	TER	58.65	58.99	58.23	62.56
	METEOR	0.55	0.56	0.56	0.45
	ChrF	52.4	49.93	56.08	23.1
	ROUGE-1	0.58	0.71	0.61	0.1
	ROUGE-2	0.4	0.51	0.43	0.09
	ROUGE-L	0.55	0.6	0.57	0.1
Marian	BLEU	34.09	29.72	30.48	14.79
	BERTSc	0.85	0.83	0.84	0.76
	TER	56.82	61.64	61.09	71.92
	METEOR	0.58	0.53	0.52	0.36
	ChrF	53.91	46.44	53.0	14.45
	ROUGE-1	0.62	0.72	0.58	0.12
	ROUGE-2	0.41	0.48	0.39	0.09
	ROUGE-L	0.59	0.59	0.54	0.12

Table 26 continued from previous page

MT	Metrics	de-en	de-vi	de-fr	de-zh
----	---------	-------	-------	-------	-------

Table 27: **Extra Results: Cascaded ST Baselines. German to X** results are reported in this table. Supplement of Table 5 in the main paper.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

MT	Metrics	zh-en	zh-vi	zh-fr	zh-de
Decoder					
Llama -3.1-8B	BLEU	19.01	17.65	13.84	11.13
	BERTSc	0.78	0.75	0.75	0.74
	TER	101.13	104.68	106.77	109.71
	METEOR	0.48	0.45	0.39	0.35
	ChrF	40.96	34.68	36.97	32.37
	ROUGE-1	0.47	0.58	0.39	0.34
	ROUGE-2	0.27	0.36	0.21	0.16
	ROUGE-L	0.43	0.46	0.34	0.3
Qwen -2.5-7B	BLEU	25.36	26.31	17.84	12.61
	BERTSc	0.82	0.81	0.79	0.78
	TER	79.48	80.4	84.11	91.04
	METEOR	0.53	0.53	0.43	0.4
	ChrF	45.87	41.37	40.88	34.86
	ROUGE-1	0.53	0.67	0.45	0.42
	ROUGE-2	0.32	0.45	0.26	0.2
	ROUGE-L	0.48	0.55	0.4	0.37
Mistral -v0.3-7B	BLEU	20.17	08.01	12.58	7.14
	BERTSc	0.8	0.73	0.76	0.72
	TER	83.87	85.69	87.85	98.87
	METEOR	0.48	0.27	0.37	0.29
	ChrF	40.58	21.71	34.68	25.42
	ROUGE-1	0.48	0.53	0.39	0.3
	ROUGE-2	0.26	0.27	0.2	0.13
	ROUGE-L	0.43	0.4	0.34	0.27
Encoder-decoder					
mBart -large-50	BLEU	11.88	18.04	12.3	9.64
	BERTSc	0.89	0.79	0.76	0.75
	TER	83.71	80.4	89.41	96.86
	METEOR	0.36	0.42	0.34	0.32
	ChrF	32.19	34.81	34.69	30.11
	ROUGE-1	0.4	0.64	0.38	0.35
	ROUGE-2	0.18	0.38	0.19	0.15
	ROUGE-L	0.35	0.49	0.33	0.3
M2M100 -418M	BLEU	16.65	21.83	16.94	13.06
	BERTSc	0.76	0.83	0.79	0.78
	TER	85.02	78.88	84.12	90.04
	METEOR	0.44	0.46	0.41	0.4
	ChrF	39.11	37.38	41.01	37.33
	ROUGE-1	0.44	0.62	0.45	0.42
	ROUGE-2	0.22	0.4	0.25	0.2
	ROUGE-L	0.4	0.49	0.39	0.37
Marian	BLEU	8.5	13.37	8.39	5.73
	BERTSc	0.75	0.77	0.74	0.73
	TER	93.56	86.57	97.14	109.11
	METEOR	0.32	0.36	0.28	0.27
	ChrF	28.26	30.13	30.09	27.23
	ROUGE-1	0.34	0.61	0.33	0.3
	ROUGE-2	0.12	0.31	0.14	0.11
	ROUGE-L	0.3	0.44	0.27	0.25

Table 27 continued from previous page

MT	Metrics	zh-en	zh-vi	zh-fr	zh-de
----	---------	-------	-------	-------	-------

Table 28: **Extra Results: Cascaded ST Baselines. Chinese to X** results are reported in this table. Supplement of Table 5 in the main paper.

All cascaded models use Whisper_{small-mono} as ASR model (Whisper ASR is fine-tuned monolingually - on each source language separately). Its WER on test set is 29.6%, 33.8%, 31.3%, 26.3%, 45.7% for Vietnamese, English, Chinese, German and French respectively.

F.4 Qualitative Results

F.4.1 Vietnamese to German Speech Translation



ASR Transcripts:

1. estrogen và progesterone cái thuốc tránh thai thì áp dụng được
2. hai nhóm nguyên nhân là nguyên nhân thực thể và nguyên nhân chức năng có nguyên nhân
3. rồi là thường nó có đau đớn gì hết trơn ở nhà bình thường nhưng mà đúng khi mà nó gãy xương và một biến chứng gây xương nằm



Human-translated Transcripts:

1. Östrogen- und Progesteron-Verhütungsmittel können angewendet werden
2. Es gibt zwei Hauptursachen: physische Ursachen und funktionelle Ursachen
3. Normalerweise tut es zu Hause nicht weh, aber wenn es bricht, wird es eine Frakturkomplikation haben.

Llama-3.1-8B

1. Und Parotidrot, eine Fertilitätsmedikation, Anwendungen der Wirkungen von Anwendung der Wirkungen von Anwendung der Wirkungen von Anwendung der Wirkungen von Anwendung der Wirkungen von
2. Zwei Gruppen von Ursachen sind Ursachen der Krankheit und Ursachen der Energie.
3. Normalerweise ist es nur schmerzhaft, aber wenn es zu einem Knochenbruch oder einer Knochenbruch-Komplikation kommt, wird es im Bett bleiben. Die häufige Schmerzen, die normalerweise zu Hause sind, aber wenn es zu einem

Mistral-v0.3-7B

1. Und die Parotis, die Schwangerschaftsmittel, dann die Wirkung der Schwangerschaftsmittel.
2. Es gibt zwei Gruppen von Ursachen, nämlich die Ursachen der Krankheit und die Ursachen der Stoffe.
3. Es ist normal, dass es Schmerzen gibt, wenn es zu Hause ist, aber wenn es sich wirklich um eine Knochenbruchverletzung handelt oder ein Komplikation einer Knochenbruchverletzung ist, die zu Hause liegt, dann ist es nicht

Qwen-2.5-7B

1. und die Parotidrot-Tabletten, dann werden die Wirkungen der Tabletten angewendet.
2. Zwei Gruppen von Ursachen sind die Krankheitsursache und die Ursache der Leistung.
3. Normalerweise ist es in der Regel schmerzhaft, aber wenn es sich bricht oder eine Knochenbruchkomplikation eintritt, liegt man im Krankenhaus. ersetzt durch: Normalerweise ist es schmerzhaft, aber wenn es sich bricht oder eine



Figure 33: **Qualitative Results.** Vietnamese to German ST

F.4.2 Vietnamese to English Speech Translation



ASR Transcripts:

1. estrogen và progesterone cái thuốc tránh thai thì áp dụng được
2. hai nhóm nguyên nhân là nguyên nhân thực thể và nguyên nhân chức năng có nguyên nhân
3. rồi là thường nó có đau đến gì hết tron ở nhà bình thường nhưng mà đúng khi mà nó gãy xương và một biến chứng gây xương nằm



Human-translated Transcripts:

1. Estrogen and progesterone contraceptives can be applied
2. There are two main causes: physical causes and functional causes
3. Usually it doesn't hurt at home, but when it fractures, it will have a fracture complication.

