

KIT's Offline Speech Translation and Instruction Following Submission for IWSLT 2025

Sai Koneru^{*♣}, Maïke Züfle^{*♥}, Thai-Binh Nguyen, Seymanur Akti, Jan Niehues, Alexander Waibel

Karlsruhe Institute of Technology

firstname.lastname@kit.edu

Abstract

The scope of the International Workshop on Spoken Language Translation (IWSLT) has recently broadened beyond traditional Speech Translation (ST) to encompass a wider array of tasks, including Speech Question Answering and Summarization. This shift is partly driven by the growing capabilities of modern systems, particularly with the success of Large Language Models (LLMs). In this paper, we present the Karlsruhe Institute of Technology's submissions for the Offline ST and Instruction Following (IF) tracks, where we leverage LLMs to enhance performance across all tasks. For the Offline ST track, we propose a pipeline that employs multiple automatic speech recognition systems, whose outputs are fused using an LLM with document-level context. This is followed by a two-step translation process, incorporating an additional refinement step to improve translation quality. For the IF track, we develop an end-to-end model that integrates a speech encoder with an LLM to perform a wide range of instruction-following tasks. We complement it with a final document-level refinement stage to further enhance output quality by using contextual information.

1 Introduction

This paper provides an overview of the systems submitted by the Karlsruhe Institute of Technology (KIT) to the [Offline Speech Translation](#) (ST) and the [Constraint Long Instruction-Following](#) (IF) tasks of IWSLT 2025. For the Offline track, we participate in the unconstrained setting for the *English→German* language pair. For the IF task, we participate in the constrained-long track, aiming to perform Automatic Speech Recognition (ASR), Speech Translation (ST), Spoken Question Answering (SQA), and Speech Summarization (SSUM) across various languages.

A growing research trend in the field is the application of Large Language Models (LLMs) to speech processing tasks ([Tang et al., 2023](#); [Züfle and Niehues, 2024](#); [Chu et al., 2024b](#); [Abouelenin et al., 2025](#), among others), leveraging their strong general knowledge and natural language understanding capabilities. These strengths make LLMs particularly relevant to both the Offline ST and IF tracks. Accordingly, in our submissions, we explore strategies for effectively integrating LLMs into speech processing pipelines.

There are multiple approaches to leveraging LLMs in speech systems. One strategy involves incorporating LLMs as an additional step within a cascaded architecture ([Koneru et al., 2024a](#)), where they can perform task-specific refinement. This modular approach allows each component to be trained independently, benefiting from specialized data. Alternatively, LLMs can be integrated in an end-to-end fashion ([Tang et al., 2023](#); [Züfle and Niehues, 2024](#); [Chu et al., 2024b](#); [Abouelenin et al., 2025](#)), allowing for better information flow and potentially improving generalization to unseen tasks.

Although both the Offline and IF tasks fall under the umbrella of speech processing, they differ significantly in nature. In the offline setting, speed and adaptability to unseen tasks are not primary concerns. In contrast, the IF task demands flexibility and generalization, as the system must handle a variety of instructions. This has an impact on the architectures we choose for the different tracks.

For the Offline track, we utilize LLMs specialized on a specific task as refinement modules within a cascaded architecture. This is common practice; all systems submitted to IWSLT 2024 for this track employed a cascaded architecture ([Ahmad et al., 2024](#)), underlining its practical advantages in training the system due to availability in data, e.g for low-resource languages ([Liu et al., 2023](#)), and simplicity by decomposing into smaller tasks.

For the IF track, training a dedicated cascaded

^{*} Equal Contribution

[♣] Offline, [♥] Instruction-Following

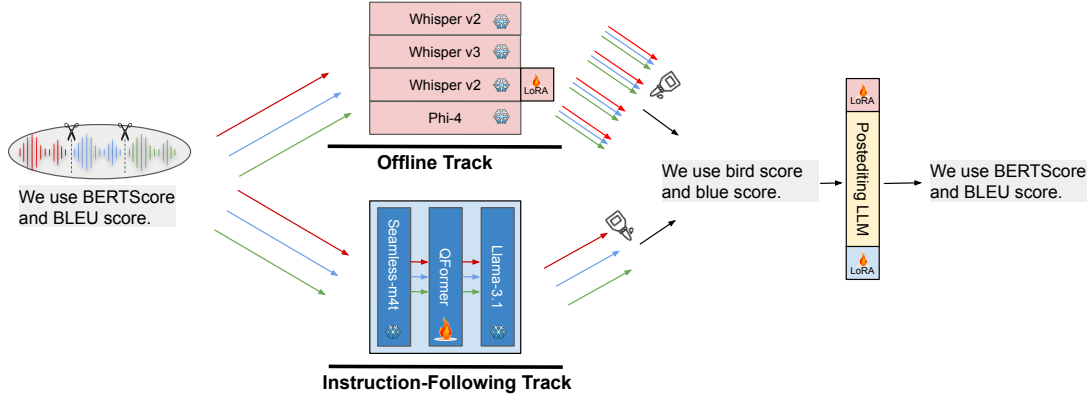


Figure 1: For the Instruction-Following track, we train an end-to-end SpeechLLM, while the Offline track relies on an ensemble of existing models. To enhance the outputs from both tracks, we apply a post-editing model that provides two main benefits: correcting scientific terminology and recovering context that may be lost due to the segmentation of long audio sequences.

system for each task is not an efficient solution, moreover, the goal of this track is to build a model that can follow different instructions. Consequently, we adopt an end-to-end approach using a Speech Large Language Model (SpeechLLM). Nevertheless, for tasks such as ASR, ST, and SSUM, we also include an additional refinement step to enhance fluency and contextual consistency in the output.

