Adaptive Inner Speech-Text Alignment for LLM-based Speech Translation

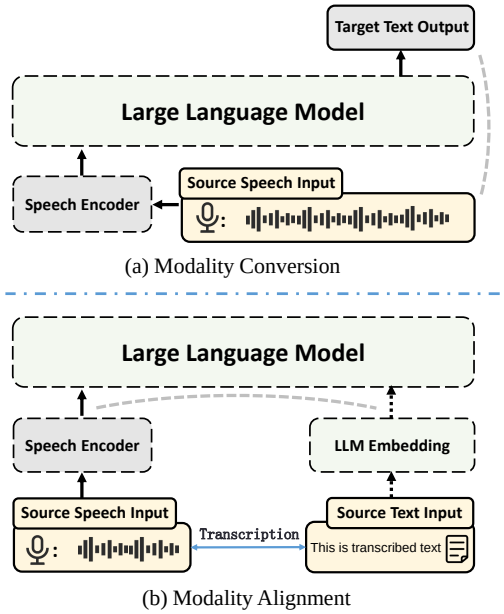
Anonymous ACL submission

Abstract

Recent advancement of large language models (LLMs) has led to significant breakthroughs across various tasks, laying the foundation for the development of LLM-based speech translation systems. Existing methods primarily focus on aligning inputs and outputs across modalities while overlooking deeper semantic alignment within model representations. To address this limitation, we propose an **A**daptive **I**nner **S**peech-**T**ext **A**lignment **(AI-STA)** method to bridge the modality gap by explicitly aligning speech and text representations at selected layers within LLMs. To achieve this, we leverage the optimal transport (OT) theory to quantify fine-grained representation discrepancies between speech and text. Furthermore, we utilize the cross-modal retrieval technique to identify the layers that are best suited for alignment and perform joint training on these layers. Experimental results on speech translation (ST) tasks demonstrate that **AI-STA** significantly improves the translation performance of large speech-text models (LSMs), outperforming previous state-of-the-art approaches. Our findings highlight the importance of inner-layer speech-text alignment in LLMs and provide new insights into enhancing cross-modal learning.[1].

1 Introduction

The emergence of large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Anil et al., 2023; Chiang et al., 2023) has achieved remarkable success across numerous natural language processing (NLP) tasks (OpenAI, 2024) and various studies extend its generative capabilities to multimodal domains (Chen et al., 2023; Zhang et al., 2023b; Li et al., 2023; Rubenstein et al., 2023; Li et al., 2024). The unprecedented capabilities of LLMs have laid the foundation for leveraging these models as the foundation for building powerful speech translation (ST) systems (Sethiya and Maurya, 2024).



(a) Modality Conversion

(b) Modality Alignment

Figure 1: Different training paradigm: *Modality Conversion* implicitly learns speech-text relationships from paired data, focusing on end-to-end mapping. While *Modality Alignment* explicitly enforces semantic consistency by aligning representations through supervised objectives.

To equip text-based LLMs with speech capabilities, recent research has investigated multiple approaches for developing large speech-text models (LSMs). These methods include expanding text-based LLMs vocabulary and adopting large-scale speech-text pre-training (Rubenstein et al., 2023; Zhang et al., 2023a), multi-task learning (Chu et al., 2023), curriculum learning (Hu et al., 2024), constructing speech instruction fine-tuning datasets (Tang et al., 2023; Wang et al., 2023). However, as illustrated in Figure 1, these approaches primarily concentrate on *Modality Conversion* paradigm, which addresses the superficial relationship between the inputs and

[1]Our dataset and code will be available upon acceptance.

outputs of different modalities. It often leads to the neglect of the deeper semantic alignment, which is essential for ensuring that both speech and text embeddings convey equivalent meanings.

Motivated by these findings, we argue that *Modality Alignment* paradigm which aligns speech and text representation is crucial for further improving the performance on ST tasks. To achieve this, we introduce optimal transport (OT) theory (Peyré et al., 2019) to capture the fine-grained representation differences between speech and text. Additionally, we propose a novel **A**daptive **I**nner **S**peech-**T**ext **A**lignment **(AI-STA)** method that dynamically selects specific layers within LLM to align speech and text representations. Experiments conducted on speech translation (ST) demonstrate that our method effectively improves the translation ability of LSM. Our main contributions are summarized as follows:

- We first explore the impact of the inner layer alignment between speech and text modalities in LLMs.
- We propose an innovative adaptive speech-text alignment method to bridge the modality gap in specific selected layers and improve the performance of ST.
- Extensive experiments demonstrate that **AI-STA** outperforms the previous state-of-the-art (SOTA) methods (Chu et al., 2024) on the CoVoST2 (Wang et al., 2021) dataset in two translation directions.

