Speech Translation and Metrics in 2026: Findings of the IWSLT Campaign

David Ifeoluwa Adelani[1,2] **Victor Agostinelli**[3] **Antonios Anastasopoulos**[4] **Luisa Bentivogli**[5] **Ondřej Bojar**[6] **Sébastien Bratières**[7] **Marine Carpuat**[8] **Fabrício Carraro**[9] **Roldano Cattoni**[5] **Mauro Cettolo**[5] **Lizhong Chen**[3] **Marcello Federico**[10] **Marco Gaido**[5] **Mahendra Gupta**[11] **HyoJung Han**[8] **Ali Hatami**[12] **Lewis C. Howe**[13] **Dávid Javorský**[6] **Yejin Jeon**[1,2] **Marek Kasztelnik**[14] **Antoine Laurent**[15] **Danni Liu**[16] **Nam Luu**[6] **Min Ma**[17] **Dominik Macháček**[6,18] **Marie Maltais**[1,2] **Evgeny Matusov**[19] **John McCrae**[12] **Chutong Meng**[4] **Chandresh Kumar Maurya**[20] **Mohammad Mohammadamini** [15] **Yasmin Moslem**[21] **Kenton Murray**[4] **Satoshi Nakamura**[22] **Matteo Negri**[5] **Jan Niehues**[16] **Atul Kr. Ojha**[12] **John E. Ortega**[23] **Siqi Ouyang**[24] **Sara Papi**[5] **Peter Polák**[19,6] **Fabian Retkowski**[16] **Stephanny Sánchez**[25] **Beatrice Savoldi**[5] **Claytone Sikasote**[26] **Matthias Sperber**[27] **Sebastian Stüker**[28] **Katsuhito Sudoh**[29] **Marie Tahon**[15] **Marco Turchi**[28] **Alex Waibel**[24] **Patrick Wilken**[19] **Rodolfo Zevallos**[30] **Vilém Zouhar**[31] **Maike Züfle**[16]

[1]McGill [2]Mila [3]Oregon State U. [4]GMU [5]FBK [6]Charles U. [7]Translated [8]UMD [9]Barcelona SC [10]Amazon [11]GPC Anuppur [12]U. Galway [13]U. Georgia [14]ACC Cyfronet AGH [15]Le Mans U. [16]KIT [17]Google DeepMind [18]U. of Edinburgh [19]AppTek [20]IIT Indore [21]ADAPT Centre [22]CUHK Shenzhen [23]Northeastern U. [24]CMU [25]U. Ing. Nac. Perú [26]U. Cape Town [27]Apple [28]Zoom [29]Nara Women's U. [30]U. Pompeu Fabra [31]ETH Zurich

Abstract

This paper reports on the outcomes of the shared tasks organized as part of the 23rd International Workshop on Spoken Language Translation (IWSLT). The workshop covered ten major challenges in spoken language translation, including speech-to-text translation for both high-resource and low-resource language pairs, customized speech translation, speech generation, instruction-following speech processing, and the evaluation of speech translation systems. The shared tasks received strong participation, with more than 30 teams submitting runs. This year's edition broadened the range of tasks, placing particular emphasis on speech generation and evaluation metrics.

Sections

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 336–422

1 Introduction

The International Conference on Spoken Language Translation (IWSLT) stands as the leading annual scientific conference dedicated to advancing all aspects of spoken language translation (SLT). Operating under the auspices of the Special Interest Group on Spoken Language Translation (SIGSLT), the conference receives support from three prestigious organizations: the Association for Computational Linguistics (ACL), the International Speech Communication Association (ISCA), and the European Language Resources Association (ELRA).