An overview of both systems exploiting LLMs internally or via post-editing for refinement, can be found in Fig. 1. We describe the details of each system in the following sections. First, we present the Offline ST track system in Section 2. Then, we discuss the IF track system in Section 3.

2 Offline Track

The goal of the Offline ST track is to generate high-quality translations across diverse domains without latency constraints. Recent work has highlighted the potential of LLMs for this task (Ahmad et al., 2024; Koneru et al., 2024a). Building on these insights, we integrate LLMs at multiple stages of our speech translation pipeline. Below, we present a high-level overview, with each component detailed in the following sections.

We begin with long-form audio inputs, which may span several minutes to hours. Due to memory limitations and the lack of training data for such durations, our ASR and MT systems cannot handle these directly. Thus, we first segment the audio into manageable chunks using a Voice Activity Detection (VAD)-based method, which is effective even in noisy conditions.

The segmented audio is then transcribed into En-

glish using ASR. Rather than relying on a single model, we adopt a fusion strategy, combining outputs from multiple ASR systems—including both pre-trained models and a fine-tuned variant. This approach, akin to model ensembling, leverages the complementary strengths of different systems to reduce errors.

We fuse the ASR outputs using an LLM, which processes the combined hypotheses at the document level. This allows for the incorporation of broader context, resulting in more coherent and accurate transcriptions.

The English text is then segmented into sentences using the `nltk` tokenizer and translated into German. For this, we fine-tune a translation LLM on high-quality parallel data. To ensure quality, we use a quality estimation model to filter out noisy sentence pairs, keeping only high-confidence examples.

Finally, both the source transcript and the machine-translated output are passed to an Automatic Post-Editing (APE) model. This model refines the translations, producing polished final outputs.

2.1 Segmentation

The segmentation module breaks long-form audio into manageable segments for the ASR pipeline. We explored two strategies: fixed-window chunking and content-aware segmentation.

Fixed-window chunking applies a uniform sliding window and relies on transcript overlap to stitch adjacent chunks. While effective on clean audio, it often fails in noisy settings like the ITV or EPTV datasets, leading to fragmented or duplicated text.

Content-aware segmentation uses audio cues to find natural cut points. Basic methods rely on VADs like Silero (Team, 2021) or py-webrtcvad (Wiseman, 2019), which work well in clean conditions but struggle with noise. Instead, we use an end-to-end speaker segmentation model from Bredin and Laurent (2021), trained for noisy scenarios and capable of tracking up to three speakers. While methods like SHAS (Tsiamas et al., 2022) use wav2vec embeddings, they underperform in the presence of background noise.

Even with smarter cut-point detection, uncontrolled segment lengths can hurt ASR performance. Inspired by WhisperX (Bain et al., 2023), we enforce length constraints by post-processing VAD segments: overly long segments are split at their lowest-confidence point, while overly short ones are merged with neighbors (even across non-speech gaps) until they reach the desired duration *Chunk Size*.

Chunk Size	<i>Peloton</i>	<i>EPTV</i>	<i>ITV</i>	<i>ACL</i>
5	13.62	15.79	21.49	14.38
10	12.61	14.63	18.8	12.03
15	12.23	14.08	17.71	11.43
20	12.27	13.98	17.29	11.71
25	11.98	13.98	16.62	11.49

Table 1: Impact of chunk size during segmentation for ASR. We report the WER scores using Whisper-v3 with different chunk sizes. Best scores for each test set are highlighted in **bold**.

To determine the optimal *chunk size*, we perform a grid search using test sets from various domains, with results shown in Table 1. We use the **Whisper v3 model**¹ (Radford et al., 2023) and evaluate it on the Peloton, EPTV, and ITV subsets from the IWSLT 2024 development sets (Ahmad et al., 2024), as well as the ACL 60/60 test set (Salesky et al., 2023). A chunk size of 25 consistently yields the best performance. We hypothesize that this is due to the larger chunk size offering more contextual information, aligning with prior work on the benefits of long-form decoding in noisy conditions (Koneru et al., 2024a; Yan et al., 2024).

2.2 Automatic Speech Recognition

After segmenting the audio into smaller chunks, we send them to the ASR system for transcription.

¹openai/whisper-large-v3

Since we participated in the language direction *English*→*German*, the audio needs to be transcribed in English, a high-resource language. Many publicly available pre-trained models excel at English transcription, and we first evaluated several of them individually. Specifically, we considered the Whisper variants v2² and v3³ (Radford et al., 2023), as well as the recently developed multimodal LLM Phi-4⁴ (Abouelenin et al., 2025).

To build a robust model for noisy scenarios, such as those found in TV series, we further fine-tuned Whisper Large v2 on the Bazinga dataset (Lerner et al., 2022). The Word-Error-Rate (WER) for these models on ITV and ACL 60/60 are reported in Table 2.

Model	<i>ITV</i>	<i>ACL</i>
Whisper v2	17.04	11.55
Whisper v2 + Bazinga	16.87	11.23
Whisper v3	16.62	11.49
Phi-4	20.64	9.71
LLM-Fuse	17.03	10.77

Table 2: WER scores of ASR models on the ITV and ACL test sets. LLM-Fuse indicates the post-edited output of all ASR systems at document-level. Best scores for each test set are highlighted in **bold**.

As shown in Table 2, there is no clear winner across the two test sets. Our manual analysis further reveals that different models tend to make different types of errors, suggesting that combining these systems could be a promising strategy.