2 Related Work

2.1 LLM-based Speech Translation

LLM demonstrate in-context learning (ICL) capabilities through large-scale data training, making them effective tools for solving ST tasks (Sethiya and Maurya, 2024). Inspired by the aforementioned advantages, recent studies have leveraged the capabilities of LLMs to address a variety of downstream speech tasks. The prevailing method involves feeding discretized speech units into the LLM and expanding its vocabulary to enable understanding and generation of speech (Rubenstein et al., 2023; Zhang et al., 2023a; Wang et al., 2024b). Another common approach is to connect a speech encoder to a backbone LLM, enabling effective processing of speech inputs Chu et al., 2023; Du et al., 2023; Chu et al., 2024; Hu et al., 2024; Fang et al., 2024. These models support a wide range of multi-modal speech tasks while achieving comparable performance with task-specific ST models.

LST (Zhang et al., 2023c) employs Wav2vec 2.0 (Baevski et al., 2020) as the fronted speech encoder and Llama 2 (Touvron et al., 2023) as LLM, achieving high performance on the MuST-C dataset (Di Gangi et al., 2019). (Huang et al., 2023) further incorporates the Chain-of-Thought (CoT) (Wei et al., 2022), enabling a step-by-step approach using LLMs. LLaST (Chen et al., 2024) proposed a dual-LoRA optimization strategy rendering it a strong baseline for the CoVoST2 (Wang et al., 2021) in X->En translation direction.

However, these studies primarily focus on modality conversion, while the intrinsic semantic correlation between input speech and its transcript text has not been fully exploited. In this work, we emphasize the role of modality alignment between input speech and text and propose explicit supervision signals to guide the model in learning their underlying semantic relationships.

2.2 Speech-Text Cross-Modality Alignment

Cross-modal alignment aims to establish semantically consistent mappings between different modalities (Liang et al., 2022). Early cross-modal alignment methods for speech and text modalities were mostly implicit, relying on parameter-sharing encoding mechanisms and performing multi-task learning on paired speech-text data to align the speech and text spaces (Ao et al., 2021; Bapna et al., 2021; Tang et al., 2022). Additionally, various approaches have been proposed to address modality differences by designing different loss functions and training objectives, such as connectionist temporal classification (Liu et al., 2020; Wang et al., 2020; Xu et al., 2021), contrastive learning (Ye et al., 2022; Ouyang et al., 2022; Fang et al., 2022), adversarial learning (Alinejad and Sarkar, 2020), and optimal transport (Zhou et al., 2023; Le et al., 2023; Tsiamas et al., 2024). These methods have primarily been explored within the encoder-decoder architecture and applied to the embedding or encoder layers.

Recent works have explored the alignment between speech and LLMs' text embeddings. For example, (Wang et al., 2024a) employ CFormer to address the speech-text length mismatch and introduce a KL-divergence loss to enhance the alignment of output distributions. (Nguyen et al.,

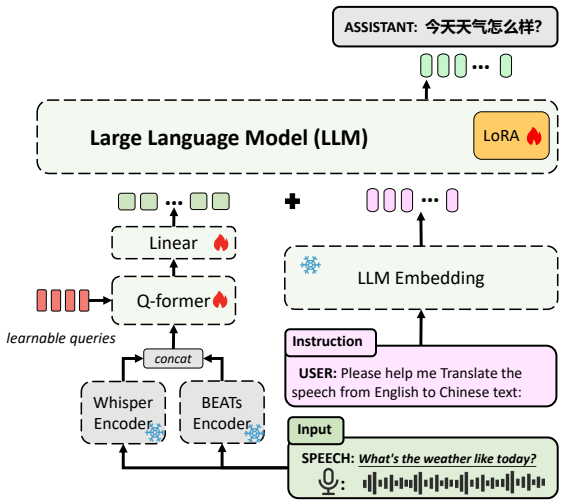


Figure 2: Model Architecture of Our LSM.

2025) conducts continuous training using mixed speech and text sequences, enabling the model to effectively learn cross-modal tasks.

In this study, we extend the cross-modal alignment to the decoder-only architecture and introduce a cross-modal retrieval task to investigate whether different hidden layers can effectively contribute to representation alignment.