Llama-3.1-8B

1. And parotidarot, the pregnancy-inducing drug, then applies the effects of...
2. There are two reasons, one is the reason for the disease, the other is the reason for the energy.
3. It usually hurts to move around at home, but when it breaks a bone or has a complication, it needs to be hospitalized.

Mistral-v0.3-7B

1. And parotidarot, the pregnancy drug, applies the effects of...
2. There are two groups of causes, one is pathogenic factors, the other is pathogenic factors.
3. Usually it hurts in the lower back, but when it's a bone fracture or a complication of a bone fracture, it can lie in bed..

Qwen-2.5-7B

1. And parotidot, the drug for inducing labor, applies its effects.
2. There are two causes of illness: genetic cause and physiological cause.
3. It usually hurts at home, but when it breaks or there's a complication of breaking a bone, it has to be taken to the hospital. It's normal to have pain at home, but if it breaks or there's a complication of breaking a bone, you need to go to the hospital. It's



Figure 34: **Qualitative Results.** Vietnamese to English ST

F.4.3 Vietnamese to French Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. estrogen và progesterone cái thuốc tránh thai thì áp dụng được
2. hai nhóm nguyên nhân là nguyên nhân thực thể và nguyên nhân chức năng có nguyên nhân
3. rồi là thường nó có đau đến gì hết trơn ở nhà bình thường nhưng mà đúng khi mà nó gãy xương và một biến chứng gây xương nằm



Human

Human-translated Transcripts:

1. œstrogènes et progestérone, les contraceptifs peuvent être appliqués
2. Il existe deux causes principales : les causes physiques et les causes fonctionnelles
3. Habituellement, il n'y a aucune douleur à la maison, mais quand il se fracture, il aura une complication de fracture.

Llama-3.1-8B

1. Et la parotidarot, le médicament de stérilité, est appliqué pour ses effets.
2. Les deux groupes de causes sont la cause de la maladie et la cause de la fatigue.
3. Généralement, il n'y a pas de douleur, mais lorsque la fracture se produit, il est généralement hospitalisé. La fracture peut entraîner diverses complications, telles que l'immobilité. Lorsque la fracture se produit, la douleur peut être

Mistral-v0.3-7B

1. Et pour la parotidarot, le médicament pour la grossesse, appliquer les effets de...
2. Il y a deux groupes de causes, les causes pathologiques et les causes chimiques.
3. Il y a souvent des douleurs, mais elles sont généralement bien tolérées à la maison, mais lorsqu'il y a une fracture ou une complication d'une fracture, elle se produit à la maison.

Qwen-2.5-7B

1. et la glande parotide, les médicaments pour le traitement de l'insuffisance thyroïdienne, ont des effets sur la grossesse.
2. Les deux causes sont les causes de la maladie et les causes physiologiques.
3. Il est généralement douloureux et glissant à la maison, mais lorsque le os se casse ou qu'il y a une complication de fracture, il faut aller au lit. Il faut aller au lit. C'est-à-dire que la douleur devient insupportable



Robot

Figure 35: **Qualitative Results.** Vietnamese to French ST

F.4.4 Vietnamese to Chinese Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. estrogen và progesterone cái thuốc tránh thai thì áp dụng được
2. hai nhóm nguyên nhân là nguyên nhân thực thể và nguyên nhân chức năng có nguyên nhân
3. rồi là thường nó có đau đến gì hết trơn ở nhà bình thường nhưng mà đúng khi mà nó gây xương và một biến chứng gây xương nằm



Human-translated Transcripts:

1. 雌激素和孕激素避孕药可以应用
2. 有两个主要原因: 实体原因和功能原因
3. 通常它在家没什么疼痛, 但它骨折时会发生骨折并发症。

Llama-3.1-8B

1. 和甲状腺刺激剂, 生育药, 应用其的作用。
2. 有两种原因, 一种是病因, 一种是能量因素。这两种原因是导致疾病的主要原因。这两种原因是导致
3. 它通常在家里会有疼痛, 但正是在骨骼破裂或是骨骼破裂并发症时才会出现。这时它就会使得你无法

Mistral-v0.3-7B

1. 并且喉咙腺泌液, 妊娠药物应用其作用的作用。
2. 两个原因是病原和化学原因。
3. 通常, 它会在家里感到疼痛, 就像一般情况一样, 但是当它患上骨

Qwen-2.5-7B

1. 和腰门的妊娠药物, 应用其治疗作用。
2. 两个原因是疾病的原因和动能的原因。
3. 通常它会在家里感到疼痛, 但是当骨折或发生骨折并发症时, 就会要求医



Figure 36: **Qualitative Results.** Vietnamese to Chinese ST

F.4.5 English to Vietnamese Speech Translation



ASR Transcripts:

1. medical history and your general health, so thank you for having me on today. So, DVT and PE, it is a very common condition actually and most people are not aware of that. And I usually joke around
2. because affects everyone differently. So it can cause tumors throughout the body, on nerves terminations,



Human-translated Transcripts:

1. Hãy cho tôi biết về tiền sử bệnh và sức khỏe tổng thể của bạn, cảm ơn bạn đã cho tôi gặp hôm nay. Vì vậy, DVT và PE, thực sự đó là một tình trạng rất phổ biến và hầu hết mọi người không biết điều đó. Và tôi thường hay nói đùa
2. Bởi vì nó ảnh hưởng đến mỗi người khác nhau. Vì vậy, nó có thể gây ra khối u trên toàn bộ cơ thể, trên các đầu dây thần kinh.

Llama-3.1-8B

1. Lịch sử y tế và sức khỏe tổng thể của bạn. Vì vậy, cảm ơn bạn đã cho tôi được tham gia hôm nay. Vì vậy, DVTMP, đó là một tình trạng rất phổ biến thực sự và hầu hết mọi người không biết điều đó. Và tôi nghĩ, tôi thường đùa
2. Bởi vì nó ảnh hưởng đến mọi người theo cách khác nhau. Vì vậy, nó có thể gây ra các khối u ở mọi nơi trong cơ thể, đặc biệt là ở các đầu dây thần kinh.

Mistral-v0.3-7B

1. Lịch sử bệnh và sức khỏe chung của bạn. Vì vậy, cảm ơn bạn đã đưa tôi đến ngày hôm nay.
2. bởi vì nó ảnh hưởng mọi người khác nhau. Vì vậy, nó có thể gây ra khối u trên toàn bộ cơ thể ở cả

Qwen-2.5-7B

1. Lịch sử y tế và tình trạng sức khỏe tổng quát của bạn. Vì vậy, cảm ơn vì đã mời tôi tham gia hôm nay. Vì vậy, DVTMP, thực ra nó là một tình trạng rất phổ biến và hầu hết mọi người không nhận thức được điều đó. Và tôi nghĩ
2. Bởi vì nó ảnh hưởng đến mọi người một cách khác nhau. Vì vậy, nó có thể gây ra khối u ở khắp cơ thể trên các điểm kết thúc của thần kinh.