2.2.1 Fusing with LLM

To fuse the ASR outputs, token-level ensembling is a viable approach—provided the vocabularies of the systems are compatible. However, the vocabulary used by Phi-4 differs from that of the Whisper variants, limiting the effectiveness of this method. Alternative techniques such as re-ranking offer some promise but are unable to leverage document-level context.

To overcome these limitations, we employ an LLM to generate the final transcript based on the outputs from individual ASR systems. Thanks to their ability to process long contexts, LLMs enable us to concatenate hypotheses from multiple chunks and refine them collectively.

²openai/whisper-large-v2

³openai/whisper-large-v3

⁴microsoft/Phi-4-multimodal-instruct

However, an off-the-shelf LLM may not perform optimally for this specific task. To improve, we propose fine-tuning the model using a dataset generated through data augmentation. For this purpose, we use monolingual English text from the Europarl v7 and v10 datasets (Koehn, 2005), News-Commentary v16, OpenSubtitles (Lison and Tiedemann, 2016), and the NUTSHELL dataset⁵ (Züfle et al., 2025). With the exception of NewsCommentary, the other datasets contain document-level structure—episodes in the case of OpenSubtitles and abstracts in the case of NUTSHELL.

We then employ the Text-to-Speech model VITS (Kim et al., 2021) to synthesize audio from the selected texts. This generated audio is subsequently transcribed using Phi-4 and the Whisper variants. As a result, we obtain ASR hypotheses for the synthesized speech along with their corresponding ground-truth references.

Next, we convert this data into a prompt format, as described in Appendix App. A. We fine-tune the LLM Llama 3 8B⁶ (Grattafiori et al., 2024) using LoRA (Hu et al., 2022), training it to predict the reference transcription given the hypotheses produced by the different ASR systems. We illustrate this in Fig. 1. We also report the ASR performance of the LLM fusion approach in Table 2 and observe that it does not outperform the individual systems. However, as we demonstrate in the following sections, this fusion proves to be highly beneficial when computing the final ST scores.

2.3 Speech Translation

The next step in the pipeline, after performing ASR, is to translate the transcriptions into German. Since the transcriptions are produced at the chunk level, they often contain multiple sentences, some of which may be incomplete. To address this, we first concatenate all the text from a given talk and then segment it into sentences using the NLTK tokenizer. This ensures that only complete sentences are passed to the MT system, aligning with the way such systems are typically trained.

2.3.1 Gold vs ASR Transcripts

Recently, several translation-focused LLMs have been introduced, demonstrating strong performance on high-quality input (Xu et al., 2024a; Alves et al., 2024). However, their effectiveness on noisy input—such as ASR-generated tran-

⁵Our submission is unconstrained by using this dataset.

⁶meta-llama/llama-3-8B

Model	Chrf2 (↑)	MetricX (↓)	COMET (↑)
Gold Transcript ●			
Tower 7B	68.7	2.02	83.31
GemmaX2 9B	70.5	2.08	83.62
Whisper v3 ASR (Chunk size=25) ●			
Tower 7B	66.1	2.46	81.01
GemmaX2 9B	66.4	2.65	80.74
Phi-4 ASR ●			
Tower 7B	64.9	2.73	79.25
GemmaX2 9B	65.4	2.9	79.12

Table 3: Translation quality comparison between Gold and ASR transcripts on the ACL 60/60 test set. Note that higher is better for chrf2 and COMET scores and lower for MetricX scores.

scripts—remains uncertain. To assess this, we first evaluate the out-of-the-box translation quality of two leading models: Tower⁷ (Xu et al., 2024a) and GemmaX2⁸ (Cui et al., 2025). We use the COMET⁹ (Rei et al., 2022a), MetricX¹⁰ (Juraska et al., 2024), and ChrF2 (Popović, 2015) metrics, with results reported in Table 3 for the ACL 60/60 test set.

GemmaX2 outperforms Tower on gold transcripts in terms of COMET scores, but its performance drops significantly on ASR-generated input. Interestingly, translation quality is lower when using transcripts from the Phi-4 ASR model, despite it having the lowest WER in Table 2. We hypothesize that this is due to inconsistencies in punctuation and casing, which are not captured by WER but can impact translation quality. This highlights that lower WER does not always correlate with better translations. As a result, we choose Tower 7B as our base model for subsequent enhancements, given its superior robustness to noisy input.

2.3.2 Quality-Filtered Finetuning for MT

Tower 7B is a multilingual model and we only focus on English → German in our submission. Therefore, we adapt it to this specific language pair. While plenty of data is available for fine-tuning, these also include low quality translation pairs.

Recent studies have demonstrated the importance of high-quality data during fine-tuning (Finkelstein et al., 2024; Ramos et al., 2024; Xu et al., 2024b). To this end, we leverage the Europarl

⁷Unbabel/TowerInstruct-7B-v0.2

⁸ModelSpace/GemmaX2-28-2B-v0.1

⁹Unbabel/wmt22-comet-da

¹⁰google/metricx-24-hybrid-xl-v2p6

Model	ITV		ACL		
	Chrf2 (↑)	MetricX (↓)	Chrf2 (↑)	MetricX (↓)	COMET (↑)
Whisper v3 ASR (Chunk size=25) ●					
Tower 7B	41.4	4.25	66.1	2.46	81.01
Tower 7B Finetuned	41.5	4.19	67.7	2.27	82.05
LLM-Fuse ○					
Tower 7B Finetuned	41.7	4.12	68	2.01	83.07
Tower 7B Finetuned + Tower 13B APE	42.1	4.03	69.6	1.84	83.31

Table 4: Analysis of translation quality of our ST system with different enhancements on the ITV and ACL test sets. Note that higher is better for chrf2 and COMET scores and lower for MetricX scores. Best scores for each metric per test set are highlighted in **bold**.

v7 and v10 datasets (Koehn, 2005), NewsCommentary v16, and OpenSubtitles (Lison and Tiedemann, 2016) to extract high-quality translation pairs. We employ the XCOMET¹¹ quality estimation model (Guerreiro et al., 2024) to rank the translation pairs and select the top 500k based on quality scores. Tower 7B is then fine-tuned on this curated dataset using LoRA adapters (Hu et al., 2022), adapting it for generating German translations.