3 Preliminary

3.1 Model Architecture

As illustrated in Figure 2, the model architecture in this study is identical to SALMONN (Tang et al., 2023). We use the Whisper-Large encoder (Radford et al., 2023) as the speech encoder, BEATs (Chen et al., 2022) as the audio encoder, and employ Vicuna-13B-v1.1 (Chiang et al., 2023) or Qwne2-7B (Yang et al., 2024a) as the backbone LLM. Q-Former (Li et al., 2023) serves as the connection module followed by a linear layer to project speech representation to the text representation space. The output sequence integrated with the text instructions will be fed into the LLM with LoRA adapters (Hu et al., 2021) to generate the text response. LoRA as a widely used parameter-efficient fine-tuning method for LLM adaptation, introduces additional trainable parameters. The trainable parameters of our LSM include those from LoRA adapters, Q-Former, and the linear layer, while the backbone LLM and two encoders remain frozen during training.

3.2 Optimal Transportation

OT has recently been applied in ST, primarily for finding alignments between speech and text (Zhou et al., 2023), enhancing the effectiveness of speech pre-training (Le et al., 2023), or integrating the speech encoder to the text space of the machine translation (MT) model (Tsiamas et al., 2024). While previous work has concentrated on encoder architectures, we extend this approach to a decoder-only architecture. To this end, we utilize OT to integrate the speech representation space into the text representation space within LLMs.

To align a speech representation $\mathbf{h}^{\mathrm{s}} \in \mathbb{R}^{n \times d}$ with the text representation $\mathbf{h}^{\mathrm{t}} \in \mathbb{R}^{m \times d}$, we minimize their Wasserstein loss (Frogner et al., 2015) using OT theory (Le et al., 2023; Zhou et al., 2023; Tsiamas et al., 2024). We assume the mass of each position in the speech and text representations are two uniform probability distributions. The optimized objective is defined as:

$$W_\delta = \min_{\mathbf{Z}} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{Z}_{ij} \mathbf{C}_{ij}, \quad \text{s.t.} \sum_{j=1}^{m} \mathbf{Z}_{:,j} = \frac{1}{n}, \sum_{i=1}^{n} \mathbf{Z}_{i,:} = \frac{1}{m}. \tag{1}$$

The Wasserstein distance W_δ is defined as the minimum transportation cost of all possible transportation plans $\mathbf{Z}$. $\mathbf{C}$ represents a squared euclidean cost between two vectors, where $\mathbf{C}_{ij} = ||h_i^s - h_j^t||^2$.

4 Methodology

4.1 Speech Pre-training Stage

To enable LLM to initially comprehend speech inputs and mitigate the discrepancy between pre-trained parameters and randomly initialized parameters, we utilize extensive datasets focusing on recognition and annotation tasks. This phase aims to establish a foundational ability to handle spoken language rather than deeply understanding the textual content within these speeches.

Let $\mathbf{S}$ represent the speech input and $\mathbf{T}$ as its corresponding target text sentence. The speech encoder and audio encoder transform the speech input $\mathbf{S}$ into representations $\boldsymbol{R}'$ and $\boldsymbol{R}''$, respectively. Since both encoders have the same output frame rate of 50Hz, we finally get $\mathbf{R}$ by a frame-by-frame concatenation operation along the feature dimension. Then we use the window-level Q-Former (Tang et al., 2023) to segment $\mathbf{R}$ into L-sized window representations and outputs textual tokens $\boldsymbol{E}^{\boldsymbol{S}}$. The main training objective for the speech-text pair ($\mathbf{S}$, $\mathbf{T}$) is:

