LLaST: Improved End-to-end Speech Translation System Leveraged by Large Language Models

Xi Chen[1], **Songyang Zhang**[2], **Qibing Bai**[1], **Kai Chen**[2], **Satoshi Nakamura**[1,3]
[1]The Chinese University of Hong Kong, Shenzhen
[2]Shanghai AI Laboratory
[3]Nara Institute of Science and Technology, Japan
{xichen7@link,snakamura@}.cuhk.edu.cn

Abstract

We introduces **LLaST**, a framework for building high-performance **L**arge **La**nguage model based **S**peech-to-text **T**ranslation systems.We address the limitations of end-to-end speech translation (E2E ST) models by exploring model architecture design and optimization techniques tailored for LLMs. Our approach includes LLM-based speech translation architecture design, ASR-augmented training, multilingual data augmentation, and dual-LoRA optimization. Our approach demonstrates superior performance on the CoVoST-2 benchmark and showcases exceptional scaling capabilities powered by LLMs. We believe this effective method will serve as a strong baseline for speech translation and provide insights for future improvements of the LLM-based speech translation framework[1].

1 Introduction

The speech-to-text translation (ST) task, which transcribes spoken language into written text in a different language, is pivotal for bridging communication barriers. This capability has a wide array of applications, including facilitating global communication, enabling automatic subtitling, and aiding in language learning.

Conventional ST systems are typically composed of two distinct components: an *automatic speech recognition* (ASR) module that transcribes spoken speech into written text in the source language, and a *machine translation* (MT) module that subsequently translates this text into the target language. These modules can be trained using paired ASR and text-to-text translation data, significantly enhancing the overall performance of ST systems. Despite their modular design, cascade systems are prone to error accumulation, where inaccuracies from the ASR stage are compounded in the MT phase, often leading to sub-optimal translations. Recently, the focus has shifted towards the development of end-to-end speech translation (E2E ST) models that bypass the need for separate automatic speech recognition (ASR) and machine translation (MT) modules by directly converting spoken input into text in the target language. Nonetheless, these approaches often necessitate extensive training datasets and are contingent upon sophisticated model architectures to achieve strong performance.

Speech translation is intrinsically linked to natural language processing (NLP), as it involves the conversion of spoken language into written text in a target language, necessitating a deep understanding of both the source and target languages' linguistic structures and semantics. The unprecedented capabilities that large language models (LLMs) have demonstrated across a variety of NLP tasks (Touvron et al., 2023a,b; Achiam et al., 2023) have opened up new possibilities to construct potent speech translation systems by leveraging these LLMs as a foundation. Recent research has seen some preliminary attempts exploring this direction (Chu et al., 2023; Wu et al., 2023; Huang et al., 2023). Despite these advancements, the question remains on how to most effectively harness the vast potential of LLMs to develop a high-performance ST system in an efficient manner, without compromising on quality or scalability.

In this study, we focus on the exploration of best practices for constructing an effective speech translation system powered by Large Language Models (LLMs), which we term **LLaST**. The paper delves into the core aspects of the development process, specifically the *model architecture design* and *optimization techniques*. Our exploration begins with the creation of a minimalist model architecture, examining the selection of key modules such as the speech encoder and LLMs. Subsequently, we investigate training strategies, including *ASR-augmented training* and *dual-LoRA optimization*. Moreover, to

[1]We release the data, code and models in https://github.com/openaudiolab/LLaST

Findings of the Association for Computational Linguistics: ACL 2024, pages 6976–6987
August 11-16, 2024

deepen our understanding of scaling laws in LLM-based ST, we also scrutinize the impact of model size variations. Through these concerted efforts, we aim to uncover insights that can significantly enhance the performance and training efficiency of LLaST.

Our contributions are listed as follows.

• We explore the LLMs-based speech translation method, including model architecture design, training strategies, and data recipe.

• Extensive evaluations demonstrate the superiority of our approach, surpassing the previous SOTA methods (Barrault et al., 2023) and achieving **45.1** BLEU on the `fr`$\rightarrow$`en` test set of CoVoST-2.

• We are dedicated to making all data recipes, training methodologies, and model weights associated with LLaST openly accessible to the community. By doing so, we foster transparency, collaboration, and advancement in the field of LLM-based speech translation technology.