2.3.3 Automatic Post-Editing Translations

As a final step, we aim to correct translation errors through APE (Koneru et al., 2024b). To achieve this, we fine-tune Tower 13B¹² on a synthetically generated APE dataset. Using our previously fine-tuned model, we generate 100k (*source*, *hypothesis*, *reference*) triplets by sampling a subset from the top 500k high-quality sentence pairs. Then, we transform into the prompt format as shown in App. A. We choose the larger 13B model for this task, as we expect it to be adaptable to correct the output with limited fine-tuning. To train within resource constraints, we follow the same approach as before and fine-tune using LoRA adapters.

We present an overview of the ST scores in Table 4 for the ITV and ACL 60/60 test sets. The results show that fusing system hypotheses using an LLM leads to improved ST performance on both test sets (from 4.19 → 4.12 for ITV and 2.27 → 2.01 for ACL in MetricX). Additionally, applying Automatic Post-Editing (APE) further enhances translation quality. As a result, our final pipeline integrates multiple ASR systems fused via an LLM, followed by initial translation generation and post-editing to ensure high-quality output.

¹¹Unbabel/XCOMET-XL

¹²Unbabel/TowerInstruct-13B-v0.1

2.4 Future Directions and Potential Improvements

There are several potential avenues for improving our approach in future iterations of the shared task. First, while we did not explore it in this work, it is unclear how well SHAS segmentation performs when trained on noisy data. Semantic segmentation of noisy inputs could yield performance gains. Second, incorporating LLM specific to the target language (e.g. German LLM) for APE at the document level could offer promising improvements. Lastly, we experimented with Quality-Aware Decoding (Koneru et al., 2025), which showed benefits primarily when the quality of the ASR output was high. Future research could focus on adapting the quality estimation component to perform robustly under noisy or imperfect segmentation conditions.

3 Instruction Following Long Track

The Instruction-Following (IF) Speech Processing track in the scientific domain aims to benchmark foundation models that can follow natural language instructions—an ability well-established in text-based LLMs but still emerging in speech-based counterparts. The track covers four tasks: Automatic Speech Recognition (ASR), Speech Translation (ST), Spoken Question Answering (SQA), and Spoken Summarization (SSUM). ASR is evaluated on English, ST on English → German, Chinese, and Italian (en→{de, it, zh}), and SQA/SSUM across all four directions (en→{en, de, it, zh}).

We participate in the Constrained Long track, which focuses on long-form speech inputs (5–10 minutes). This track enforces limitations on both model selection and training data. Specifically,

only SeamlessM4T-Large¹³ (Communication et al., 2023) and LLaMA-3.1-8B-Instruct¹⁴ (Grattafiori et al., 2024) are permitted as base models.

Our approach employs an end-to-end speech model trained under these constraints, enhanced with a post-editing stage for improved output quality similar to the Offline track.

3.1 Data

Data in the Constrained Setting For ASR and ST, the provided datasets include EuroParl-ST (Iranzo-Sánchez et al., 2020) and CoVoST 2 (Wang et al., 2020). For the SQA task, the only resource available is the extractive Spoken-SQuAD (Lee et al., 2018). For SSUM, NUTSHELL (Züfle et al., 2025), an abstract generation dataset for scientific talks, is provided. As development data, the ACL 60/60 benchmark (Salesky et al., 2023) is made available. Notably, the only in-domain datasets, i.e., those based on scientific talks, are NUTSHELL and ACL 60/60. Moreover, no multilingual data is provided for SQA and SSUM.

Data Augmentation To address the limitations of the constrained setting, we apply task-specific data augmentation strategies¹⁵:

ASR: To introduce domain-specific data, we augment the ASR training data using scientific abstracts from NUTSHELL (Züfle et al., 2025). The abstracts are split into sentences with `nltk` and then converted to synthetic speech using SeamlessM4T-Large.

ST: We do not augment the ST training data, but construct an artificial en-it test set for the ACL 60/60 dataset, which lacks Italian. We translate the English ACL 60/60 transcripts into Italian using both SeamlessM4T-Large and LLaMA-3.1-8B-Instruct, and evaluate translation quality using COMETKiwi (Rei et al., 2022b). SeamlessM4T-Large achieves a slightly higher score (82.55 vs. 81.07), and is therefore used to generate the final test set translations. The translation prompts for LLaMA-3.1-8B-Instruct are detailed in App. B.3.

SQA: For SQA, we aim to: (1) support all language pairs, (2) adapt to the scientific domain, and (3) include abstractive QA, as required by the track. Therefore, we transcribe NUTSHELL dev talks using SeamlessM4T (audio split into 15-second

chunks at silence regions). We then use LLaMA-3.1-8B-Instruct to generate two answerable and one unanswerable QA pair per segment for all language pairs. We balance the dataset by ensuring that unanswerable questions comprise 5% of the final set. Additionally, we generate a 250-sample test set from a subset of the NUTSHELL test data. Prompt templates are included in App. B.1

SSUM: To enable multilingual evaluation of speech summarization, we translate the full NUTSHELL dataset ($\text{en} \rightarrow \{\text{de}, \text{it}, \text{zh}\}$) using LLaMA-3.1-8B-Instruct. Prompt details are provided in App. B.2. As with SQA, we also generate a 250-sample multilingual test set.