Maintaining its 23-year tradition, the 2026 conference was preceded by a comprehensive evaluation campaign designed to address critical scientific challenges in SLT. This paper presents the outcomes of the 2026 IWSLT Evaluation Campaign, which comprised ten distinct shared tasks organized into five primary research areas:

- **Speech-to-Text Translation**
  - **Offline track**, with focus on unconstrained speech-to-text translation of audio from a variety of diverse domains in mostly high-resource languages in two tracks (language-aware and language-agnostic).
  - **Low-resource SLT**, focusing on the translation of recorded speech from a variety of domains in low-resource and generally under-served languages, covering 10 language pairs. It also included a data track, inviting participants to submit newly collected speech translation datasets of under-resourced language pairs.
- **Customized ST**
  - **Model compression**, with focus on speech-to-text translation of recorded scientific presentations, TV series, and business news from English to German and Chinese, achieved by reducing the size of a large multilingual speech-to-text foundation model.
  - **Subtitling track**, with focus on speech-to-subtitle translation of audio-visual documents from English to five languages (Arabic, German, Chinese, Japanese, and Spanish).
  - **Simultaneous track**, focusing on speech-to-text translation of streamed audio of conferences and interviews from English to German, Italian and Chinese, and from Czech to English.
- **Speech Generation**
  - **Indic S2S track** focuses on speech-to-speech translation between English and three Indic languages (Hindi, Marathi, Punjabi) in both directions.
  - **African/Celtic S2S track** focuses on speech-to-speech translation from three under-resourced languages (Hausa, Igbo, Yorùbá) to English (targets being both text and speech for two tracks).
  - **Cross-Lingual Voice Cloning track**, which requires systems to synthesize speech in a target language while preserving the voice characteristics of a speaker from source language audio. Unlike traditional speech translation tasks, the cross-lingual voice cloning task focuses on transferring voice identity across languages while maintaining naturalness and intelligibility.
- **Instruction-following Speech Processing task** It aims to test general-purpose multimodal models (known as multimodal LLMs or audio-LLMs) for the speech modality. Covers the downstream tasks of Speech Recognition, Translation, Question Answering, Summarization, Audio Chaptering, and a surprise task, with focus on Scientific talk audios from English to German, Italian, and Chinese.
- **Speech Translation Metrics task**, focusing on Quality Estimation for speech translation, a reference-free evaluation of speech translation quality. Participants assess the quality of translations produced in other IWSLT shared tasks, and system outputs will be evaluated based on their correlation with human judgments.

2 Cross-Task Evaluation

The evaluation campaign features both automatic and human evaluation. To support automatic evaluation, we developed a dedicated evaluation server this year, as detailed in Section 2.1. The server was piloted in the *Offline*, *Model Compression*, and *Instruction Following* tracks. For the other tracks, submission and evaluation processes were managed by the respective organizers, following the

Team	Organization	Tracks	Reference
ADAPT-MTU	Munster Technological University		Sonawane and Afli (2026)
APG	individual		Palomino (2026)
ARN-SPA	Universidad Nacional Autónoma de México		Martínez et al. (2026)
AppTek	AppTek		
Aura-ST	Kitami Institute of Technology and RBG AI Research		HB et al. (2026)
BSC	Barcelona Supercomputing Center		Pareras et al. (2026)
CATENG	Universitat Pompeu Fabra, Universitat Politècnica de Catalunya, Barcelona Supercomputing Center, Northeastern University		Zevallos et al. (2026)
CPII-HK	Centre for Perceptual and Interactive Intelligence HK		Xing et al. (2026)
CUHKSZ	The Chinese University of Hong Kong, Shenzhen		Yang and Nakamura (2026); Sun et al. (2026)
CUNI-Pocket	Charles University		Ortega and Macháček (2026)
CUNI-Align	Charles University		Fuxa and Macháček (2026)
CISCO	Cisco		Gupta (2026)
ETH	ETH Zürich		Zarzu and Zouhar (2026)
Fbk	Fondazione Bruno Kessler, Italy		Cettolo et al. (2026a); Xie et al. (2026)
FLEURS-Badini	LIUM, Le Mans University and University of Duhok		Mohammadamini et al. (2026)
Hw-Tsc	Huawei Translation Services Center, China		Huang et al. (2026b); Lan et al. (2026a); He et al. (2026)
OSU	The Ohio State University		Krahn and Fosler-Lussier (2026)
IIIT-BGP	Indian Institute of Information Technology Bhagalpur		Singh et al. (2026)
IIT-Patna	IIT-Patna		Ahtasam et al. (2026)
individual	University of Washington		Pong (2026)
kit	Karlsruhe Institute of Technology		Akti and Waibel (2026); Liu et al. (2026); Ugan et al. (2026); Dinh and Niehues (2025)
KHU	Kyung Hee University		Shah et al. (2026)
	Langswap		Shigabeev et al. (2026)
lium	LIUM, Le Mans University		Mohammadamini and Tahon (2026)
MLLP-VRAIN UPV	Universitat Politècnica de València		Iranzo-Sánchez et al. (2026)
NeMo	NVIDIA		Grigoryan et al. (2026)
NLE	NAVER LABS Europe		Boito et al. (2026)
pinch-ast	Pinch		Bentes and Safka (2026)
QUESPA	Northeastern University, Universitat Pompeu Fabra, Barcelona Supercomputing Center, Universidad Nacional de Ingeniería, Peru, Universtiy of Georgia		Ortega et al. (2026a)
SIT-TCD	Shaggar Institute of Technology and Trinity College Dublin		Abebe and Moslem (2026)
TalTech	TalTech		
velo	individuals		Callañaupa et al. (2026)