2 Related Work

2.1 Cascaded Speech Translation

Historically, the construction of speech translation systems has been approached in a cascading fashion, incorporating both an ASR and an MT subsystem (Stentiford and Steer, 1988; Ney, 1999; Nakamura et al., 2006). The procedure involves initially converting the input speech into text in the source language, which is subsequently translated into the target language. The primary objective of this line of research has been to mitigate error accumulation, including the use of multiple recognition outputs and the development of robust MT models (Casacuberta et al., 2008; Kumar et al., 2014; Sperber et al., 2017). Sperber et al. (2019b) introduces a self-attention mechanism to handle the lattice inputs, and Zhang et al. (2019) proposes a lattice transformer, equipped with a controllable lattice attention mechanism, to derive latent representations. Lam et al. (2021) establishes a feedback cycle in which the downstream performance of the MT system serves as a signal to enhance the ASR system via self-training.

2.2 End-to-End Speech Translation

The development of end-to-end speech translation (E2E ST) models, which bypass the requirement for intermediary stages such as ASR outputs and lattices, has been a significant stride in mitigating error propagation. Research indicates that these E2E ST models demonstrate encouraging results and offer performance on par with cascaded models (Sperber et al., 2019a; Ansari et al., 2020; Bentivogli et al., 2021; Ye et al., 2021). Moreover, these models present additional benefits such as lower latency and the potential to be applied to languages that lack a written form (Bérard et al., 2016).

Data scarcity and the modeling burden are recognized as two significant obstacles impeding the performance of E2E ST (Xu et al., 2023). Firstly, the intrinsic complexity of speech translation, which integrates transcription and translation, presents a challenge in optimizing a single model to accomplish these cross-modal and cross-lingual tasks in one step. Secondly, ASR datasets are typically less extensive than MT datasets, and the extension to ST datasets further exacerbates this size discrepancy. To address this issue of data scarcity, researchers have employed strategies such as data augmentation (Tsiamas et al., 2023; Lam et al., 2022), pre-training (Wang et al., 2020c; Ao et al., 2022), and knowledge distillation (Liu et al., 2019), which leverage external datasets.

To alleviate the modeling burden, a variety of multi-task learning strategies have been investigated (Zhang and Yang, 2018). Originating from the multi-task encoder-decoder architecture (Weiss et al., 2017), some researchers have chosen to split the decoder into two separate components (Liu et al., 2020a; Anastasopoulos and Chiang, 2018): one dedicated to transcription and the other to translation. Parallel research efforts (Liu et al., 2020b; Cheng et al., 2023) have similarly decoupled the encoder, with further work showing that a shared encoder can be independently partitioned (Tang et al., 2021; Ye et al., 2022) to make better use of ASR data. In addition, non-autoregressive (NAR) modeling has been explored as a means to decrease latency (Inaguma et al., 2021; Chuang et al., 2021).

Significantly, recent advancements have also delved into multi-tasking within the context of large-scale training, leading to impressive results on ST benchmarks. For instance, Whisper (Radford et al., 2023) and SeamlessM4T (Barrault et al., 2023) have incorporated 680k and 470k hours of multilingual speech data in their training.

2.3 LLM-based Speech Translation

Inspired by the robust linguistic capabilities of LLMs (Brown et al., 2020; Touvron et al., 2023b), recent initiatives have sought to harness the

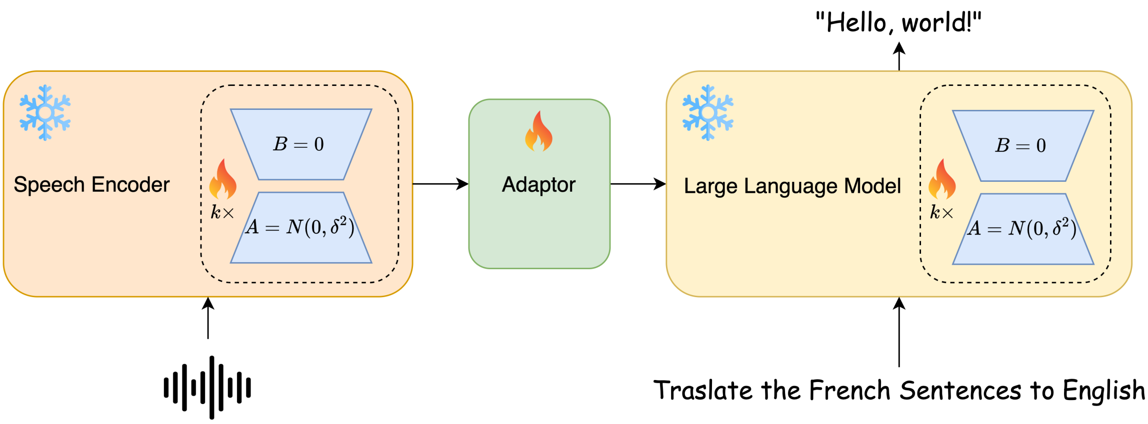

Figure 1: **Model Architecture of LLaST** We introduce *dual-LoRA* in the optimization, and keep weights of the speech encoder and LLM frozen. We use a 3-layer MPLs for adaptor and fine-tune its parameters together with dual-LoRA.