3.2 Model

In the constrained setting of the track, only the speech foundation model SeamlessM4T-Large¹³ (Communication et al., 2023) and LLaMA-3.1-8B-Instruct¹⁴ (Grattafiori et al., 2024) are permitted.

Architecture To integrate the speech encoder and LLM in an end-to-end architecture, we use Q-Former (Li et al., 2023; Tang et al., 2024) as a projector. Specifically, we use a four transformer layers and four learnable query tokens to bridge the modality gap between the features from SeamlessM4T and LLaMA. During training, only the projector is trained and the speech encoder and LLM remain frozen.

Training We explore three training strategies: (1) Direct fine-tuning on all available training data, (2) ASR pretraining followed by fine-tuning, and (3) contrastive pretraining, as proposed by Züfle and Niehues (2024), followed by fine-tuning.

For contrastive pretraining, we use ASR data and experiment with cosine similarity and Wasserstein loss functions (Peyré and Cuturi, 2019; Le et al., 2023). As shown in Table 5, contrastive pretraining yields notable improvements over the other training strategies. Consequently, this approach is adopted for the final model submissions. Hyperparameter details are given in Table 10 in App. B.4.

During initial experiments, our model struggled to distinguish answerable from unanswerable SQA questions. To improve this, we apply chain-of-thought prompting: the model first tags the question as answerable or not, then generates an answer only if applicable. This stepwise approach improves both classification and answer quality.

¹³facebook/seamless-m4t-v2-large

¹⁴meta-llama/llama-3.1-8B-Instruct

¹⁵Augmented Dataset available at HuggingFace: [maikezu/data-kit-sub-iwslt2025-if-long-constraint](https://huggingface.co/maikezu/data-kit-sub-iwslt2025-if-long-constraint)

Model	ASR ●	ST ●			SQA ○	SSUM ○
	ACL 60/60 WER en-en	ACL 60/60 COMET en-de	en-it*	en-zh	Sp.-SQuAD BERTScore en-en	NUTSHELL BERTScore en-en
~no pretrain	25.1	72.49	73.61	76.93	80.88	83.89
~ASR pretrain	21.42	76.72	79.73	80.62	82.48	85.97
~contr. cos.	18.82	77.31	80.27	80.76	82.53	86.07
~contr. wasser.	19.07	77.33	80.06	81.34	82.66	86.6

~ Model not trained on multilingual SSUM and SQA

● Gold segmentation

○ No segmentation (full audio used)

Table 5: Ablation studies on different pretraining methods for the instruction following task: No pretraining, ASR pretraining and contrastive pretraining with either cosine similarity (*contr. cos.*) or Wasserstein distance (*contr. wasser.*). Test sets marked with * are automatically generated due to lack of availability for this language pair (see Section 3.1).

Segm.	max secs.	ASR (WER)	ST (COMET)		
		en-en	en-de	en-it*	en-zh
●	N/A	18.77	77.15	80.65	81.83
●	5	45.52	57.55	51.47	72.73
●	10	20.73	65.55	56.88	76.97
●	15	20.74	68.92	58.24	77.44
●	20	20.63	69.94	59.01	77.45
●	25	25.48	71.61	75.74	78.04
●	30	-	70.79	58.99	76.16
●	35	-	67.54	56.88	76.5

● Gold segmentation

● VAD segmentation

Table 6: Ablation study on Voice Activity Detection (VAD) segmentation using the *IF contr. cos. model.* on the ACL 60/60 dataset. Test sets marked with * are automatically generated due to lack of availability for this language pair (see Section 3.1). For ASR, segmenting audio into chunks of up to 20 seconds yields the best results, while for ST, 25-second chunks perform best.

3.3 Handling long audio

The IF Constrained Long track involves processing audio inputs from five to ten minutes in duration.

ASR and ST Initial experiments revealed that our model struggled with full-length audio inputs for ASR and ST, even when trained with artificially concatenated long-form sequences. To address this, we segment the input audio prior to inference.

We use a Voice Activity Detection (VAD) approach (Sohn et al., 1999) to segment audio, as due to track constraints, SHAS (Tsiamas et al., 2022) is not permitted. For ASR, segmenting into chunks of up to 20 seconds yields best performance and for ST, segments of up to 25 seconds are more effective. Ablation results are provided in Table 6.

Post-editing context	ASR (WER)	ST (COMET)		
	en-en	en-de	en-it*	en-zh
No Post-Editing ●	20.63	71.61	75.74	78.04
1 ●	21.09	70.54	75.0	77.22
3 ●	20.96	71.91	75.88	77.17
5 ●	20.43	71.64	75.69	77.20
10 ●	21.88	71.90	75.53	77.14
15 ●	50.07	71.95	75.88	77.19
20 ●	50.12	71.82	75.55	77.20

● VAD segmentation































































Table 7: Ablation study on the context size of the post-editing model using the *IF contr. cos. model.* on the ACL 60/60 dataset. For ASR, a context size of 5 yields the best results, for ST, a context size of 15. For en→zh, post-editing does not lead to an improvement.




SQA and SSUM For SQA and SSUM, we use the full audio. To handle long-form audio, we segment audio into 60-second chunks. Each chunk is encoded, and the embeddings are concatenated before being passed to the Q-Former and LLM, following Züfle et al. (2025). This strategy maintains full end-to-end trainability. For audios exceeding 26.7 minutes, we truncate the input to fit within memory constraints.