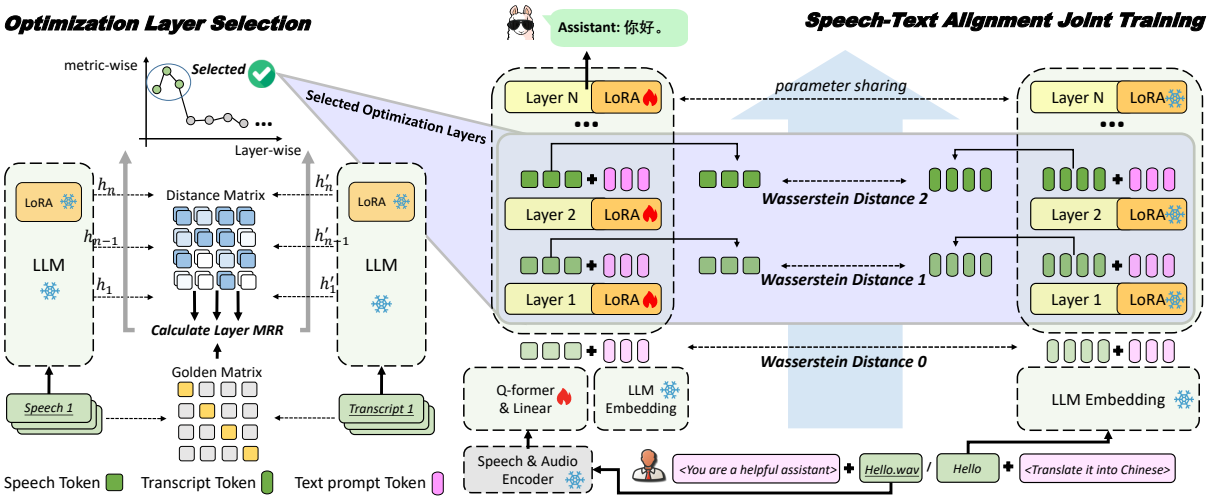


Figure 3: Overview of second and third stages of the proposed AI-STA. The **left** part first chose specific layers within the LLM according to its cross-modal retrieval ability. Then the **right** part obtains hidden states by separately forwarding speech or transcribed text concatenated with the same prompts and optimizes the LSM by combining alignment loss (computed via Wasserstein distance between hidden states) with cross-entropy loss.

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta}\left(-\log P\left(T\,\middle|E^S, I, \theta\right)\right), \\ &= \arg\min_{\theta}\left(-\sum_{m=1}^{M}\log P(T_m|T_{<m}, E^S, I, \theta)\right),\end{aligned} \tag{2}$$

where M is the length of the target token, T_m is the m-th target token, and I is the embedding of task instruction. We employed the standard causal language modeling loss as our training loss, which is designed to predict the subsequent token based on the previous token. We using the same prompt template as described by SALMONN (Tang et al., 2023) for Vicuna and Qwen2-Audio (Chu et al., 2024) for Qwen.

4.2 Optimization Layer Selection Stage

Figure 3 **left** part depicts the process of this stage. To determine the most suitable LLM layers for representation alignment, we conduct experiments after the speech pre-training stage. We randomly sample 1,000 parallel speech-text pairs from Librispeech test-clean set (Panayotov et al., 2015). Each data pair, concatenated with the same instruction, is fed into the LLM to extract hidden states from all layers. Let $h^s_{i,l}$ denote the l-th layer and i-th sample speech representation and $h^t_{i,l}$ denote the corresponding text representation. We then compute the Wasserstein Distance (Frogner et al., 2015) for each speech-text pair, constructing a distance matrix that facilitated speech-to-text retrieval, as described by the following equation:

$$D^{(l)}_{i,j} = Wasserstein(h^s_{i,l}, h^t_{j,l}). \tag{3}$$

Specifically, we rank the text samples according to their Wasserstein Distance from the speech samples and calculate the mean reciprocal rank (MRR) of the golden match across all 1,000 samples. The metrics are expressed as follows:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{r_i}, \tag{4}$$

where Q is the number of speech-text pairs. For each speech utterance, define r_i as the actual rank of the ground-truth paired text among all text samples. Subsequently, we compute the average MRR for each layer:

$$\mathcal{I} = \{l | MRR^{(l)} > threshold, l \in [0, num\_layer]\}, \tag{5}$$

where num_layer is determined by the LLM we used. The 0-th layer corresponds to the embedding layer. As shown in Figure 4, we observe that in the Vicuna-13B version, the MRR scores remain relatively high (above 0.5) from layer 0 (the embedding layer) to layer 5. However, starting from layer 6, the scores drop sharply, falling below 0.01. Similarly, in the Qwen2-7B version, layers 0 and 1 demonstrate higher scores but exhibit a steep decline in subsequent layers. This pattern indicates that the shallow layers of LLM play a

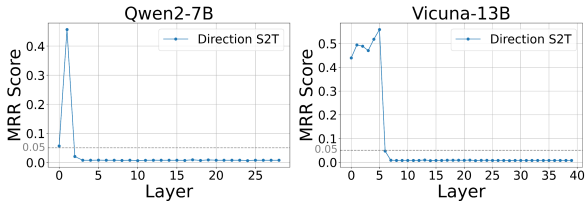


Figure 4: Layer-wise trends of average mean reciprocal rank (MRR) in two distinct backbone LLMs for speech-to-text retrieval evaluation.

crucial role in capturing the semantic properties of speech inputs. Based on empirical practice, a threshold value of 0.05 was set to select our optimal layers for optimization.