power of LLMs to address various speech tasks, aided mainly by instruction tuning. The prevailing method involves integrating an LLM (backend) with a speech encoder (frontend). Models like LauraGPT (Chen et al., 2023) and Qwen-audio (Chu et al., 2023) support a range of multimodal speech tasks, demonstrating performance comparable to task-specific E2E ST models. ViOLA (Wang et al., 2023) employs a neural codec model (Défossez et al., 2022) to discretize the speech input while tuning the LLM. Similarly, AudioPaLM (Rubenstein et al., 2023) discretizes the speech input and achieves commendable results on CoVoST-2 (Wang et al., 2020b).

Salmonn (Tang et al., 2023) employs two encoders as the frontend and uses LoRA (Hu et al., 2021) for efficient fine-tuning. However, the extent of its performance improvement on ST remains largely unexplored. Some recent studies (Wu et al., 2023; Zhang et al., 2023a) specifically target the ST task and delve into efficient tuning strategies, but their performance enhancements have been somewhat limited. In an industrial study focusing on translation between Chinese and English, Huang et al. (2023) additionally incorporates the Chain-of-Thought (CoT) technique (Wei et al., 2022), enabling a step-by-step approach using LLMs.

3 Method

This section presents our method in detail. We begin by introducing the problem setting of the speech-to-text translation task in Sec. 3.1. Then, we explain the structure of the proposed model in Sec. 3.2, followed by the description of the training and inference processes in Sec. 3.3.

3.1 Problem Setting

We now present the problem setting of speech translation. Given a speech translation dataset $\mathcal{D} = \{(\mathbf{S}, \mathbf{Y}_{src}, \mathbf{Y}_{tgt})\}$, the source language speech $\mathbf{S}$'s acoustic features (e.g., mel-spectrogram) are denoted as $\mathbf{X}_s$, and we have:

$$\mathbf{X}_s = \mathcal{F}_a(\mathbf{S}), \quad \mathbf{X}_s = \{x_1, x_2, \dots, x_T\}$$

where $\mathcal{F}_a$ is the acoustic feature extraction operation, and T is the timesteps of the input features. $\mathbf{Y}_{src}$ and $\mathbf{Y}_{tgt}$ are the transcripts of $\mathbf{S}$ in the source and target languages, respectively. The goal of speech translation is to generate the prediction text of target language $\hat{\mathbf{Y}}_{tgt}$ from the source speech $\mathbf{S}$.

We can formulate the whole process as:

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

and $\mathcal{F}$ represents the entire ST system. Performance of ST system is typically assessed by comparing the predicted output $\hat{\mathbf{Y}}_{tgt}$ with the ground truth $\mathbf{Y}_{tgt}$ using metrics like BLEU (Papineni et al., 2002).

3.2 Model Architecture

Our objective is to develop the LLaST model with a simple architecture, as depicted in Figure 1. The design of LLaST comprises three key components: a speech encoder to process the input speech, an adaptor that projects these speech features into the compatible feature space for Large Language Models (LLMs), and finally, a decoder-only LLM for multi-modality decoding.

Example of Speech-text Prompt for LLaST

Speech Translation Prompt:
<audio><AudioInputs></audio> Translate the French sentence to English. Transcripts of AudioInputs is "Bonjour le monde."

Expected Output:
Hello world.

Automatic Speech Recognition Prompt:
<audio><AudioInputs></audio> Transcribe the French sentence to French. Transcripts of AudioInputs is "Bonjour le monde."

Expected Output:
Bonjour le monde.

Figure 2: **An example for training data.**

Speech Encoder Acoustic features $\mathbf{X}_s$ encapsulate a wealth of information, including speaker traits, emotions, prosody, background noise, and more. The role of the speech encoder is to disentangle these variabilities and generate robust linguistic representations, denoted as $\mathbf{Z}_s$. We define this process mathematically:

$$\mathbf{Z}_s = \mathcal{F}_{se}(\mathbf{X}_s)$$

where $\mathcal{F}_{se}$ represents the speech encoder function. Our work investigates various options for the speech encoder, with a focus on mHubert (Hsu et al., 2021; Lee et al., 2021) and Whisper (Radford et al., 2023). For an in-depth analysis and discussion on the speech encoder selection, please refer to Sec. 5.1.