Figure 37: **Qualitative Results.** English to Vietnamese ST

F.4.6 English to German Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. medical history and your general health, so thank you for having me on today. So, DVT and PE, it is a very common condition actually and most people are not aware of that. And I usually joke around
2. because affects everyone differently. So it can cause tumors throughout the body, on nerves terminations,



Human-translated Transcripts:

1. Erzählen Sie mir bitte von Ihrer Krankengeschichte und Ihrem allgemeinen Gesundheitszustand, danke, dass Sie mich heute empfangen. Also, TVT und PE, das ist tatsächlich eine sehr häufige Erkrankung, und die meisten Menschen sind sich dessen nicht bewusst. Und ich scherze normalerweise
2. Denn es betrifft jeden unterschiedlich. So kann es Tumoren im ganzen Körper verursachen, an Nervenenden.

Llama-3.1-8B

1. Ihre medizinische Vorgeschichte und Ihre allgemeine Gesundheit. Also, vielen Dank, dass Sie mich heute haben. DVTMP, es ist eine sehr häufige Erkrankung, tatsächlich, und die meisten Menschen wissen nicht darüber. Und ich denke,
2. Denn es wirkt auf jeden unterschiedlich. Es kann also Tumoren überall im Körper auf Nervenenden verursachen.

Mistral-v0.3-7B

1. Ihre medizinische Vorgeschichte und Ihr allgemeiner Gesundheitszustand. Also danke, dass Sie mich heute hier haben. Also DVTMP, es ist eine sehr häufige Erkrankung, und die meisten Menschen sind davon nicht bew
2. Denn es wirkt sich unterschiedlich auf jeden aus. Es kann also Tumoren im gesamten Körper an Nervenendigungen verursachen, ähnlich wie bei einem Neurofibromatose-Typ 1.

Qwen-2.5-7B

1. Medizinische Geschichte und Ihre allgemeine Gesundheit. Also danke für die Einladung heute. DVTMP ist also tatsächlich eine sehr häufige Erkrankung, und die meisten Menschen sind nicht darüber im Bilde. Und ich denke, ich mache das oft ein bisschen spa
2. Weil es jeder anders beeinflusst. Es kann also Tumoren in allen Teilen des Körpers verursachen, auf Nervenenden.



Figure 38: **Qualitative Results.** English to German ST

F.4.7 English to French Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. medical history and your general health, so thank you for having me on today. So, DVT and PE, it is a very common condition actually and most people are not aware of that. And I usually joke around
2. because affects everyone differently. So it can cause tumors throughout the body, on nerves terminations,



Human-translated Transcripts:

1. Parlez-moi de vos antécédents médicaux et de votre état de santé général, merci de me recevoir aujourd'hui. Donc, la TVP et l'EP, c'est en fait une condition très courante et la plupart des gens ne le savent pas. Et j'ai l'habitude de faire des blagues
2. Car cela affecte chacun différemment. Cela peut donc causer des tumeurs dans tout le corps, sur les terminaisons nerveuses.

Llama-3.1-8B

1. Votre histoire médicale et votre santé générale. Donc, merci de m'avoir invité aujourd'hui. Le DVTMP, c'est une condition très courante en fait et la plupart des gens ne sont pas au courant. Et je pense, je plaisante souvent là-dessus
2. Parce qu'il affecte tout le monde différemment. Donc, cela peut provoquer des tumeurs dans tout le corps sur les terminaisons nerveuses, 42. 42. 42. 42. 42. 42. 42. 42.

Mistral-v0.3-7B

1. Votre histoire médicale et votre état général. Alors merci de m'avoir invité aujourd'hui. Donc DVTMP, c'est une condition très courante en fait, et la plupart des gens ne s'en rendent pas
2. car il affecte chacun différemment. Il peut donc provoquer des tumeurs dans tout le corps sur les terminaisons nerveuses.

Qwen-2.5-7B

1. Histoire médicale et état général de santé. Donc, merci de m'avoir invité aujourd'hui. DVTMP, c'est en fait une maladie très courante et la plupart des gens ne sont pas au courant de cela. Et je pense que, quand
2. Parce qu'il affecte tout le monde différemment. Il peut donc causer des tumeurs dans tout le corps, sur les terminaisons nerveuses.



Figure 39: **Qualitative Results.** English to French ST

F.4.8 English to Chinese Speech Translation



ASR Transcripts:

1. medical history and your general health, so thank you for having me on today. So, DVT and PE, it is a very common condition actually and most people are not aware of that. And I usually joke around
2. because affects everyone differently. So it can cause tumors throughout the body, on nerves terminations,



Human-translated Transcripts:

1. 请您告诉我一下您的病史和整体健康状况,感谢您今天能来。深静脉血栓形成和肺栓塞实际上是一种非常常见的疾病,大多数人对此并不了解。我通常会开玩笑说
2. 因为对每个人影响都不一样。所以它会导致全身肿瘤,神经末梢肿瘤。

Llama-3.1-8B

1. 你的医疗历史和你的普遍健康状况。所以,感谢你今天接受我。因此, DVTMP, 实际上是一种非常常见的
2. 因为它对每个人的影响不同。因此, 它可能导致全身各部位的神经终端肿瘤。1950年代, 人类发现

Mistral-v0.3-7B

1. 您的医疗史和您的一般健康。所以感谢您今天邀请我。所以DVTMP, 它实
2. 因为它对每个人都有不同的影响。因此, 它可能导致全身各处神经结束处

Qwen-2.5-7B

1. 医疗历史和您的一般健康情况。所以, 谢谢你今天让我来。DVTMP实际上是
2. 因为它对每个人的影响都不同。所以它可能导致全身神经终端处出现肿



Figure 40: **Qualitative Results.** English to Chinese ST

F.4.9 German to Vietnamese Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. kein Asthma, wenn man mal stehen bleiben muss? Zumindest nicht so wie Sie es beschreiben. Und dass Ihre Beine so geschwollen sind ... Das wundert mich ja auch. Ich dachte,
2. die hohen Blutfettwerte. Aber das liegt bei uns ja beinah in der ganzen Familie, mütterlicherseits. Wissen Sie, ob eine spezielle Fettstoffwechselstörung bei dieser familiären Häufung bekannt ist? Ich weiß nur, dass die Schwester



Human-translated Transcripts:

1. Không bị hen suyễn, ngay cả khi bạn phải dừng lại? Ít nhất là không giống như cách bạn mô tả. Và chân của bạn bị sưng lên như vậy... Điều đó cũng làm tôi ngạc nhiên. Tôi nghĩ
2. Chỉ số mỡ máu của tôi cao. Nhưng đó là vấn đề gần như phổ biến trong cả gia đình, từ phía mẹ tôi. Ông có biết liệu có rối loạn chuyển hóa lipid đặc biệt nào được biết đến với sự tập trung gia đình này không? Tôi chỉ biết rằng chị gái tôi

Llama-3.1-8B

1. Không có hen suyễn, nếu bạn phải dừng lại. Tôi thiếu là không giống như bạn mô tả. Và rằng chân của bạn rất sưng... Điều này cũng khiến tôi bối rối. Tôi nghĩ,
2. Giá trị mỡ trong máu cao. Nhưng điều này gần như phổ biến trong cả gia đình của tôi. Bạn biết không, có một rối loạn chuyển hóa chất béo cụ thể nào trong gia đình này được biết đến là có xu hướng gia tăng? Tôi chỉ biết rằng chị gái... của

Mistral-v0.3-7B

1. Không phải ca hô hấp, nếu bạn phải đứng dậy một lúc. Ít nhất là như bạn mô tả. Và vì tay bạn
2. Các mức mỡ trong máu cao. Nhưng điều đó thực sự gần gấp đôi khảo với gia đình của chúng t