3.4 Post-Editing

To improve output quality, we use a post-editing model that works on document level. This helps to correct scientific terminology and it restores contextual coherence that may be lost due to segmentation of long audio inputs.

For ASR, we train the post-editing model on the SeamlessM4T-Large transcriptions of the TTS-generated scientific abstracts from NUTSHELL,

Model	ASR 	ST 				SQA 				SSUM 			
	ACL 60/60 WER	ACL 60/60 COMET				NUTSHELL BERTScore				NUTSHELL BERTScore			
	en-en	en-de	en-it*	en-zh	en-en	en-en*	en-de*	en-it*	en-zh*	en-en	en-de*	en-it*	en-zh*
Phi-4 ¹⁶	16.8 	79.19 	83.43 	83.23 	82.56	91.78	76.85	78.41	74.41	86.27	67.71	69.76	57.03
Qwen2 Audio ¹⁷	20.14	72.35 	74.23 	77.19 	87.35	89.0	71.81	73.7	69.62	84.88	63.06	64.17	51.79
Whisper ¹⁸ + Llama 3.1 ¹⁴	14.67 	78.04 	81.93 	77.66 	82.39	91.18	71.87	72.58	54.62	86.62	57.16	58.64	49.31
Seamless V2 ¹³	18.91 	73.64 	78.78 	75.26 	—	—	—	—	—	—	—	—	—
IF contr. cos.	18.77 	77.15 	80.65 	81.83 	82.83	93.04	79.73	82.08	79.8	86.83	68.46	71.01	71.22
IF contr. cos. tag	19.82	76.95 	80.69 	81.75 	82.86	93.17	80.81	82.49	80.53	86.52	68.31	71.06	71.09
IF contr. wasser.	17.93	72.47 	79.12	80.88	82.79	93.42	80.65	82.46	80.47	86.86	68.45	71.08	71.37
IF contr. wasser. tag	17.78 	74.06 	78.87 	81.10 	82.80	93.24	80.87	82.76	80.32	86.89	68.76	71.16	71.54
IF contr. cos.	20.63 	71.61 	75.74 	78.04 	82.83	93.04	79.73	82.08	79.8	86.83	68.46	71.01	71.22
+ post-edit	20.43 	71.95 	75.88 	77.19 	×		×			86.85	68.61	71.22	71.17
IF contr. cos. tag	33.24	69.37 	73.36 	75.83 	82.86	93.17	80.81	82.49	80.53	86.52	68.31	71.06	71.09
+ post-edit	33.53 	70.39 	73.14 	73.20 	×		×			86.54	68.42	71.16	71.01
IF contr. wasser.	21.88 	71.61 	76.78 	78.21 	82.79	93.42	80.65	82.46	80.47	86.86	68.45	71.08	71.37
+ post-edit	33.51 	71.12 	77.23 	68.02 	×		×			86.88	68.68	71.12	71.24
IF contr. wasser. tag	22.07 	71.84 	76.29 	78.24 	82.80	93.24	80.87	82.76	80.32	86.89	68.76	71.16	71.54
+ post-edit	19.76 	72.29 	76.75 	77.21 	×		×			86.90	68.95	71.30	71.41

 Gold segmentation
 Voice Activity Detection (VAD) segmentation
 No segmentation (full audio used)

— Not supported by model
× post-editing not applied, because context is not available

Table 8: Results for baseline models and our end-to-end trained instruction-following models (*IF*), developed for the Constraint Instruction Following Long track. The *IF* models are pretrained using contrastive learning, with either cosine similarity (*contr. cos.*) or Wasserstein distance (*contr. wasser.*). To improve performance on question answering, we also experiment with tagging answers to indicate whether the question is answerable (+ *tag*). Test sets marked with * are automatically generated due to lack of availability for this language pair and task (see Section 3.1). The *IF contr. wasser. tag + post-edit* model was submitted to the shared task.

paired with the original text. For ST, we use the ACL 60/60 development set, transcribed by our *IF* model. The post-editing model setup is adapted from Section 2.2.1, with two key differences: in compliance with the constrained setting, we use LLaMA-3.1-8B-Instruct¹⁴ (Grattafiori et al., 2024) as the base model, and we predict the reference using only a single system output, since in the *IF* track we do not employ an ensemble.

We conduct experiments to examine the effect of context size on post-editing performance. For ASR, a context window of five sentences provides the best results, while ST benefits from a 15-sentence context. For en→zh, no performance gains are achieved. These results are summarized in Table 7. We also apply the post-editing model to SSUM outputs, using the full summary as context.

3.5 Baselines

We compare our system to four baseline models. We include two end-to-end Speech-LLMs: Phi-4¹⁶ (Abdin et al., 2024) and Qwen2 Audio¹⁷ (Chu et al., 2024a), using default parameter settings provided on Hugging Face model cards and following the prompts specified by the shared task. We also evaluate a cascaded baseline using Whisper-

large-v3¹⁸ (Radford et al., 2023), and LLaMA-3.1-8B-Instruct¹⁴ (Grattafiori et al., 2024) to follow the instructions. Lastly, for ASR and ST, we include SeamlessM4T-Large¹³ (Communication et al., 2023), given that it also serves as the speech encoder in our own end-to-end architecture.

3.6 Evaluation

We evaluate ASR with WER using JiWER, ST using COMET¹⁹ (Rei et al., 2022a), and SQA and SSUM using BERTScore (Zhang et al., 2020).

3.7 Development Results

All results can be found in Table 8. We evaluate our approach against the baselines from Section 3.5, as well as four end-to-end trained instruction-following models (*IF*). Among these, we compare two contrastive pretraining strategies (*contr. cos.* and *contr. wasser.*), as outlined in Section 3.2. For the SQA task, we also explore a chain-of-thought variant (*tag*), as detailed in Section 3.2.