Adaptor The adaptor acts as a bridge between the speech encoder and the Large Language Model (LLM), consisting of a lightweight set of trainable parameters. Fine-tuning these parameters aligns speech features more effectively with the LLM's representation space. Its function is to project the extracted linguistic representations, $\mathbf{Z}_s$, into the embedding realm of the LLM, thus yielding $\mathbf{H}_s$:

$$\mathbf{H}_s = \mathcal{F}_{ada}(\mathbf{Z}_s)$$

This transformation process facilitates a smooth integration of speech data into the LLM's text-based context. We adopt a 3-layer multilayer perceptrons(MLPs) for adaptor.

Large Language Model Equipped with the projected speech feature $\mathbf{H}_s$, our objective is to utilize the Large Language Model (LLM) for generating the translated text of the original speech. To facilitate this, we construct a speech-text prompt input for the LLM. The text component of this prompt, denoted as $\mathbf{X}_q$, conveys the specific translation task instruction, such as "`Translate the French sentence into English`". Post-tokenization and embedding, $\mathbf{X}_q$ is transformed into the LLM's input representation, $\mathbf{H}_q$. Subsequently, the LLM generates translation predictions based on the concatenated speech-text features (for simplicity, we omit `bos` and `eos` tokens in the equation below):

$$\hat{\mathbf{Y}}_{tgt} = \mathcal{F}_{llm}([\mathbf{H}_s, \mathbf{H}_q])$$

This process allows the model to fuse speech and textual information effectively to produce translations.

In summary, the entire process can be expressed as:

$$\hat{\mathbf{Y}}_{tgt} = \mathcal{F}(\mathbf{S}) = \mathcal{F}_{llm}([\mathcal{F}_{ada}(\mathcal{F}_{se}(\mathcal{F}_a(\mathbf{S}))), \mathbf{H}_q])$$

3.3 Training and Inference

This section delves into the optimization techniques employed in LLaST and elucidates its inference methodology.

Optimization with Dual-LoRA Fintuning To enhance training efficiency, we employ the LoRA (Hu et al., 2021) tuning method for model optimization. This technique significantly reduces trainable parameters by introducing trainable rank decomposition matrices to each Transformer layer, while keeping the pre-trained weights frozen.

In LLaST, we introduce the *dual-LoRA fine-tuning*, applying LoRA separately to both the speech encoder (S-LoRA) and the Large Language Model (L-LoRA). This approach ensures effective adaptation to speech translation tasks with minimal parameter updates. Specifically, we perform instruction-tuning on prediction tokens using the original auto-regressive training objective of LLM. For a target translation result $\mathbf{Y}_{tgt}$ of length N, its

Model	Speech Encoder	Adaptor	LLM
LLaST-2B	Whisper-large-v2	MLPs	TinyLlama-1.1B-Chat
LLaST-8B	Whisper-large-v2	MLPs	Llama2-7B-Chat
LLaST-14B	Whisper-large-v2	MLPs	Llama2-13B-Chat

Table 1: **Configurations of LLaST models**. We use Whisper(large-v2) and 3 layers MLPs for all LLaST models.

probability is calculated as:

$$P(\mathbf{Y}_{tgt}|\mathbf{X_s}, \mathbf{X}_q) = \prod_{i=0}^{N} P_\theta(y_i|\mathbf{X}_s, \mathbf{X}_q, \mathbf{Y}_{tgt,<i})$$

This strategy allows us to efficiently tune LLaST without extensive retraining, maintaining both computational efficiency and task-specific effectiveness.

Training with ASR-augmentation To enhance the performance of LLaST, we adopt the strategy from prior work (Barrault et al., 2023; Radford et al., 2023) to incorporate Automatic Speech Recognition (ASR) tasks for data augmentation during training. Given the structural similarity between ASR and ST tasks—both involve converting speech to text, we can simply modify the ASR prompt to match ST objectives, such as `"Transcribe the French sentence into English"`. The examples of prompts are listed in Fig. 2. This ASR-augmentation significantly boosts the effectiveness of LLaST across various language pairs, as detailed in Sec. 5.2.

Inference Methodology During inference, we construct prompts in the same format as depicted in Fig. 1. To generate translation text sequences $\hat{\mathbf{Y}}_{tgt}$, we employ a beam search algorithm with a beam size of 5.

4 Experiments

In this section, we conduct a series of experiments to validate the effectiveness of our method. We start by detailing experimental configurations in Sec. 4.1, followed by an overview of quantitative results in Sec. 4.2.

4.1 Configurations

Datasets Our speech translation models are trained and evaluated on CoVoST-2 (Wang et al., 2020b), a large-scale multilingual dataset that supports translations between English and 15 other languages, as well as from 21 languages into English. For monolingual experiments, we utilize six subsets with source languages translating to English, focusing on French-English for training and testing. In the multilingual setup, we employ `Fr`→`En`, `Es`→`En`, `De`→`En`, `It`→`En`, `Zh`→`En`, and `Ja`→`En` subsets and three English-to-X subsets: `En`→`Zh`, `En`→`Ja`, and `En`→`De` . Audio samples are downsampled from 48kHz to 16kHz in all experiments.