4.3 Speech-Text Alignment Joint Training Stage

As illustrated in Figure 3 **right** part, we fine-tune the model for downstream speech tasks while simultaneously training for speech-text alignment. Specifically, the transcribed text and instruction are concatenated and passed through the LLM, with their representation positions recorded to extract the corresponding text and speech representations. For the selected optimization layers, pairwise Wasserstein loss between these representations was minimized. The gradients from text representations do not contribute to the optimization process. Therefore, the final loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{CE}} &= -\log P\left(T \,\middle|\, E^S, I, \theta\right), \\ \mathcal{L} &= \alpha \mathcal{L}_{\text{CE}} + \sum_{l \in \mathcal{I}} \frac{1-\alpha}{|\mathcal{I}|} \mathcal{L}_{\text{Wass}}^{(l)}, \end{aligned} \tag{6}$$

where $\mathcal{L}_{\text{Wass}}^{(l)}$ is equivalent to the Wasserstein distance between speech and text representations in the l-th layer, and α is a hyperparameter to balance the relative importance between two loss.

5 Experiment

5.1 Training Data

In the speech pre-training stage, we use LibriSpeech training set (Panayotov et al., 2015) and GigaSpeech M-set (Chen et al., 2021) for automatic speech recognition (ASR), as well as WavCaps (Mei et al., 2024) (with audio clips longer than 180 seconds removed), AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) dataset for automatic audio captioning (AAC).

In the joint training stage, we chose the ST task for further training. CoVoST2 (Wang et al., 2021) is a large-scale multilingual dataset that supports translations between English and 15 other languages, as well as from 21 languages into English. To align with the previous method, we select two translation directions including English-Chinese and English-Japanese. All the datasets we used are listed in the Table 1.

Task	Data Source	#Hours	#Samples
ASR	LibriSpeech	960	280K
	GigaSpeech M-set	1000	680K
AAC	WavCaps	2800	370K
	AudioCaps	-	45K
	Clotho	-	4K
ST	CoVoST2 En2Zh	364	289K
	CoVoST2 En2Ja	364	289K

Table 1: Training data used in all stages.

5.2 Training Setup

Our model employs the encoder part of Whisper-Large-v2 (Radford et al., 2023) model as the speech encoder, the fine-tuned BEATs (Chen et al., 2022) encoder as the audio encoder, and a Vicuna-13B-v1.1 (Chiang et al., 2023) or a Qwen2-7B (Yang et al., 2024a) as the backbone LLM. In the Q-Former block, we set $N = 1$ for a single trainable query and use $L = 17$ which is approximately 0.33 seconds per window. The OT loss weight is empirically set to 0.01 based on practical experience. We freeze speech encoder, audio encoder, and LLM when training, leading 28 million (M) or 64M trainable parameters, depending on the scale of the backbone LLM parameters. Detailed training hyperparameters are available in Appendix A.1.

5.3 Evaluation

We evaluated the model using the CoVoST2 test set for English-Chinese and English-Japanese translations, employing SacreBLEU (Post, 2018) score as the evaluation metric. Audio samples are all resampled to 16kHz in the experiments.

5.4 Baselines

We compare our LSM and method with the following four baselines.

SALMONN (Tang et al., 2023) integrates a pre-trained text-based LLM with a speech encoder and audio encoder to process audio inputs. It excels in tasks like speech recognition, translation,

Method	CoVoST2	
	En-Zh	En-Ja
Baseline Models		
SALMONN	33.1	22.7
BLSP-KD	41.3	21.3
Qwen-Audio	41.5	23.5
Qwen2-Audio	<u>45.2</u>	<u>28.6</u>
Our LSM with Vicuna-13B-v1.1		
base	36.5	29.8
w/ AI-STA	37.6	30.2
Our LSM with Qwen2-7B		
base	45.3	31.0
w/ AI-STA	**46.0**	**31.4**

Table 2: Speech translation BLEU scores on CoVoST2. We conducted experiments in English (En)-to-Chinese (Zh), En-to-Japanese (Ja). For each result, We use <u>underline</u> to highlight the previous SOTA baseline, and use **bold** to highlight surpassing the SOTA performance.

and music captioning while showcasing emergent abilities.

BLSP-KD (Wang et al., 2024a) leverages CFormer architecture to tackle the speech-text length discrepancy, while incorporating a KL-divergence loss mechanism to optimize output distribution alignment. It also introduces a partial LoRA strategy to facilitate efficient LLM fine-tuning with speech inputs.

Qwen-Audio (Chu et al., 2023) is Alibaba's multi-modal LLM, accepting diverse audio and text inputs to output text. It proposes a multi-task learning framework and incorporates a word-level time-stamp prediction training task while yielding strong performance across various tasks.

Qwen2-Audio (Chu et al., 2024) is the latest progress of Qwen-Audio. It further boosts instruction-following capability and adopts direct preference optimization to align with human preferences achieving SOTA in AIR-Bench (Yang et al., 2024b).