Qwen-2.5-7B

1. Anh không bị hen suyễn, ít nhất là không như bạn mô tả. Và những chân của bạn bị sưng như vậy... Tôi cũng ngạc nhiên về điều đó. Tôi nghĩ rằng...
2. Một lượng mỡ máu cao. Nhưng gần như cả gia đình mẹ tôi đều bị. Anh có biết liệu rối loạn chuyển hóa mỡ máu đặc biệt nào đã được báo cáo trong gia đình này không? Tôi chỉ biết rằng chị gái của tôi ước tính... (không rõ) 10%



Figure 41: **Qualitative Results.** German to Vietnamese ST

F.4.10 German to English Speech Translation



ASR Transcripts:

1. kein Asthma, wenn man mal stehen bleiben muss? Zumindest nicht so wie Sie es beschreiben. Und dass Ihre Beine so geschwollen sind ... Das wundert mich ja auch. Ich dachte,
2. die hohen Blutfettwerte. Aber das liegt bei uns ja beinah in der ganzen Familie, mütterlicherseits. Wissen Sie, ob eine spezielle Fettstoffwechselstörung bei dieser familiären Häufung bekannt ist? Ich weiß nur, dass die Schwester



Human-translated Transcripts:

1. No asthma, even if you have to stop? At least not the way you describe it. And your legs are so swollen... That surprises me too. I thought
2. My blood fat levels are high. But that's almost in the whole family, on my mother's side. Do you know if there is a specific lipid metabolism disorder known for this familial aggregation? I only know that my sister

Llama-3.1-8B

1. No asthma, if you have to stop. At least not as you describe it. And that your legs are so swollen... That surprises me too. I thought, I thought, I thought...
2. High cholesterol levels. But that's almost a family trait on my mother's side. You know if there's a specific lipid metabolism disorder known to be prevalent in this family? I only know that my sister has...

Mistral-v0.3-7B

1. No asthma, if you have to stand still. At least not like you describe. And that your legs are so swollen... That surprises me too. I thought, ...
2. High cholesterol levels. But that's almost the whole family on the mother's side. Do you know if there is a specific fat metabolism disorder in this family? I only know that the sister ... ?

Qwen-2.5-7B

1. No asthma, if you have to stop. At least not like you describe it. And that your legs are swollen... That surprises me too. I thought,
2. High blood fat levels. But that's almost the whole family on my mother's side. Do you know if there's a specific lipid metabolism disorder known in this family? I only know that the sister...



Figure 42: **Qualitative Results.** German to English ST

F.4.11 German to French Speech Translation



ASR Transcripts:
1. kein Asthma, wenn man mal stehen bleiben muss? Zumindest nicht so wie Sie es beschreiben.
Und dass Ihre Beine so geschwollen sind ... Das wundert mich ja auch. Ich dachte,
2. die hohen Blutfettwerte. Aber das liegt bei uns ja beinah in der ganzen Familie, mütterlicherseits.
Wissen Sie, ob eine spezielle Fettstoffwechselstörung bei dieser familiären Häufung bekannt ist? Ich
weiß nur, dass die Schwester



Human-translated Transcripts:
1. Pas d'asthme, même si vous devez vous arrêter ? Au moins pas comme vous le décrivez. Et vos
jambes sont si gonflées... Cela me surprend aussi. Je pensais
2. Mes taux de lipides sanguins sont élevés. Mais c'est presque le cas dans toute la famille, du côté
maternel. Savez-vous s'il existe un trouble spécifique du métabolisme des lipides connu pour cette
agrégation familiale ? Je sais juste que ma sœur

Llama-3.1-8B
1. Pas d'asthme, si vous devez vous arrêter un instant. En tout cas, pas comme vous le décrivez. Et
que vos jambes soient si gonflées... Ça me surprend aussi. Je pensais, Je ne comprends pas
pourquoi vous êtes aussi
2. Des taux de lipides élevés. Mais c'est presque une tradition familiale de ce côté maternel.
Savez-vous si une stôrage lipidométrique spécifique est connue dans cette famille? Je ne sais que le
fait que la sœur... La traduction est incom

Mistral-v0.3-7B
1. Pas d'asthme, quand on doit rester debout. Au moins pas comme vous le décrivez. Et votre jambes
sont tellement gonflées... Je suis aussi surpris. Je pensais,
2. des taux élevés de lipides. Mais cela se passe dans la famille, presque partout. Savez-vous si une
telle maladie héréditaire du métabolisme lipidique est connue dans cette famille ? Je ne sais que la
sœur a

Qwen-2.5-7B
1. Pas d'asthme, même si vous devez vous arrêter. Au moins pas comme vous le décrivez. Et que
vos jambes soient si gonflées... Ça m'étonne aussi. Je pensais,
2. Des taux de lipides élevés. Mais cela se produit pratiquement dans toute notre famille maternelle.
Savez-vous si il y a une maladie métabolique du lipide connue dans cette famille ? Je ne sais que ça,
c'est sa sœur... (Le médecin semble



Figure 43: **Qualitative Results.** German to French ST

F.4.12 German to Chinese Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. kein Asthma, wenn man mal stehen bleiben muss? Zumindest nicht so wie Sie es beschreiben. Und dass Ihre Beine so geschwollen sind ... Das wundert mich ja auch. Ich dachte,
2. die hohen Blutfettwerte. Aber das liegt bei uns ja beinah in der ganzen Familie, mütterlicherseits. Wissen Sie, ob eine spezielle Fettstoffwechselstörung bei dieser familiären Häufung bekannt ist? Ich weiß nur, dass die Schwester



Human-translated Transcripts:

1. 没有哮喘，即使你必须停下来？至少不像你描述的那样。而且你的腿肿了……这也很奇怪。我以为
2. 我的血脂很高。但这是我们家族，尤其是母系这边普遍的问题。您知道这种家族聚集性是否与某种特定的脂代谢紊乱有关吗？我只知道我的姐姐

Llama-3.1-8B

1. 如果你停下来，你不会哮喘。至少不像你描述的那样。而且你的腿都肿了……我也很奇怪。我以前认为，
2. 高脂血症。但是，这在我的母亲一侧家庭里几乎是普遍的。您知道这个家庭中是否存在特定的脂质代谢

Mistral-v0.3-7B

1. 没有哮喘，只是站不住了。至少不像您所说的那样。而且您的腿都
2. 血脂水平较高。但这与我们家几乎所有成员都是母亲那边。您知道这家人

Qwen-2.5-7B

1. 没有哮喘，即使需要站立。至少不会像你描述的那样。而且你的腿这
2. 他有高血脂，但这是在我们家里几乎是母亲一边的所有人。您知道这个家庭中是否



Figure 44: **Qualitative Results.** German to Chinese ST

F.4.13 French to Vietnamese Speech Translation



ASR Transcripts:

1. fait que vous avez pris énormément tranquilisant. vous pouvez pas rester dans un lit sans surveillance. je veux pas rester. je veux plus vivre. oui après je [...] un impossible pour
2. maladie est extrêmement rare et que la mutation est très peu présente dans la population. mais pourtant on estime qu' environ 4% de la population mondiale est



Human-translated Transcripts:

1. Đó là do bạn đã uống rất nhiều thuốc an thần. Bạn không thể ở lại giường mà không có người giám sát. Tôi không muốn ở lại. Tôi không muốn sống nữa. Vâng, sau khi tôi ... một điều bất khả thi ...
2. Bệnh rất hiếm gặp và đột biến này rất ít phổ biến trong dân số. Tuy nhiên, ước tính khoảng 4% dân số thế giới mang đột biến này.