Model Architecture Tab. 1 presents the three LLaST model configurations. Each model utilizes a Whisper-large-v2 speech encoder, contributing approximately 1B parameters. The adaptor is a compact multilayer perceptron with three layers, ingesting 1280-dimensional inputs and adjusting its output dimensions to match those of the subsequent LLMs. Consequently, the overall parameter count is predominantly influenced by the LLM component. Hence, we denote our models as LLaST-2B, LLaST-8B, and LLaST-14B.

Hyperparameters All models are optimized with AdamW, setting $\beta_1 = 0.9$ and $\beta_2 = 0.98$. A warmup-then-linear decay learning rate schedule is adopted, peaking at 0.0002. Training spans one epoch for each model. By default, the rank of S-LoRA (Whisper LoRA) is set to 128, while L-LoRA (LLM LoRA) rank is 512 unless specified otherwise. The LLaST-8B and LLaST-14B models are trained using 32 NVIDIA A100 GPUs, each with a batch size of 32, while the smaller LLaST-2B model is trained on a setup consisting of 8 A100 GPUs, maintaining the same batch size per GPU.

4.2 Main Results

Comparisons with Other Models Tab. 2 presents a comparison between our proposed LLaST models and previous methods, with SacreBLEU scores evaluated across six language pairs: `Fr`→`En`, `Ja`→`En`, `De`→`En`, `Zh`→`En`, `Es`→`En`, and `It`→`En`. Notably, LLaST-2B outperforms SeamlessM4T(medium) and demonstrates competi-

Model	Parms.	X→ English					
		French	Japanese	German	Chinese	Spanish	Italian
Baseline Models							
S2T_Transformer (Wang et al., 2020a)	0.04B	27.2	N/A	18.2	N/A	25.1	N/A
SpeechLLaMA (Wu et al., 2023)	7B	25.2	19.9	27.1	12.3	27.9	25.9
Whisper-small (Radford et al., 2023)	0.25B	27.3	17.3	25.3	6.8	33.0	24.0
Whisper-large-v2 (Radford et al., 2023)	1.6B	36.4	26.1	36.3	18.0	40.1	30.9
Qwen-audio (Chu et al., 2023)	8B	38.5	N/A	33.9	15.7	39.7	36.0
SeamlessM4T(medium) (Barrault et al., 2023)	1.2B	38.4	15.2	34.7	18.0	38.7	36.5
SeamlessM4T(large-v2) (Barrault et al., 2023)	2.3B	42.1	23.8	39.9	22.2	42.9	40.0
Our Models							
LLaST-2B	2B	41.2	24.2	36.8	19.2	43.2	39.3
LLaST-8B	8B	44.1	24.4	40.8	23.3	45.3	42.1
LLaST-14B	14B	**45.1**	**28.8**	**41.2**	**24.8**	**46.1**	**43.0**

Table 2: **Performance comparison on CoVoST-2 X→ English test set.** We use SacreBLEU scores as metrics for all experiments, the models are trained with multi-lingual data.

Speech Encoder	LLM	BLEU
mHuBERT	TinyLlama	24.4
Whisper-base	TinyLlama	28.7

Table 3: **Influence of different speech encoders.** For speech encoder, mHuBERT-base(95M) and Whisper-base(74M) share the similar model size. We use TinyLlama-1.1B-Chat (Zhang et al., 2024) in this study. We report SacreBLEU scores on CoVoST-2 fr→ en test set for all experiments.

tive performance against SeamlessM4T(large-v2). LLaST-8B significantly excels by improving upon the Qwen-audio model of similar scale with an impressive **5.6** BLEU point gain on the Fr→En task. Furthermore, LLaST-14B achieves state-of-the-art (SOTA) results, attaining a BLEU score of **45.1** on CoVoST-2's Fr→En subset, surpassing SeamlessM4T(large-v2) by **3.0** BLEU points. These results convincingly demonstrate the superiority of LLaST and highlight the promising potential of exploring LLMs for speech translation tasks.

5 Ablation Analysis

In this section, we delve into a meticulous ablation study and analysis of LLaST. We begin by examining the impact of model architecture in Sec. 5.1, followed by an exploration of optimization strategies in Sec. 5.2. Finally, we investigate the relationship between model scale and performance in Sec. 5.3.