6 Results

6.1 Main Result

Table 2 presents a comparison of our base LSM, our LSM with AI-STA method, and previous methods, reporting SacreBLEU scores evaluated on two language pairs: En-Zh, and En-Ja. Notably, our base LSM with Qwen2-7B achieves SOTA with the BLEU of 45.3 on En-Zh and 31.0 on En-Ja translation direction. Further with our AI-STA method, our LSM outperforms previous SOTA for 0.8 BLEU in En-Zh and 2.8 BLEU in En-Ja. Additionally, the AI-STA method provides a noticeable boost in performance for all models, with BLEU score gains of approximately 0.8 for LSM with Vicuna-13B-v1.1 and 0.6 for LSM with Qwen2-7B. We also observe that the performance gain with AI-STA is greater for En-Zh (0.9 BLEU) than for En-Ja (0.4 BLEU), suggesting that our alignment method may benefit more from target languages with richer training resources. These results convincingly demonstrate the superiority of AI-STA and highlight the promising potential of exploring LLMs for speech translation tasks.

Aligning Position	BLEU
base	45.3
Layer 0	45.7
Layer 0-5	45.5
Layer 1	X
Layer 0-1(Selected)	**46.0**

Table 3: The impact of different layer optimization selection strategies on the performance of CoVoST2 En-Zh using Qwen2-7B as backbone LLM.

6.2 Effect of Optimization Layer Selection

Tabel 3 shows that different aligning layer selections have a great impact on CoVoST2 En-Zh performance. Only aligning speech and text representation in layer 0 (embedding layer) gains 0.4 BLEU improvement. Once we further align at the inner layers, the performance begins to decline (45.7 -> 45.5). Especially when not aligning layer 0, the training loss fails to converge leading to catastrophic failure. The performance is further enhanced by aligning with selected alignment layers obtained through our optimization layer selection strategy (45.7 -> 46.0), highlighting the necessity of our layer selection strategy.

6.3 Comparison of Aligning Methods

For connectionist temporal classification (CTC) (Graves et al., 2006), we apply it at the token level using backbone LLM's tokenizer to encode the transcript of source speech as the golden token and train an independent classification layer for matching with LLM's vocabulary size. For contrastive learning (CL), we treat golden paired speech-text samples as positive pairs, and the others as negative pairs and apply a multi-class N-pair contrastive loss (Sohn, 2016). Both alignment methods only operate at the embedding layer. As

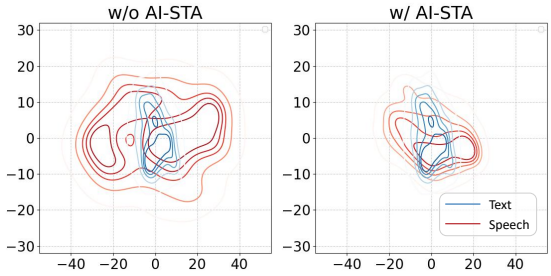


Figure 5: t-SNE visualization of speech and text representation from LSMs trained with or without AI-STA methods.

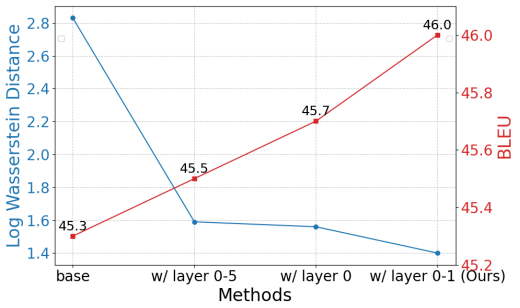


Figure 6: The correlation between alignment degree and performance in the CoVoST2 En-Zh direction across different training methods. Lower logarithmic Wasserstein distance indicates a higher degree of alignment, while a higher BLEU score corresponds to better performance in the ST task.

Aligning method	BLEU
base	45.3
↪ w/ CTC	44.6
↪ w/ CL	45.1
↪ w/ AI-STA	**46.0**

Table 4: The impact of different aligning methods on CoVoST2 En-Zh performance.

shown in Table 4, employing either CTC or CL results in performance degradation. By contrast, our method yields a 0.7 BLEU improvement.

We argue that CTC is not suitable for adapters like Q-Former that incorporate attention mechanisms. CTC is a forced alignment method where each output position must align precisely with a specific token, which can lead to conflicts when applied after the Q-Former. When applying contrastive learning (CL) such an alignment method at the overall semantic level, yields limited effectiveness and fails to further capture the fine-grained relationships between words. Both methods cause conflicting training objectives and hinder the training process.