Llama-3.1-8B

1. Bởi vì việc bạn đã sử dụng một lượng thuốc an thần lớn, bạn không thể ở trong giường mà không được giám sát.
2. Từ đó, chúng ta có thể suy luận rằng bệnh này rất hiếm và đột biến rất ít trong dân số. Tuy nhiên, chúng ta ước tính rằng cứ một triệu người thì có khoảng 4 người mắc bệnh này.

Mistral-v0.3-7B

1. Bởi vì việc bạn đã uống thuốc tỉnh mạnh như vậy, bạn không thể ở trong một giường không theo dõi.
2. Được từ đó, chúng ta có thể suy luận rằng bệnh này rất quen với, và đột kích thước này rất ít xuấ

Qwen-2.5-7B

1. Bởi vì việc bạn đã sử dụng một lượng lớn thuốc an thần, bạn không thể nằm trong giường mà không có sự giám sát.
2. Nhận ra rằng bệnh này rất hiếm và biến đổi gen rất ít trong dân số. Nhưng mà, người ta ước tính rằng khoảng 4% dân số bị bệnh này. Ừm, vậy thì chúng ta có thể kết luận gì? Vậy thì chúng ta có thể kết luận rằng bệnh này



Figure 45: **Qualitative Results.** French to Vietnamese ST

F.4.14 French to German Speech Translation



ASR Transcripts:

1. fait que vous ayez pris énormément tranquilisant. vous pouvez pas rester dans un lit sans surveillance. je veux pas rester. je veux plus vivre. oui après je [...] un impossible pour
2. maladie est extrêmement rare et que la mutation est très peu présente dans la population. mais pourtant on estime qu' environ 4% de la population mondiale est



Human-translated Transcripts:

1. Das liegt daran, dass Sie sehr viele Beruhigungsmittel genommen haben. Sie können nicht unbeaufsichtigt im Bett bleiben. Ich möchte nicht bleiben. Ich möchte nicht mehr leben. Ja, danach... ein Unmöglich...
2. Die Krankheit ist extrem selten und die Mutation kommt in der Bevölkerung nur sehr selten vor. Dennoch wird geschätzt, dass etwa 4% der Weltbevölkerung die Mutation trägt.

Llama-3.1-8B

1. Weil Sie eine enorme Menge Beruhigungsmittel genommen haben, können Sie nicht in einem Bett bleiben, ohne überwacht zu werden. Sie müssen in einer Einrichtung untergebracht werden, wo Sie engmaschig überwacht werden können. Dies ist für Ihre Sicherheit un
2. Daraus können wir schließen, dass diese Krankheit sehr selten ist und die Mutation sehr selten in der Bevölkerung vorkommt. Dennoch schätzen wir, dass etwa 4% der Bevölkerung... 4% der Bevölkerung.

Mistral-v0.3-7B

1. Denn wenn Sie viel Schlafmittel eingenommen haben, können Sie nicht ohne Überwachung in einem Bett bleiben.
2. Man kann schließen, dass diese Krankheit extrem selten ist und die Mutation sehr selten in der Bevölkerung vorkommt. Doch trotzdem schätzt man, dass in etwa 4% der Bevölkerung diese Mutation vorhanden ist.

Qwen-2.5-7B

1. Denn die Tatsache, dass Sie viele Beruhigungsmittel genommen haben, können Sie nicht ohne Überwachung im Bett bleiben.
2. Es lässt sich ableiten, dass diese Krankheit äußerst selten ist und die Mutation in der Bevölkerung sehr selten vorkommt. Aber trotzdem schätzt man, dass etwa 4 % der Bevölkerung... (der Rest der Aussage ist unlesbar



Figure 46: **Qualitative Results.** French to German ST

F.4.15 French to English Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. fait que vous ayez pris énormément tranquilisant. vous pouvez pas rester dans un lit sans surveillance. je veux pas rester. je veux plus vivre. oui après je [...] un impossible pour
2. maladie est extrêmement rare et que la mutation est très peu présente dans la population. mais pourtant on estime qu' environ 4% de la population mondiale est



Human-translated Transcripts:

1. It's because you took a lot of tranquilizers. You can't stay in bed unsupervised. I don't want to stay. I don't want to live anymore. Yes, after I... an impossible...
2. The disease is extremely rare and the mutation is very uncommon in the population. However, it is estimated that about 4% of the world's population carries the mutation.

Llama-3.1-8B

1. Because the fact that you have taken a lot of tranquilizers you can't stay in a bed without supervision.
2. To deduce that this disease is extremely rare and that the mutation is very rare in the population. But yet we estimate that about 1 in 4% of the population...

Mistral-v0.3-7B

1. Because the fact that you have taken a large amount of tranquilizers, you can't stay in a bed without supervision.
2. Therefore, it can be deduced that this disease is extremely rare and that the mutation is very rare in the population. However, it is estimated that about 4% of the population... -Translated from French by a model.

Qwen-2.5-7B

1. Because you have taken a lot of tranquilizers, you can't stay in bed without supervision.
2. It can be deduced that this disease is extremely rare and that the mutation is very rarely present in the population. But yet it is estimated that about 4% of the population... 4% of the population. 4% of the population. 4% of the population. 4% of the population.



Figure 47: **Qualitative Results.** French to English ST

F.4.16 French to Chinese Speech Translation



ASR Transcripts:
1. fait que vous ayez pris énormément tranquilisant. vous pouvez pas rester dans un lit sans surveillance. je veux pas rester. je veux plus vivre. oui après je [...] un impossible pour
2. les médecins soupçonnent cette fois non pas une bronchiolite mais une épiglottite, c'est à dire une laryngite aiguë qui bloque la respiration.



Human-translated Transcripts:
1. 导致你服用了大量的镇静剂。你不能独自躺在床上。我不想留在这里。我不想再活了。是的，之后我.....不可能.....
2. 这次医生怀疑的不是支气管炎，而是会厌炎，也就是急性喉炎，会导致呼吸道阻塞。

Llama-3.1-8B
1. 因为你用了很多镇静剂，你不能在床上没有监护。因为你用了很多镇静剂，你不能在床上没有监护。因为
2. 这次医生担忧不是支气管炎，而是喉咽病，也就是说喉咽的一个小部分发炎了。他们还担忧他们的

Mistral-v0.3-7B
1. 因为你服用了大量的安眠剂，你不能在没有监护的情况下卧床。
2. 这次医生猜测不是支气管炎，而是粒喉炎，也就是粒喉

Qwen-2.5-7B
1. 因为你吃了很多安神药，所以你不能在没有监护的情况下卧床休养。
2. 医生这次怀疑的不是支气管炎，而是喉咽炎，就是说.....



Figure 48: **Qualitative Results.** French to Chinese ST

F.4.17 Chinese to Vietnamese Speech Translation



ASR Transcripts:

1. 如果手术的范围大的话 它一定会有照明上的死角。那无影灯它其实是组合式的光源。它可以在四面八方的灯聚焦的状况之下
2. 这个胰腺癌呢 有20%是genetics 80%是sporadic 就是说 她是因为有些genetic mutation 造成的。



Human-translated Transcripts:

1. Đèn phẫu thuật không hắt bóng (đèn Led phẫu thuật) thực chất là một nguồn sáng kết hợp. Nó có thể giảm thiểu bóng của bàn tay chúng ta cầm dụng cụ hoặc chính dụng cụ đó trong quá trình phẫu thuật
2. Ung thư tuyến tụy này, 20% là do di truyền, 80% là tự phát, nghĩa là nó được gây ra bởi một số đột biến gen.