5.1 Model Architecture Design

Choice of Speech Encoder We experiment with various speech encoder architectures, including mHuBERT (Hsu et al., 2021; Lee et al., 2021) and Whisper (Radford et al., 2023) model. For the mHuBERT, we adhere to the preprocessing approach from (Dong et al., 2023; Lee et al., 2021) to extract semantic units. For a fair comparison, we select the Whisper-base model, which is comparable in size to the mHuBERT model. Performances reported in Tab. 3 indicate that the Whisper model yields superior performance, demonstrating a **4.3** BLEU score improvement over mHuBERT. This improved performance can be attributed to the fact that Whisper has been trained on significantly more data, thus generating more representative linguistic features.

Choice of Large Language Models We examine the impact of different large language models within LLaST to discern how variations in language modeling performance affect its speech translation capabilities. We present X→en results in Figure 3. Notably, Qwen achieves a score of 47.3 on the en→zh test set, outperforming Llama2 (Touvron et al., 2023a) by **4.9** BLEU points. Similarly, InternLM2(Cai et al., 2024) surpasses Llama2 by **5.0** BLEU points. These findings suggest that Chinese-oriented LLMs notably enhance performance on Chinese-related ST tasks, exemplified by En→Zh and Zh→En. The LLaST model, when coupled with Llama2, demonstrates exceptional performance particularly in the Fr→En and De→En language pairs. This intriguing observation underscores the potential of LLM-based ST approaches, as they allow for effortless integration of diverse LLM strengths tailored to specific languages or tasks.

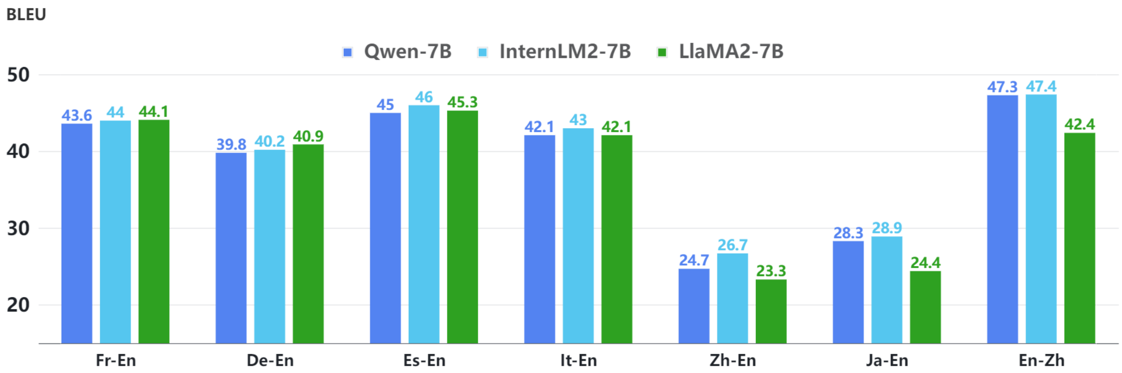

Figure 3: **Influence of different language models.** We use Whisper-large-v2 as speech encoder and report SacreBLEU scores on CoVoST-2 test set for all experiments.

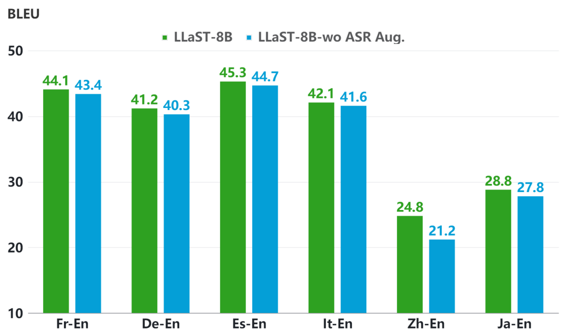

Figure 4: **Influence of different LLMs and ASR-augmentation.** We report SacreBLEU scores on CoVoST-2 test set for all experiments.

Speech Encode	Multi-Ling.	BLEU
Whisper-large-v2	✗	42.5
Whisper-large-v2	✓	**44.1**

Table 4: **Study of training with multilingual data**. We use Llama2-7B-Chat for LLMs and report SacreBLEU scores on CoVoST-2 `fr`→ `en` test set for all experiments.

5.2 Optimization

Training with ASR Augmentation Automatic Speech Recognition (ASR) is a task akin to speech translation, as both involve converting speech into text. Prior research has leveraged ASR tasks as auxiliary objectives for ST training (Zhang and Yang, 2018; Ye et al., 2022; Zhang et al., 2023b), or used models pre-trained on ASR data (Wang et al., 2020a). In LLaST, we adopt this concept and incorporate ASR tasks to optimize LLaST performance. An example of the speech-text prompt structure can be found in Fig. 2, where ST and ASR samples are randomly mixed during training, with the focus remaining on the ST task at inference time. The results presented in Fig. 4 demonstrate the efficacy of ASR augmentation in optimizing LLaST. We observe across nearly all test sets that ASR augmentation improves ST performance, suggesting that leveraging ASR or multi-task training within LLM-based ST frameworks is a promising direction with significant potential for future work.