ASR and ST Using gold segmentation, we compare our *IF* models against the baselines. Phi-4¹⁶ (Abdin et al., 2024) achieves the strongest performance on ST, while Whisper¹⁸ (Radford et al., 2023) performs best for ASR. However, our

¹⁶microsoft/Phi-4-multimodal-instruct

¹⁷Qwen/Qwen2-Audio-7B-Instruct

¹⁸openai/whisper-large-v3

¹⁹Unbabel/wmt22-comet-da

Model	ASR 	ST 			SQA 				SSUM 			
	WER	COMET			BERTScore (normalized)				BERTScore (normalized)			
	en-en	en-de	en-it	en-zh	en-en	en-de	en-it	en-zh	en-en	en-de	en-it	en-zh
Phi-4 ¹⁶ (baseline)	0.17 	0.55 	0.56 	0.51 	0.42	0.35	0.36	0.39	0.17	0.16	0.19	0.04
IF contr. wasser. tag (ours)	0.15 	0.74 	0.77 	0.77 	0.41	0.35	0.39	0.41	0.23	0.21	0.25	0.37

• Voice Activity Detection (VAD) segmentation

• No segmentation (full audio used)

Table 9: Official evaluation results for the IWSLT 2025 IF Speech Processing track in the long and constrained setting.

IF models consistently outperform both Qwen2 Audio¹⁷ (Chu et al., 2024a) and SeamlessM4T-Large¹³ (Communication et al., 2023). The latter result confirms that our end-to-end architecture is able to improve over the speech foundation model.

Under VAD segmentation, which is also used for the shared task testset, we observe a performance drop across all IF models, as expected. Applying post-editing partially mitigates this drop. For ASR, post-editing only improves *IF contr. cos* and *IF contr. wasser. tag*, bringing them close to their gold-segmented counterparts. In ST, post-editing yields consistent improvements for en→de and en→it, but not for en→zh, likely due to the limited Chinese capabilities of the post-editing model and sparse training data in that language.

SQA and SSUM On the SQA-NUTSHELL dataset, all IF models outperform the baselines, whereas on Spoken-SQuAD (which is extractive and out-of-domain), this is not the case. For SSUM, IF models consistently surpass the baselines, particularly in en→it and en→zh. Post-editing yields slight gains for SSUM as well, though similar to ST, no improvement is observed for en→zh.

Final Model We select *IF contr. wasser. tag + post-edit* for our final submission. It offers the best performance for ASR, SQA, and SSUM, and is competitive with the other IF models in ST.

3.8 Results on IWSLT Official Test Set

Table 9 shows the performance of our final system on the official IWSLT 2025 test sets provided by the organizers (Abdumumin et al., 2025). Our system outperforms the baseline in ASR, ST, and SSUM, and achieves stronger results in SQA across all language pairs except for en→en.

4 Conclusion

This system paper presents KIT’s submissions to the Offline and the IF Long tracks. By inte-

grating LLMs into both cascaded and end-to-end architectures for speech processing, we demonstrate their potential in handling a range of spoken language tasks. For future work, we aim to explore a unified architecture capable of producing high-quality translations while also supporting instruction-following capabilities.

Acknowledgments

Part of this work received support from the European Union’s Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BetWEEN People). Part of this work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 technical report](#). *Preprint*, arXiv:2412.08905.
- Idris Abdumumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 30 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria (in-person and online). Association for Computational Linguistics. To appear.
- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao,

- Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, and 25 others. 2024. **FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN**. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.
- Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.
- Hervé Bredin and Antoine Laurent. 2021. **End-to-end speaker segmentation for overlap-aware resegmentation**. In *Interspeech 2021*, pages 3111–3115.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024a. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024b. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. **Seamlessm4t: Massively multilingual & multimodal machine translation**. *Preprint*, arXiv:2308.11596.
- Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.
- Mara Finkelstein, David Vilar, and Markus Freitag. 2024. **Introducing the NewsPaLM MBR and QE dataset: LLM-generated high-quality parallel data outperforms traditional web-crawled data**. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1355–1372, Miami, Florida, USA. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. **The llama 3 herd of models**. *Preprint*, arXiv:2407.21783.
- Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. **MetricX-24: The Google submission to the WMT 2024 metrics shared task**. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.
- Philipp Koehn. 2005. **Europarl: A parallel corpus for statistical machine translation**. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Sai Koneru, Thai Binh Nguyen, Ngoc-Quan Pham, Danni Liu, Zhaolin Li, Alexander Waibel, and Jan Niehues. 2024a. **Blending LLMs into cascaded speech translation: KIT’s offline speech translation system for IWSLT 2024**. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 183–191, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.
- Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024b. **Contextual refinement of translations: Large language models for sentence and**