Llama-3.1-8B

1. Nếu phạm vi phẫu thuật rộng, nó chắc chắn sẽ có bóng tối. Và đèn không có bóng là một nguồn ánh sáng kết hợp.
2. Ung thư tuyến tụy này, 20% do di truyền. 80% là do đột biến di truyền, nghĩa là nó xuất hiện ngẫu nhiên. Nó được tạo ra bởi một số đột biến di truyền. 1 0 0 0 0

Mistral-v0.3-7B

1. Nếu khu vực phẫu thuật lớn, nó chắc chắn sẽ có dây chằng chụp sáng. Cái gì đó này, nó thực sự
2. Cái tuyến tụy này, nó có hơn 20 genetici. 80% là do tá trợ hoặc genetici. Điều này có nghĩa là nó là do mẹ

Qwen-2.5-7B

1. Nếu phạm vi phẫu thuật lớn, nó chắc chắn sẽ có một khu vực chết về ánh sáng. Vì vậy, đèn vô trùng thực chất là một nguồn sáng kết hợp.
2. Nhiều ung thư tuyến tụy có liên quan đến gen, chiếm 20%. 80% là do gen ngẫu nhiên, nghĩa là chúng được gây ra bởi một số đột biến gen. 80% là do gen ngẫu nhiên, nghĩa là chúng được gây ra bởi



Figure 49: **Qualitative Results.** Chinese to Vietnamese ST

F.4.18 Chinese to English Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. 如果手术的范围大的话 它一定会有照明上的死角。那无影灯它其实是组合式的光源。它可以在四面八方的灯聚焦的状况之下
2. 这个胰腺癌呢 有20%是genetics 80%是sporadic 就是说 她是因为有些genetic mutation 造成的。



Human

Human-translated Transcripts:

1. The shadowless surgical lamp (LED surgical light) is actually a combination light source. It can minimize the shadows cast by our hands holding instruments or the instruments themselves during surgery.
2. This pancreatic cancer, 20% is genetics, 80% is sporadic, meaning it's caused by some genetic mutations.

Llama-3.1-8B

1. If the surgical scope is large, it will definitely have a dark spot in the lighting. The xenon lamp is actually a combination of light sources.
2. This pancreatic cancer, it's twenty percent genetic. Eighty percent is sporadic, meaning it's caused by genetic mutations. It's not inherited, it's not caused by anything we can control. It just happens. And then there are some cases where it's caused by a genetic mutation. It

Mistral-v0.3-7B

1. If the surgical field is large, it will definitely have surgical adhesive. The headlamp, it is actually a composite light source..
2. This pancreatic cancer, it has about 20,000 genetic mutations. 80,000 means it's spread, it means it's spread because of some genetic mutations.

Qwen-2.5-7B

1. If the scope of surgery is large, it will definitely have dead space in terms of lighting. An operating room light is actually a combination of light sources.
2. This pancreatic cancer has 20% genetic causes. 80% is sporadic, meaning it is caused by some genetic mutations.



Figure 50: **Qualitative Results.** Chinese to English ST

F.4.19 Chinese to French Speech Translation



ASR Transcripts:

1. 如果手术的范围大的话 它一定会有照明上的死角。那无影灯它其实是组合式的光源。它可以在四面八方的灯聚焦的状况之下
2. 这个胰腺癌呢 有20%是genetics 80%是sporadic 就是说 她是因为有些genetic mutation 造成的。



Human-translated Transcripts:

1. il y aura forcément des angles morts d'éclairage. La lampe sans ombre est en fait une combinaison de sources lumineuses. Elle peut concentrer la lumière de toutes parts, réduire au minimum les ombres de nos mains tenant les instruments pendant la chirurgie, ou les ombres des instruments eux-mêmes.
2. Ce cancer du pancréas, 20% est génétique, 80% est sporadique, ce qui signifie qu'il est causé par des mutations génétiques.

Llama-3.1-8B

1. Si la portée de la chirurgie est large, il y aura certainement des zones d'ombre. Le loup de la lumière est en fait un éclairage composé. 1 0 0 0 0 0 0 0 0 0
2. Ce cancer du pancréas, il est de type génétique, 20%. 80% sont sporadiques, ce qui signifie qu'ils sont causés par des mutations génétiques. 1 0 0 。 1 0 0 。 1

Mistral-v0.3-7B

1. Si l'aire de la chirurgie est large, il y aura certainement un adhésif lumineux. Un flash, c'est en fait une source de lumière combinée.
2. Ce cancer du pancréas a 20 gènes. 80% signifie qu'il s'agit d'un cancer sporadique, c'est-à-dire qu'il est causé par certaines mutations génétiques.

Qwen-2.5-7B

1. Si la portée de l'opération est grande, il y aura nécessairement un endroit où la lumière sera faible. L'éclairage chirurgical, en fait, est une source lumineuse composée.
2. Cancer du pancréas est responsable de 20% des cas génétiques. 80% des cas sont dits sporadiques, ce qui signifie qu'ils sont causés par des mutations génétiques.



Figure 51: **Qualitative Results.** Chinese to French ST

F.4.20 Chinese to German Speech Translation



Automatic Speech Recognition

ASR Transcripts:

1. 我们新安装的那个心门是在这里。然后我们看一下有没有倒漏的问题。有没有狭窄的那个情况。
2. 所以一切都做得非常好 没有什么大碍。手术之后感觉就是比较舒畅一点。



Human

Human-translated Transcripts:

1. Die neue Herzklappe, die wir installiert haben, ist hier. Lassen Sie uns prüfen, ob es Lecks oder Verengungen gibt.
2. Alles ist also sehr gut verlaufen, es gibt nichts Ernstes. Ich fühle mich nach der Operation viel wohler.

Llama-3.1-8B

1. Unsere Herzklappe, die wir zuversichtlich sind, befindet sich hier. Schauen wir nach, ob es eine Wegstörung gibt, und ob es eine Engstelle gibt. Es gibt keine Störung und keine Enge. Es gibt keine Störung und keine
2. Also alles war sehr gut, es gab nichts zu kritisieren. Nach der Massage fühlte ich mich etwas lockerer. 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Mistral-v0.3-7B

1. Diese, die die wir im Herzen installiert, sind die Tore, die wir hier sehen. Schauen Sie mal, ob es einen Weg zur Frage gibt. Es gibt auch keine engefachte Situation.
2. Daher aus diesem Grunde ist alles, was es immer besser macht, nichts ein Wort. Nachdem ich die Handgebärde abgeschlossen habe, fühlt es sich etwas locker.

Qwen-2.5-7B

1. Die Herzklappe, die wir einsetzen, befindet sich hier. Lassen Sie uns nachsehen, ob es Probleme mit den Straßen gibt, auch keine Enge. Es gibt keine enge Situation.
2. Also ist alles sehr gut gemacht worden, es gab keine Probleme. Nach der Gestik fühlte ich mich etwas entspannter.



Figure 52: **Qualitative Results.** Chinese to German ST

G Ethical Statements

G.1 Fair Use

G.1.1 Fair Use Considerations

The concept of fair use is critical when creating, curating, and utilizing medical ST datasets (Sobel, 2017). Fair use provides a legal framework allowing limited use of copyrighted material without obtaining permission from the copyright holders (Yankwich, 1954). However, its application is context-dependent and often requires careful analysis of specific factors to ensure compliance with the law, particularly when dealing with sensitive domains like healthcare and multilingual communication. Below is an in-depth discussion of how fair use applies in the context of medical ST datasets.