Multilingual Data Augmentation In our experiments, we explore both monolingual and multilingual settings. Specifically, for the monolingual setup, we employ the `Fr`→`En` language pair. In the multilingual scenario, we introduce additional language pairs while maintaining the `Fr`→`En` data identical to that in the monolingual experiment.

The results presented in Tab. 4 reveal that incorporating other language pairs indeed benefits the `Fr`→`En` translation task, with a **1.6** BLEU score improvement observed upon adding multilingual data augmentation. This finding aligns with similar phenomena reported in LLM research (Team, 2023; Zeng et al., 2022), where exposure to multilingual corpora has been shown to enhance the language modeling capabilities of these models.

Dual-LoRA Optimization We investigate the impact of employing dual-LoRA for both speech encoders and large language models. In the ablation experiments, we utilize *Whisper-large-v2* and *Llama2-7B*. The results from scenarios without any LoRA, with LoRA applied only to Whisper, LoRA applied only to Llama2, and dual-LoRA are reported in Table 5. From these outcomes, it is evident that even with a lightweight adaptor, leveraging a strong speech encoder and LLM can yield commendable performance. We also discover that

Adaptor	S-LoRA	L-LoRA	BLEU
✓	✗	✗	40.5
✓	✓	✗	41.3
✓	✗	✓	43.6
✓	✓	✓	**44.1**

Table 5: **Ablation study of dual-LoRA optimization strategy**. S-LoRA means LoRA used in Whisper, and L-LoRA means the LoRA used in LLM. We use Whisper-large-v2 and Llama2-7B-Chat for speech encoder and LLMs, respectively. And we report SacreBLEU scores on CoVoST-2 fr→ en test set for all experiments.

applying single LoRA to either Whisper or Llama2 separately leads to substantial gains, improving scores from 40.5 to **41.3** and **43.6**, respectively. More notably, when dual-LoRA is used to jointly optimize both speech encoder and large language model, an additional improvement is achieved, culminating in a **44.1** BLEU score on test set.

5.3 Impact of Model Scale

Different Size of Speech Encoder We maintain a constant language model, **Llama2-7B**, and vary the size of Whisper models acting as speech encoders to examine the effect of encoder size on performance. The range of encoder sizes spans from 40M to 800M parameters. As shown in Table 6, we observe that as the encoder size increases, the BLEU score of the model consistently improves; however, the rate of improvement diminishes with each incremental increase in size. The base encoder achieves a BLEU score of 37.0, while the large encoder attains a peak score of **44.1**. This considerable leap underscores the importance of scaling up speech encoders for better speech-to-text translation. However, future research should consider the trade-offs between model size, computational efficiency, and overall performance to strike the right balance for practical applications.

Different Size of LLMs We further investigate the impact of varying LLM sizes on speech translation performance. With the speech encoder consistently set as *Whisper-large-v2*, we assess three different scale LLMs: TinyLlama-1B, Llama2-7B, and Llama2-13B. The outcomes are presented in Tab. 2. Our findings reveal that there is a positive correlation between the size of the language model and the BLEU scores across all test sets. As the capacity of the LLM increases, so does the overall performance in terms of translation quality, indicating that larger models can capture more nuanced linguistic patterns and generate more accurate translations.

Speech Encoder	Encoder Size	BLEU
Whisper-base	∼40 M	37.0
Whisper-small	∼120 M	41.2
Whisper-medium	∼390 M	43.1
Whisper-large-v2	∼800 M	**44.1**

Table 6: **Ablation study of model size of Whisper model**. We use Llama2-7B-Chat for LLM and report SacreBLEU scores on CoVoST-2 fr→ en test set.

Comparison between the Encoder Scaling and Decoder Scaling Given the Tab. 6 and Tab. 2, we observe some interesting phenomena. Despite Whisper-small+Llama2-7B-Chat and LLaST-2B demonstrating nearly equivalent performance on the fr->en subset, the former operates with approximately 7B parameters, whereas LLaST-2B functions with only about 2B parameters. This suggests that, in terms of parameter efficiency, scaling the encoder is a more effective strategy. It also indicates that, in these experiments, the encoder may play a more significant role. Meanwhile, the performance of the LLM-based system has yet to converge with respect to scale. To draw more comprehensive conclusions, we may need to continue scaling up the Whipster model and experiment with LLMs larger than 13B. For instance, in the domain of vision-language models, LLaVA (Liu et al., 2023) and InternVL (Chen et al., 2024) demonstrate that achieving optimal performance with a larger vision encoder (6B) necessitates employing correspondingly larger LLMs, such as Yi-34B (AI et al., 2024).