- document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.
- Sai Koneru, Matthias Huck, Miriam Exel, and Jan Niehues. 2025. Quality-aware decoding: Unifying quality estimation and decoding. *arXiv preprint arXiv:2502.08561*.
- Phuong-Hang Le, Hongyu Gong, Changan Wang, Juan Pino, Benjamin Lecouteux, and Didier Schwab. 2023. [Pre-training for speech translation: Ctc meets optimal transport](#). *Preprint*, arXiv:2301.11716.
- Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Proc. Interspeech 2018*, pages 3459–3463.
- Paul Lerner, Juliette Bergoënd, Camille Guinaudeau, Hervé Bredin, Benjamin Maurice, Sharleyne Lefevre, Martin Bouteiller, Aman Berhe, Léo Galmant, Ruiqing Yin, and Claude Barras. 2022. [Bazinga! a dataset for multi-party dialogues structuring](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3434–3441, Marseille, France. European Language Resources Association.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning*.
- Pierre Lison and Jörg Tiedemann. 2016. [OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).
- Danni Liu, Thai Binh Nguyen, Sai Koneru, Enes Yavuz Ugan, Ngoc-Quan Pham, Tuan Nam Nguyen, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2023. [KIT’s multilingual speech translation system for IWSLT 2023](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 113–122, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Gabriel Peyré and Marco Cuturi. 2019. [Computational optimal transport: With applications to data science](#). *Foundations and Trends® in Machine Learning*, 11:355–206.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- 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.
- Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. [Aligning neural machine translation models: Human feedback in training and inference](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. [Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. [A statistical model-based voice activity detection](#). *IEEE Signal Processing Letters*, 6(1):1–3.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. [SALMONN: Towards generic hearing abilities for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, MA Zejun, and Chao Zhang. 2023. [Salmonn: Towards generic hearing abilities for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Silero Team. 2021. Silero models: pre-trained enterprise-grade stt / tts models and benchmarks. <https://github.com/snakers4/silero-models>.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. [SHAS: Approaching optimal Segmentation for End-to-End Speech Translation](#). In *Proc. Interspeech 2022*, pages 106–110.

Changhan Wang, Anne Wu, and Juan Pino. 2020. [Covost 2: A massively multilingual speech-to-text translation corpus](#). *Preprint*, arXiv:2007.10310.

John Wiseman. 2019. Wiseman/py-webrtcvad. *GitHub repository*, Nov.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024a. X-agma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *International Conference on Machine Learning*, pages 55204–55224. PMLR.

Brian Yan, Patrick Fernandes, Jinchuan Tian, Siqi Ouyang, William Chen, Karen Livescu, Lei Li, Graham Neubig, and Shinji Watanabe. 2024. [CMU’s IWSLT 2024 offline speech translation system: A cascaded approach for long-form robustness](#). In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 164–169, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Maike Züfle, Sara Papi, Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, and Jan Niehues. 2025. Nutshell: A dataset for abstract generation from scientific talks. *arXiv preprint arXiv:2502.16942*.

Maike Züfle and Jan Niehues. 2024. [Contrastive learning for task-independent speechllm-pretraining](#). *Preprint*, arXiv:2412.15712.

Maike Züfle, Sara Papi, Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, and Jan Niehues. 2025. [Nutshell: A dataset for abstract generation from scientific talks](#). *Preprint*, arXiv:2502.16942.

A Offline Track - Prompts

LLM Fuse Prompt

Post-Edit the Automatic Speech Recognition Transcripts from different systems understanding the context.

ASR Transcripts:

```
System1: {Whisper v2 Hyps}
System2: {Whisper v2 FT Hyps}
System3: {Phi-4 Hyps}
System4: {Whisper v3 Hyps}
```

Post-Edited Transcript:
{Reference}

MT APE Prompt

```
<|im_start|>user
Post-Edit the German Translation
of the English sentence.
English:
{src}
German:
{mt}
<|im_end|>
<|im_start|>assistant
Post-Edited German:
{ref}
```

B Instruction-Following Track - Prompts

B.1 Data Augmentation Prompts SQA

System Prompt:

You are a professional question generator. Given a transcript, you will create three questions: two that can be answered based on the transcript and one that cannot be answered (but is relevant to the topic). The answers should be full sentences in the target language specified.

Your response must be in valid JSON format, with keys for 'questions' and 'answers'. Do not include any explanations or additional text.\n

Prompt:

```
<Transcript>\n
Based on the transcript, generate
a JSON dictionary with the
following structure.
The questions and answers must be
in <trg lang>:\n
{{\n
```



```
'  "questions": [\n'
'    {"q1": "First question in <trg lang>", "a1": "Full-sentence\nanswer in <trg lang>"},\n'
'    {"q2": "Second question in <trg lang>", "a2": "Full-sentence\nanswer in <trg lang>"},\n'
f'    {"q3": "Third question in\n<trg lang>", "a3": "N/A"}]\n'
  ]\n
}\n
Ensure the response is a valid\nJSON object with properly\nformatted keys and values.
```

B.2 Data Augmentation Prompts SSUM

System Prompt:

A chat between a curious user and a professional system for translating ACL abstracts.\n

Prompt:

<abstract>\nTranslate this abstract to <trg lang>. Do not provide any explanation or additional text.

B.3 Data Augmentation Prompts ST

System Prompt:

You are a professional translator. Your task is to provide accurate, fluent, and natural translations without adding explanations, comments, or extra content.

Prompt:

Translate the following English text into <trg lang>. Do not provide any explanation or additional text.\n<text>

B.4 Hyperparameters Model Training

training parameters	Q-Former Num Query Token	4
	Q-Former Num Hidden Layers	4
	Q-Former Num Attention Heads	12
	Q-Former Seconds per Window	1/3
	num GPUs	4
	learning rate	1e-4
	warmup ratio	0.03
	optimizer	adamw_torch
	learning rate scheduler type	cosine
	model max length	2048
pretraining specific	gradient clipping	1
	num epochs	5
	per device batch size	10
	gradient accumulation steps	2
	contrastive τ cos + wasser	0.1
	contrastive τ nwp	0.5
	sinkhorn loss p	2
	sinkhorn loss blur	0.5
finetuning specific	num epochs	2
	per device batch size	2
	gradient accumulation steps	10

Table 10: Hyperparameters for the trainings, which are conducted on four NVIDIA GH200 96GB GPUs, mostly following [Züfle and Niehues \(2024\)](#).