Table 1: List of participants to the IWSLT 2026 shared tasks (Offline track; Simultaneous track; Subtitle track; Compression track; Low-resource track; Indic S2S track; African/Celtic S2S track; Instruction-following track; Cross-Lingual Voice Cloning; Metrics track.

procedure used in previous campaigns. In addition, we performed a human evaluation across several tracks as described in 2.2

2.1 SPEECHM-IWSLT2025 Evaluation Server

The Evaluation Server is a suite of datasets and metrics designed to measure and monitor the performance of task-specific systems. It is part of the "SPEECHM" platform, developed under the Meetween European Project.[1] For the IWSLT-2025 Evaluation Campaign, a dedicated instance—SPEECHM-IWSLT2025[2]—was created. This instance features a web-based user interface that allows participants to submit system outputs and track their performance via a leaderboard. The implemented evaluation metrics depend on the task: COMET, BLEURT, BLEU and CharacTER are used in the Offline and the Model Compression tasks, while WER, COMET and BERT scores are used Instruction Following task.

The Evaluation Server is described in detail in Appendix B.1.

2.2 Human Evaluation

Similar to last year's round, a human evaluation through direct assessment is performed on the primary submissions of each participant of offline, compression, and instruction following tasks in order to verify the soundness and completeness of the results. We mostly follow Sperber et al. (2024)'s approach to handle the automatically segmented long-form speech in a robust manner. A key difference is that human evaluation was done based on segmented source audio, not segmented source transcripts. Details are provided in Appendix A.

[1] www.meetween.eu
[2] iwslt2025.speechm.cloud.cyfronet.pl

Track I Offline track

1 Introduction

The Offline speech translation task represents the longest-standing tradition within the IWSLT evaluation campaigns, serving as a primary benchmark for tracking and fostering technological progress in spoken language translation. Unlike tasks subject to strict temporal or structural constraints, such as simultaneous translation or subtitling, the offline track focuses on *unconstrained* speech translation. While maintaining a consistent core formulation, the evaluation has progressively evolved to reflect the complexities of real-world applications by introducing more challenging scenarios, diverse language pairs, heterogeneous domains, and spontaneous speaking styles. In this spirit, the 2026 edition introduced several key advancements designed to push the boundaries of state-of-the-art architectures. First, Japanese was integrated among the target languages. Second, the multi-domain evaluation corpus, which in 2025 already featured TV series, scientific presentations, business news, and dedicated accented speech data, was further expanded in this edition. By incorporating two brand-new scenarios, call center interactions and YouTube videos, the task significantly broadened the spectrum of communicative situations and acoustic conditions, such as overlapping speakers and background noise, that systems had to navigate. Third, this year marked the debut of a novel "language-agnostic" track, designed to test a system's ability to translate speech when the source language is unknown.