6 Limitation

While our study has yielded significant findings, it is crucial to recognize the limitations that may impact the interpretation and broad applicability of our results. Although we delved into the architecture design and optimization strategies, our reliance on a relatively narrow data source and the use of short voice samples could potentially affect the generalizability of our outcomes. To address this, future research will expand to encompass a more diverse array of data. Moreover, due to the constraints of our current resources, we have not ventured into exploring larger language models or a broader range of language pairs in this study.

In speech translation, LLaST's use of LLMs raises concerns in actual application as teh following: (a)Probabilistic inaccuracy, mistranslations may occur due to nuances or dialects, impairing accuracy and cultural relevance. (b)Data imbalances, insufficient representation in training data can lead to biased translations or reduced effectiveness for underrepresented groups. (c)Deployment challenges, large model sizes and complexity may cause latency, high energy usage, and device compatibility issues. (d)Harmful content generation, despite post-processing, risks persist; ongoing monitoring, filter refinement, and expert collaboration are needed.

7 Conclusion

We presents the development and analysis of LLaST, a novel speech translation model that harnesses LLM in this work. The study demonstrates that integrating well-tuned speech encoders like Whisper with different sizes of LLMs significantly improves speech-to-text translation performance. Through meticulous ablation studies, it is shown that applying dual LoRA optimization to both speech encoders and LLMs leads to substantial gains in BLEU scores. Additionally, experiments confirm that increasing the scale of either the speech encoder or the LLM positively impacts performance, though the rate of improvement decreases as size increases. Furthermore, incorporating ASR augmentation and multilingual training further enhances the model's performance on specific language pairs. Overall, LLaST underscores the potential of large language models for advancing speech translation tasks and offers valuable insights into their effective integration.

Ethical Considerations

We use the public LLMs to build LLaST, the LLMs may produce unexpected outputs due to its size and probabilistic generation paradigm. For example, the generated responses may contain biases, discrimination, or other harmful content. Addtionally, we use ChatGPT and Grammarly to polish the writing.

References

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

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Francisco Casacuberta, Marcello Federico, Hermann Ney, and Enrique Vidal. 2008. Recent efforts in spoken language translation. *IEEE Signal Processing Magazine*, 25(3):80–88.

Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, et al. 2023. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks.

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.

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, Online. Association for Computational Linguistics.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Qianqian Dong, Zhiying Huang, Chen Xu, Yunlong Zhao, Kexin Wang, Xuxin Cheng, Tom Ko, Qiao Tian, Tang Li, Fengpeng Yue, et al. 2023. Polyvoice: Language models for speech to speech translation. *arXiv preprint arXiv:2306.02982*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*.

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.

Gaurav Kumar, Matt Post, Daniel Povey, and Sanjeev Khudanpur. 2014. Some insights from translating conversational telephone speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3231–3235. IEEE.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2021. Cascaded models with cyclic feedback for direct speech translation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7508–7512. IEEE.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 245–254.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. 2021. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proc. Interspeech 2019*, pages 1128–1132.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020a. 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. 2020b. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 517–520. IEEE.

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 of the Association for Computational Linguistics*, pages 311–318.

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.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1380–1389, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019a. 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, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019b. Self-attentional models for lattice inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1185–1197, Florence, Italy. Association for Computational Linguistics.

Fred WM Stentiford and Martin G Steer. 1988. Machine translation of speech. *British Telecom technology journal*, 6(2):116–122.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

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.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models, 2023. *URL https://arxiv.org/abs/2307.09288*.

Ioannis Tsiamas, José Fonollosa, and Marta Costa-jussà. 2023. SegAugment: Maximizing the utility of speech translation data with segmentation-based augmentations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8569–8588, Singapore. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020c. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proc. Interspeech 2017*, pages 2625–2629.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6796–6804. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-End Speech Translation via Cross-Modal Progressive Training. In *Proc. Interspeech 2021*, pages 2267–2271.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Hao Zhang, Nianwen Si, Yaqi Chen, Wenlin Zhang, Xukui Yang, Dan Qu, and Xiaolin Jiao. 2023a. Tuning large language model for end-to-end speech translation. *arXiv preprint arXiv:2310.02050*.

Pei Zhang, Niyu Ge, Boxing Chen, and Kai Fan. 2019. Lattice transformer for speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6475–6484, Florence, Italy. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43.

Yuhao Zhang, Chen Xu, Bei Li, Hao Chen, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023b. Rethinking and improving multi-task learning for end-to-end speech translation. *arXiv preprint arXiv:2311.03810*.