6.4 Can AI-STA Close the Modality Gap?

We randomly sample 1,000 speech-text transcription pairs from the test set of CoVoST2 En-Zh to explore representation alignment between speech and text in the embedding layer of our LSM with the Qwen2-7B version. The speech representation is obtained as semantic tokens after processing through the Q-Former, while the text representation is derived from the tokenization and embedding layer of the LLM. All representations are averaged along the length dimension. We apply bivariate kernel density estimation (Parzen, 1962) and utilize the T-SNE technique reducing data dimensions to a two-dimensional space for visualization (Van der Maaten and Hinton, 2008). As depicted in Figure 5, AI-STA significantly reduces the distance discrepancy between the speech and text representation spaces compared to the baseline without our method, demonstrating a strong relationship between these two modalities.

6.5 Correlation between Alignment Degree and ST Performance

To investigate the relationship between representation alignment and speech translation performance, we calculate alignment scores for LSMs trained with various alignment methods. Taking the LSM with Qwen2-7B as an example, our optimal layer selection strategy identifies the zero and first layers as particularly suitable for representation alignment. We quantify the degree of alignment by calculating the logarithmic Wasserstein distance between these two layers. Figure 6 illustrates a strong correlation between the alignment score and speech translation performance. From left to right, the alignment methods are the same with section 6.2 except for the optimization applied to layer 1. As the alignment scores decrease, we consistently observe a steady increase in the BLEU score, indicating a strong correlation between improved alignment and enhanced translation performance.

6.6 Can AI-STA Help Knowledge Transfer?

To investigate whether our method can bridge the modality gap and enable the model to understand speech modality inputs as text modality. We directly perform zero-shot text translation inference on a model that has been trained on the speech translation task. The inference prompt is identical to the training prompt. We intended to verify

Method	ST	MT
Training on ST		
base(Vicuna-13B)	36.5	34.0
↪ w/ AI-STA	37.6	39.4
base(Qwen2-7B)	45.3	51.2
↪ w/ AI-STA	**46.0**	**51.5**

Table 5: The impact of Speech-Text Alignment on zero-shot machine translation task. Demonstrates that our method facilitates knowledge transferring from speech to text modality.

whether the model has effectively utilized the correspondence between speech and text during the training process.

As shown in Table 5, we observe that in the Vicuna-13B version, the zero-shot performance gap between using and not using our method reaches up to 5.4 BLEU. This indicates that our method significantly enhances the LLM's ability to leverage ST knowledge for text-based MT tasks. In the Qwen2-7B version, the zero-shot performance gap shrinks to 0.3 BLEU. Irrespective of whether our method is applied, the translation performance in text scenarios is significantly stronger than that in speech scenarios. We attribute this phenomenon to the Qwen2-7B model's strong English and Chinese language capabilities, as well as its more precise capture of the relationship between speech and text modalities. We use the transcript and translation text pair of the CoVoST2 En-Zh test set as our MT evaluation data.

6.7 Case Study

In this section, we present several cases generated by our LSM with Vicuna-13B to compare its performance with the previous end-to-end model, SALMONN (Tang et al., 2023). The results are summarized in Figure 7. In the first case, SALMONN incorrectly translates "in no way unique" as the meaning of "nothing unique", leading to a deviation from the intended meaning. Our LSM inaccurately translates it as meaning "completely normal" which is out of context. While training with our AI-STA method, our LSM accurately translates it to the correct answer.

In the second case, SALMONN and our LSM exhibit different translation errors in this case. SALMONN fails to translate the word "portraying" as "representation". Our LSM, in turn, misinterprets the word "mascot", resulting in a significant misunderstanding. In contrast, our LSM with AI-STA correctly translates these words,

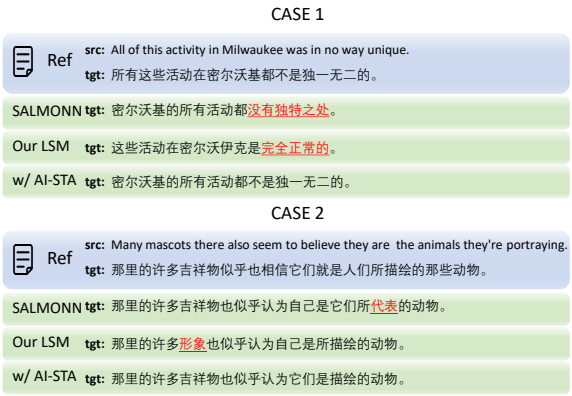


Figure 7: CoVoST2 En-Zh test cases that generated from the SALMONN, our LSM with Vicuna-13B and our LSM with AI-STA. The red underlined text indicates an incorrect answer.

yielding a more accurate overall translation than SALMONN and our base LSM. These observations highlight the ST capabilities of our LSM with AI-STA, demonstrating that AI-STA enhances the LSM's ability to capture fine-grained semantics.