2 Task Description

The 2026 round of the Offline task challenged participants to build robust translation systems capable of operating across diverse acoustic environments and domains without resorting to *ad-hoc*, domain-specialized models. To evaluate different aspects of this challenge, the campaign was structured into two distinct tracks:

- **Language-aware Track:** This track followed the traditional evaluation setup, where the source and target languages were known *a priori*. Participants could submit systems for four translation directions originating from English:
  - *English* → *German*: Evaluated on TV series, scientific presentations, call center two-person conversations, YouTube videos, business news, and accented speech.
  - *English* → *Chinese*: Evaluated on TV series, scientific presentations, call center conversations, YouTube videos, and business news.
  - *English* → *Japanese*: Evaluated on TV series, scientific presentations, call center conversations, YouTube videos, and business news.
  - *English* → *Arabic*: Focused exclusively on the business news domain.
- **Language-agnostic Track:** A newly introduced challenge designed to foster the development of universal, flexible speech translation models. Under this paradigm, systems had to process and translate input speech without any pre-defined source language labels. The evaluation covered three source languages (Czech, German, and English) and a single target language (English), which implicitly required the systems to perform English-to-English Automatic Speech Recognition (ASR) alongside translation.

Participating teams were free to submit systems for any combination of translation directions within the language-aware track, the language-agnostic setting, or both.

In line with previous campaigns, both cascade (pipeline) and end-to-end (E2E) architectures were eligible for participation. To ensure a fair comparison, E2E systems were strictly defined as those that did not employ intermediate discrete representations (such as source language transcripts) during inference. Furthermore, all parameters and components utilized during decoding had to be optimized for the direct speech-to-translation task (multitask training was permitted, whereas language model rescoring was prohibited).

To accommodate different computational constraints and resource availability, three training paradigms were defined for each translation direction: *constrained* (restricted to a specific list of medium-sized corpora), *constrained with LLMs* (allowing the standard constrained data plus any open-source Large Language Model under a permissive license), and *unconstrained* (permitting any external resource except the official evaluation sets).

Also this year, system submissions were managed via the centralized "SPEECHM" Evaluation Server.[3] While teams were allowed to upload mul-

[3] https://speechm.cloud.cyfronet.pl/0000

tiple runs, they were required to designate one *primary* system per track, with all remaining submissions treated as *contrastive*.

3 Data and Metrics

Test Data The evaluation framework for this edition featured a heterogeneous suite of test sets designed to reflect a wide array of operational domains and acoustic environments, specifically encompassing:

- **TV Series** from ITV Studios[4] – This subset comprises three distinct recordings of roughly 45 minutes each, capturing multi-party interactions across diverse scenarios. Deploying systems on this data requires managing complex phenomena such as overlapping speech, heterogeneous accents, and ambient noise.
- **Scientific Presentations** – This dataset includes 21 recordings of oral presentations, with an average length of approximately 6 minutes per recording, providing the original transcripts alongside their multilingual translations. The collection covers a broad spectrum of technical and academic topics presented by an international cohort of speakers.
- **Call Center two-person conversations** – This dataset gathers unscripted, 10-to-15-minute simulated interactions between customers and agents recorded via online streaming platforms. The dialogues cover spontaneous transactions and inquiries regarding goods or services across multiple commercial sectors.
- **YouTube videos** from YODAS[5] (YouTube-Oriented Dataset for Audio and Speech) – Sourced from large-scale YouTube collections, this dataset includes five English audio files with durations spanning between 10 and 30 minutes (amounting to approximately 1.5 hours in total). High-quality reference translations for German, Japanese, and Chinese were professionally curated by AppTek[6].
- **Business News** from Asharq Business with Bloomberg[7] – Representing the financial and economics domain, this dataset consists of a single broadcast recording (approximately 1.5 hours) from a television channel and distributed across digital and social platforms.
- **Accented English Conversations** sampled from the Edinburgh International Accents of English Corpus (EdAcc, Sanabria et al., 2023) – This dataset contributes roughly 3.5 hours of unscripted dialogues where pairs of friends discuss everyday topics, such as leisure activities and travel. It introduces a substantial phonetic challenge by featuring 33 distinct international English accents, combined with the inherent difficulties of highly spontaneous speech.
- **Synthetic TTS Audio Data** is generated from the English-German "Scientific Presentation" dataset, where the transcripts are synthesized using different TTS models. Specifically, two TTS systems are used: VITS (Kim et al., 2021) and Kokoro-82M[8]. VITS was used to generate two versions of the data — one with a male voice and one with a female voice. The purpose of this test set is to evaluate the impact of the voice quality on the performance of the submitted models.
- **Language-agnostic test data** For the language-agnostic task, we combined multiple datasets, without revealing the data source of individual test instances to the participants. The English input data comprised scientific presentations used in the other track. For German, we recorded additional scientific presentations in German and translated them into English. For the Czech-to-English condition, we used the dataset from the simultaneous track. Furthermore, motivated by (Sinhamahapatra et al., 2026), we created an additional test set containing mixed German-English dialog about scientific papers, which was translated into English.

Training and Development Data In continuity with recent rounds of the challenge, participants could submit systems built under three training data conditions: *i)* constrained, *ii)* constrained with large language models, and *iii)* unconstrained. Under the constrained setting, the list of permitted training resources[9] comprised a variety of speech, speech-to-text parallel, text-parallel, and text-monolingual datasets. In this condition, the use of any pre-trained language models was prohibited. For the Constrained with large language models condition (constrained^{+LLM}), participants were allowed to supplement the con-

005

[4] https://www.itvstudios.com/

[5] https://huggingface.co/datasets/espnet/yodas

[6] https://www.apptek.ai/

[7] https://asharqbusiness.com/

[8] https://huggingface.co/hexgrad/Kokoro-82M

[9] See https://iwslt.org/2026/offline

strained resources with freely accessible large language models released under permissive licenses. The objective of this setup was to enable the leveraging of accessible LLMs within a standardized evaluation scenario. For the unconstrained condition, the use of any resources, including pretrained language models, was permitted, with the sole exception of the official evaluation sets. The aim of this setup was to accommodate teams equipped with high computational power and capable of developing effective solutions leveraging additional in-house resources.

Two types of development data were made available. These included the material released for previous rounds of the Offline task[10] and the development data provided for the Subtitling task[11], which specifically covered the TV series, business news, and YouTube video domains.

Evaluation Metrics System performance was evaluated against human-curated target-language references using a suite of automatic metrics: COMET, BLEURT, BLEU, TER, and characTER. COMET, the metric selected for the official ranking, was computed on hypotheses automatically resegmented to match the reference text via the `mwerSegmenter`[12] tool.

4 Results

In this edition, only HW-TSC (Huang et al., 2026a) participated in the English–German and English–Chinese language-aware tracks, using a cascade system trained under the "unconstrained" condition. The model is a pipeline composed of: (1) a VAD module to segment long audio, (2) a two-pass transcription system that first produces a rough transcript with timestamps, then re-segments the audio based on these timestamps and finally runs ASR on the newly segmented audio, and (3) a translation component that generates the final translations. All components—VAD, ASR, and MT—are implemented using Qwen-family LLMs (Qwen et al., 2025) of varying sizes (0.6M, 4B, and 8B parameters). The ASR and MT components have been fine-tuned using existing datasets. The MT training data was cleaned through a series of heuristic filters to remove noise, including language identification, length-ratio constraints, character-level deduplication, and a semantic alignment step.