The Four Factors of Fair Use: Fair Use is governed by Section 107 of the Copyright Act, which provides a legal framework for evaluating whether a specific use of copyrighted material qualifies under this doctrine. The determination of whether the use of copyrighted material qualifies as fair use typically hinges on four key factors, as described in Fair Use defined by the U.S. Copyright Office³⁰. We elaborate as below:

1. Purpose and character of the use:

- The purpose and character of the use focus on whether the use is for nonprofit, educational, or research purposes, which are more likely to qualify as fair use, as opposed to commercial purposes. In the case of medical ST datasets, fair use might apply if the data is used for research aimed at advancing public health, improving medical communication, or fostering innovations in machine learning for medical applications.
- Transformative use, where the material is repurposed or recontextualized for a different objective, also strengthens the argument for fair use. For example, using patient speech data to train Artificial Intelligence (AI) models for real-time MT in healthcare settings can be considered transformative because the original purpose (e.g., doctor-patient communication) is being altered to advance language and accessibility technologies.

2. Nature of the copyrighted work:

- The nature of the copyrighted material evaluates whether the original work is creative or factual. Works that are more factual, such as medical speech recordings or transcripts used for diagnosis and treatment, are generally more likely to fall under fair use than highly creative works like fictional narratives.
- Medical ST datasets often consist of factual information, such as conversations regarding symptoms, diagnoses, or medical instructions. These factual elements weigh in favor of fair use when the data is used to support research, training, or public health interventions.

3. Amount and substantiality of the portion used:

- This factor examines both the quantity and quality of the copyrighted material used. While fair use does not provide a specific threshold for the amount of material that can be used, using only what is necessary for the intended purpose is a key principle.
- In the context of medical ST datasets, this means limiting the dataset to include only the audio or textual data required for training or evaluation. Anonymizing or redacting non-essential information, such as personally identifiable details, can further support the argument for fair use by demonstrating that the dataset minimizes unnecessary use of protected content.

4. Effect of the use on the market for the original work:

- The potential impact on the market value of the original work is a crucial consideration. If the use of the copyrighted material negatively affects the market for the original work, it may weigh against fair use. For example, using proprietary medical transcription data for commercial purposes without authorization could harm the value of the original service or content.
- Conversely, using the material for nonprofit research or educational purposes, where there is no direct competition with the original work, is less likely to harm the market. In the case of medical speech datasets, it is important to consider whether the use might substitute for a commercial service or create a competitive disadvantage for the copyright holder.

³⁰<https://www.copyright.gov/fair-use/>

G.1.2 Ensuring Fair Use Compliance

While the principles of fair use offer a framework for leveraging copyrighted material in medical ST datasets, organizations and researchers should take proactive measures to minimize legal risks and maximize the ethical integrity of their projects. These measures include:

1. Anonymization and de-identification: Removing all personally identifiable information (PII) from the dataset is critical in the healthcare domain to comply with privacy laws such as HIPAA³¹ (Health Insurance Portability and Accountability Act) and GDPR³² (General Data Protection Regulation). This step not only enhances patient confidentiality but also strengthens the argument for fair use by limiting the dataset to factual, non-identifiable content.

2. Licensing and permissions: Where possible, obtaining licenses or permissions to use copyrighted material ensures that the dataset is fully compliant with intellectual property laws. For medical ST datasets, this might involve collaborating with healthcare providers, transcription services, or language experts who can provide content under appropriate agreements.

3. Transparency and documentation: Maintaining transparent records of how the data is sourced, processed, and used can demonstrate good-faith efforts to adhere to fair use principles, as we did here. Our documentation includes information on the purpose of the dataset and the intended audience or beneficiaries of the research.

4. Limiting commercialization: Restricting the dataset's use to non-commercial purposes, such as academic research or public health initiatives, can further justify fair use, as we did here. If commercialization is pursued, ensuring that the product or service is sufficiently transformative and does not compete with the original work is critical.

G.2 Data Consent

It is essential to approach this topic with sensitivity and a clear understanding of ethical, legal, and scientific principles. Researchers working with medical data, including datasets for ST, must carefully navigate the balance between advancing scientific progress and protecting patients' privacy and rights. Below is a discussion on why, in certain circumstances, researchers may not require patients' ex-

plicit consent to use medical ST data, grounded in ethical reasoning and legal frameworks.

1. De-identification and anonymization of data: One of our central arguments for not requiring explicit patient consent is that medical ST data used for research purposes is de-identified or anonymized. This means that all PII is removed or masked in a way that makes it impossible to trace the data back to a specific individual. Under frameworks such as the HIPAA in the United States, once data is de-identified, it is no longer considered protected health information (PHI). In such cases, researchers are not legally obligated to obtain patient consent, as the data no longer poses a risk to the individual's privacy or confidentiality.

2. Public benefit and the advancement of science: Medical research, including the development of ST systems, serves a broader public good by improving healthcare delivery, accessibility, and outcomes. For instance, creating accurate and effective ST systems can help break language barriers in healthcare settings, enabling better communication between patients and providers. By allowing researchers to access de-identified data without requiring individual consent, delays in critical advancements can be avoided, ultimately benefiting society as a whole. The collective societal gain is often considered to outweigh the need for individual consent in such cases.

3. Impracticability of obtaining consent: In many cases, medical ST datasets involve large-scale online collections of voice recordings or transcripts, often spanning years. Obtaining consent from every individual whose data is included in the dataset can be logistically impossible or financially prohibitive. This impracticability is especially pronounced when dealing with legacy data or when the individuals are no longer reachable. By allowing the use of such data without requiring explicit consent, researchers can ensure that valuable information is not lost to scientific progress.

Precedents in data use for research: The use of medical data for research purposes without explicit patient consent is not unprecedented. For example, population-level studies, epidemiological research, and biobank studies often use de-identified data without seeking individual consent. These practices are typically justified by their alignment with ethical guidelines, legal frameworks, and public health objectives. ST datasets are no different in this regard, as long as they are managed under similar principles of de-identification.

³¹<https://www.hhs.gov/hipaa/for-professionals/index.html>

³²<https://gdpr-info.eu/>

Transparency and accountability: Although individual consent may not be required, transparency remains a cornerstone of ethical research. Researchers are encouraged to publicly disclose the purpose, methods, and intended outcomes of their studies, as we did here. This helps build trust with the public and ensures accountability in the use of sensitive medical data.

In conclusion, the decision not to require explicit patient consent for medical ST datasets is grounded in the ethical principles of beneficence, justice, and respect for privacy, as well as the legal standards governing de-identified data. While consent is a vital component of ethical research in many contexts, exceptions are made when data is anonymized, the research serves a compelling public interest, and robust safeguards are in place to protect individuals' privacy. By adhering to these principles, we can balance the need for scientific progress with the ethical imperative to protect patient rights.

G.3 Annotation Problem for Long-form Speech

Transcription annotation for long-form audio often suffers from timestamp mismatches, which can significantly impact the accuracy of transcriptions and the usability of the data. These mismatches arise from multiple factors, including technical limitations of ASR models, human annotation inconsistencies, and the inherent challenges of handling long-form audio. Below are some of the key reasons why timestamp errors frequently occur in long-form ASR annotation.

1. ASR model drift: ASR systems process audio sequentially, and small timing drifts accumulate over long durations. Many ASR models generate timestamps by predicting words frame by frame based on phonetic models and language models (Lee et al., 2013). However, minor deviations in phoneme alignment at the start of the audio can snowball, leading to timestamp mismatches by the middle or end of a long recording.

Frame-based decoding delays: ASR models break audio into small frames (e.g., 10-20 ms each) (Tyagi et al., 2006). Slight misalignments in early frames can lead to progressive timestamp shifts.