Although our method improves the performance of LLM-based ST, we still discovered content omission during the translation generation process in the two cases mentioned above. Some source words remain untranslated, such as "all of this" in the first case, which is not fully translated by any of the methods. As a result, the translated sentences tend to be shorter than the reference sentences. This highlights a persistent issue in current LLM-based Speech Translation systems, suggesting that there is still room for improvement.

7 Conclusions

In this study, we enhance the ST capabilities of LSMs by explicitly aligning speech and text representations. To achieve this, we introduce OT theory to quantify the discrepancy between speech and text representations and investigate the representation characteristics of different layers within LLMs. By leveraging the cross-modal retrieval technique, we identify specific model layers that are well-suited for representation alignment and perform joint training using these selected layers. Our experiments demonstrate that this method effectively reduces the distance between the speech and text representation spaces, enabling the model to better capture the relationships between the two modalities and significantly improves the performance of large speech models on speech translation tasks.

Limitation

We acknowledge that our proposed approach has several limitations: (1) We observed several intriguing phenomena, such as performance degradation when applying CTC or CL alignment methods at the embedding layer, as well as a sharp drop in retrieval performance at certain layers within LLM and remained low in subsequent layers. However, we did not thoroughly investigate the underlying principles and instead relied on intuition and empirical observations without theoretical justification or formal proof. (2) Although our method enhances the LLM-based ST performance and reaches SOTA, a performance gap remains compared to the text scenarios' machine translation. However, we believe our method provides valuable insights and encourages the development of cross-modal learning in LLMs.

References

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8014–8020.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*.

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.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Xi Chen, Songyang Zhang, Qibing Bai, Kai Chen, and Satoshi Nakamura. 2024. Llast: Improved end-to-end speech translation system leveraged by large language models. *arXiv preprint arXiv:2407.15415*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 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*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. 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 2012–2017. Association for Computational Linguistics.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

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

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni:

Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

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*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

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*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. 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 119–132.

Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, and Didier Schwab. 2023. Pre-training for speech translation: Ctc meets optimal transport. In *International Conference on Machine Learning*, pages 18667–18685. PMLR.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, et al. 2025. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. Waco: Word-aligned contrastive learning for speech translation. *arXiv preprint arXiv:2212.09359*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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*.

Nivedita Sethiya and Chandresh Kumar Maurya. 2024. End-to-end speech-to-text translation: A survey. *Computer Speech & Language*, page 101751.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

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, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, et al. 2022. Unified speech-text pre-training for speech translation and recognition. *arXiv preprint arXiv:2204.05409*.

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

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2024. Pushing the limits of zero-shot end-to-end speech translation. *arXiv preprint arXiv:2402.10422*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech*, pages 2247–2251.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, and Jiajun Zhang. 2024a. Blsp-kd: Bootstrapping language-speech pre-training via knowledge distillation. *arXiv preprint arXiv:2405.19041*.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9161–9168.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2024b. Viola: Conditional language models for speech recognition, synthesis, and translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. *arXiv preprint arXiv:2105.05752*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, et al. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024b. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. *arXiv preprint arXiv:2205.02444*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

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

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. Cmot: Cross-modal mixup via optimal transport for speech translation. *arXiv preprint arXiv:2305.14635*.

A Appendix

A.1 Hyperparameters

The training configurations for the two stages are summarized as follows:

Speech Pre-training Stage: Training employs the AdamW optimizer with hyperparameters β_1=0.9, β_2=0.999, $eps = 1e^{-8}$. The learning rate follows a cosine decay schedule, starting with a warm-up rate of $1e^{-6}$, peaking at $3e^{-5}$, and decaying to a minimum of $1e^{-5}$. Weight decay is set to 0.05, and the global batch size is 32. The model undergoes 80k training steps with 9k warm-up steps, using BFLOAT16 numerical precision. LoRA parameters include a rank of 8, alpha of 32, and dropout of 0.1.

Joint Training Stage: The training configuration

for this stage is largely the same as the mentioned above, with two differences: the warm-up steps in this stage are set to 3k, and we do not fix the total number of training steps. Instead, we determine whether to stop training based on the metrics from the validation phase conducted every 3k training steps. Training is stopped when the validation accuracy does not exceed the previous highest value for four consecutive validation phases.