The results are reported in Tables 27 and 28. Although the evaluation is limited to a single participant, these results clearly highlight the different levels of complexity across the evaluated test sets. Specifically, domains featuring multi-party interactions and challenging acoustic phenomena—such as the TV Series, Accented English Conversations (evaluated only for En-De), and YouTube videos—yield substantially lower automatic quality scores. In contrast, single-speaker audio such as the Scientific Presentations subset yields significantly higher evaluation scores, despite containing specialized terminology. These results confirm the well-known challenges that current systems face when dealing with complex, real-world acoustic conditions, and highlight the need for continued research and development in this area.

When comparing the model's performance on human-generated audio (Scientific Presentations) against its respective transcripts and references with TTS-generated audio (Synthetic TTS Audio data 1, 2, and 3), the differences in COMET scores are marginal and likely lack statistical significance. This suggests that the model is largely robust to the acoustic characteristics of the input signal, treating human-recorded and synthetically generated speech in a similar manner. Furthermore, the choice of synthesis engine and speaker voice (female vs. male) exerts no discernible impact on the final metrics, indicating that the specific acoustic properties introduced by different voice types or synthesis architectures are not a decisive factor in determining downstream translation quality.

While these findings warrant validation at a larger scale and across a broader range of languages and domains, they point to promising applications of synthetic audio data in speech translation research and development. Specifically, the viability of supplementing or substituting human-recorded audio with high-quality synthetic alternatives could significantly reduce the cost and effort required to curate benchmark datasets and training resources.

For instance, synthetic data could be used to construct evaluation benchmarks for specialized domains—such as medical, legal, or technical fields—where parallel textual corpora exist but

[10] https://huggingface.co/datasets/IWSLT/IWSLT.OfflineTask

[11] https://iwslt.org/2026/developmentData

[12] https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

corresponding audio is missing, and where collecting authentic recordings is economically or logistically prohibitive. Furthermore, TTS engines could be leveraged to synthesize extensive training subsets across diverse domains, thereby enhancing model robustness and generalization capability while bypassing the constraints of resource-intensive data collection. Overall, these initial insights suggest that synthetic audio represents a promising path toward viable and scalable resources for supporting the evaluation of speech translation systems.

Track II Low-resource SLT

1 Challenge

As with previous years, the goal of the low-resource shared task is to benchmark and promote speech translation technology for a diverse range of dialects and low-resource languages. While significant research progress has been demonstrated recently on popular datasets, many of the world's languages lack the parallel data at scale needed for standard supervised learning. The community will likely require creative approaches in leveraging disparate resources in order to make progress.

The low-resource shared task will, as last year, involve two tracks:

- Track 1: A "traditional" speech-to-text translation track focusing on 10 typologically diverse language pairs.
- Track 2: A data track, inviting participants to provide open-sourced speech translation datasets for under-resourced languages.

This year's focus was on **explicitly multilingual systems that can handle speech from as many diverse languages as possible.** We welcomed general recipes aimed at improving speech translation broadly for a wide typology of languages. Thus, while participants were free to participate in any number of language pairs that are our focus, we highly encouraged participation in as many as possible. Of course, we still welcomed dedicated systems that are designed to cater to a single language pair.

2 Data and Metrics

Irish–English (gle-eng) Irish (also known as Gaeilge; ISO code: `gle`) has around 170,000 L1 speakers and 1.85 million people (37% of the population) across the island (of Ireland) claim to be at least somewhat proficient with the language. In the Republic of Ireland, it is the national and first official language. It is also one of the official languages of the European Union (EU) and a recognised minority language in Northern Ireland with the ISO `ga` code.