2. Variable speech rates and pauses: Speakers naturally change their speaking rate, pause for effect, or speed up at certain points (Mirghafori et al., 1996). ASR systems and humans rely on predefined acoustic models or raw audio files that may not always capture these variations accurately.

- **Fast speech compression:** If a speaker speeds up, the ASR systems and humans may drop or misalign words, causing an early drift in timestamps (Mirghafori et al., 1996).
- **Extended pauses misinterpretation:** When a speaker pauses significantly, the ASR system may either insert silence markers inaccurately or assume the next word starts too early (Chen et al., 2015), leading a wrong reference for human validation and later manual translation.

3. Inconsistent segmentation strategies: Long-form audio is typically segmented into smaller chunks for processing efficiency, and different segmentation strategies can cause timestamp mismatches (Chang et al., 2021).

- **Fixed-length segmentation:** Some ASR pipelines divide audio into fixed intervals (e.g., 30-second segments) (Radford et al., 2022). If these segments do not align with sentence boundaries, words near the edges may be duplicated or omitted, leading to inaccurate timestamps.
- **Overlap handling issues:** Some ASR systems introduce small overlaps to avoid word truncation at segment boundaries (Cetin and Shriberg, 2006). This overlap can lead to duplicate word recognition and timestamp mismatches (Flynn and Ragni, 2023), also leading wrong reference for human annotators for transcription and translation.

4. Human annotation variability: Even in manually corrected ASR transcriptions, human annotators introduce timestamp errors due to cognitive biases, differences in annotation tools, and subjective interpretations of timing.

- **Different perceptions of word onset and offset:** Annotators may not consistently agree on where a word starts and ends, especially for words with soft or gradual onsets (e.g., "uhh", "well")
- **Tool latency and interface limitations:** Annotation software often has playback controls that introduce slight delays, affecting manual timestamp adjustments.
- **Multispeaker confusion:** When multiple speakers are involved, annotators may struggle to pinpoint precise timestamps for overlapping speech.

5. Background noise and acoustic challenges:

Long-form audio often includes background noise, cross-talk, and varying microphone quality, which impact the accuracy of human validation and thus timestamp precision. ([Maas et al., 2012](#))

6. Post-processing and formatting issues: After ASR transcription, further processing (such as punctuation insertion, casing normalization, or formatting corrections) can introduce wrong context for human validation.

- Text normalization adjustments: Some ASR systems correct text formatting after the initial transcription, potentially shifting timestamps ([Manohar and Pillai, 2024](#)).
- Forced alignment corrections: Post-processing often involves forced alignment techniques to match text to audio, but this can introduce additional errors when realigning sentences ([Mathad et al., 2021](#)), making synthetic MT transcripts unreliable for human validation.

H List of Abbreviations

AED	In automatic speech recognition (ASR), AED (Attention-based Encoder Decoder) is a neural model that maps speech features to text by using an encoder to process the audio sequence and an attention-guided decoder to generate the transcription step by step.. 4 , 27–29 , 33 , 34
AI	General term for systems that perform tasks requiring human-like intelligence. 120
ASR	It is the technology that converts spoken language into written text by analyzing and transcribing audio signals.. 2–10 , 23 , 24 , 27 , 29 , 30 , 32–35 , 40–42 , 55 , 58 , 61 , 65 , 69 , 73 , 77 , 91 , 93 , 95 , 97 , 99 , 122 , 123
CER	CER (Character Error Rate) is a metric that measures the accuracy of text recognition systems by calculating the ratio of character insertions, deletions, and substitutions to the total number of characters in the reference text.. 5 , 55
CNN	In automatic speech recognition (ASR), a CNN extracts local acoustic patterns from spectrograms by learning time-frequency features that help capture phonetic and speaker-invariant information.. 23 , 27 , 29
FFT	In automatic speech recognition (ASR), the Fast Fourier Transform (FFT) quickly converts audio signals from the time domain into frequency components, enabling models to analyze phonetic information.. 26
FFW	In machine learning, FFW (Feed-Forward Network) refers to a neural network where information moves only in one direction-from inputs through hidden layers to outputs-without feedback loops or recurrence.. 42–44
GQA	GQA (Grouped Query Attention) is a Transformer variant where multiple attention heads share key-value projections but keep separate query projections, reducing computation while preserving expressiveness.. 39
ITN	ITN (Inverse Text Normalization) is the process of converting normalized text (like "twenty twenty-five") back into its spoken-style or symbolic form (e.g., "2025").. 34 , 35
LLM	A Large Language Model is an AI system trained on vast text data to understand and generate human-like language.. 3 , 5 , 6 , 8 , 9 , 24 , 37–39 , 56–58
MFCC	MFCC (Mel-Frequency Cepstral Coefficients) are compact representations of speech audio that capture perceptually relevant frequency features, widely used in Automatic Speech Recognition (ASR) and as input features for speech-to-speech or speech-to-text Machine Translation (MT).. 26–29 , 44
MT	Machine translation is the automatic process of converting text or speech from one language into another using computational models.. 2–9 , 23–25 , 31 , 35 , 37 , 40–43 , 54 , 56 , 58 , 78 , 120 , 123
NMT	It is a deep learning approach that translates text directly from one language to another using end-to-end neural networks.. 23–25 , 37 , 40
RBMT	RBMT (Rule-Based Machine Translation) is a translation approach that relies on linguistic rules and bilingual dictionaries to convert text from a source language into a target language.. 24

RNN	A Recurrent Neural Network (RNN) in Automatic Speech Recognition (ASR) models sequential dependencies by processing speech frames one at a time, using past context to better predict phonemes or words.. 23
RNN-T	The Recurrent Neural Network Transducer (RNN-T) is an end-to-end sequence model that jointly learns acoustic and language modeling to directly map input audio to text without needing frame-level alignment.. 4 , 34
seq2seq	seq2seq is a neural network architecture that transforms one sequence (like a sentence) into another (like a translation) using an encoder-decoder structure.. 2 , 4 , 23 , 35 , 41
SFT	SFT (Supervised Fine-Tuning) means training a language model on parallel source-target sentence pairs so it learns to generate accurate translations in a supervised way.. 38 , 45 , 48 , 51
SMT	In machine translation, Statistical Machine Translation (SMT) is a method that generates translations by learning statistical patterns from large bilingual text corpora.. 24
SOTA	SOTA means "state of the art," the best or most advanced performance achieved in a given field at the present time.. 4
ST	Speech translation is the process of automatically converting spoken language in one tongue into spoken or written language in another, combining speech recognition, machine translation, and speech synthesis in real time.. 1–10 , 23 , 24 , 29 , 31–33 , 40 , 41 , 44–57 , 89 , 91 , 93 , 95 , 97 , 99–122
STFT	In automatic speech recognition (ASR), the Short-Time Fourier Transform (STFT) converts the raw audio waveform into a time-frequency representation by analyzing short overlapping windows, making speech features easier for models to process.. 26
WER	WER (Word Error Rate) is a common metric that measures how many words a speech recognition system got wrong compared to the reference transcript.. 5 , 55 , 61 , 65 , 69 , 73 , 77 , 91 , 93 , 95 , 97 , 99

I Contribution Statements

This list shows major contributions to this work:

1. **Khai Le-Duc** led all aspects of this work, including ideation, experimental setup, paper writing and business development.
2. **Tuyen Tran** conducted the experiments.
3. **Bach Phan Tat** led the data annotation team.
4. **Nguyen Kim Hai Bui** conducted all the data processing.