The provided Irish audio data were compiled from the news domain, Common Voice (Ardila et al., 2020),[13] and Living-Audio-Dataset.[14] The Irish-to-English corpus comprises approximately 13 hours of Irish speech data , translated into English texts.[15] This year, we also provided the participants of three synthetic audio Irish-to-English datasets comprising 196 hours (Moslem, 2024). The synthetic data was created by synthesizing audio from parallel textual datasets obtained from OPUS (Tiedemann, 2012), namely EUbookshop, Tatoeba, and Wikimedia.[16]

Bhojpuri–Hindi (bho-hin) Bhojpuri (ISO code: `bho`) belongs to the Indo-Aryan language group. It is dominantly spoken in India's western part of Bihar, the north-western part of Jharkhand, and the Purvanchal region of Uttar Pradesh. As per the 2011 Census of India, it has around 50.58 million speakers (Ojha and Zeman, 2020). Bhojpuri is spoken not just in India but also in other countries such as Nepal, Trinidad, Mauritius, Guyana, Suriname, and Fiji. Since Bhojpuri was considered a dialect of Hindi for a long time, it did not attract much attention from linguists and hence remains among the many lesser-known and less-resourced languages of India.

The provided Bhojpuri–Hindi corpus consists of approximately 24 hours of Bhojpuri speech data from the news domain, extracted from News On Air[17] and translated into Hindi texts.[18] Additionally, the participants were directed that they may use monolingual Bhojpuri audio data (with transcription) from ULCA-asr-dataset-corpus[19] as well as Bhojpuri Language Technological Resources (BHLTR) (Ojha et al., 2020; Ojha, 2019)[20] and Bhojpuri-wav2vec2 based model.[21]

Mapuzugun–Spanish (arn-spa) Mapudungun is a language isolate and the indigenous language of the Mapuche people, spoken by about 100,000-200,000 people in Chile and Argentina.

Data are based on the corpus described in this paper, providing more than 130 hours of Mapuzugun speech, along with transcriptions and translations in Spanish.

[13] commonvoice.mozilla.org/en/datasets
[14] github.com/Idlak/Living-Audio-Dataset
[15] github.com/shashwatup9k/iwslt2026_ga-eng
[16] hf.co/collections/ymoslem/irish-english-speech-translation-datasets-665dd9e8fbaa279db3474ca0
[17] newsonair.gov.in
[18] github.com/shashwatup9k/iwslt2026_bho-hi
[19] github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus
[20] github.com/shashwatup9k/bho-resources
[21] www.openslr.org/64/

Bemba–English (bem-eng) Bemba is a Bantu language, spoken by over 10 million people in Zambia and other parts of Africa. It is the most populous indigenous language spoken by over 30% of the population in Zambia where English is the lingua franca and official high-resourced language of communication. Bemba is native to the people of Northen, Luapula and Muchinga provinces of Zambia but also spoken in other parts of the country including urban areas such as Copperbelt, Central and Lusaka provinces by over 50% of the population (ZamStats, 2012).

The provided Bemba-English corpus (Sikasote et al., 2023a) consists of over 180 hours of Bemba audio data, along with transcriptions and translations in English. The dataset is comprised of recorded multi-turn dialogues between native Bemba speakers grounded on images.

In addition, we provided transcribed (28 hours) and untranscribed (60 hours) monolingual Bemba speech from Zambezi Voice (Sikasote et al., 2023b) and BembaSpeech (Sikasote and Anastasopoulos, 2022) datasets.

Central Kurdish–English (ckb-eng) This task focuses on speech-to-text translation from Central Kurdish to English. Central Kurdish (ISO 639-3) with an estimated 8 million speakers, is the second most widely spoken dialect of the Kurdish language. It is spoken mainly in the Kurdistan regions of Iran and Iraq and uses a modified version of the Arabic script.

The task is based on the COMMUTE-Kurdish corpus, which contains 30 hours of spontaneous Central Kurdish speech collected from Kurdish media. The data are manually segmented, transcribed, and translated into English. The corpus covers multiple domains, including politics, culture, economy, sports, art, and science.

The COMMUTE-Kurdish dataset, complementary datasets, baseline models, evaluation instructions, and contact information are publicly available.[22]

Igbo–English (ibo-eng) The Igbo language is a Niger-Congo language spoken by approximately 30–45 million people in Nigeria, Cameroon, and Equatorial Guinea.

Newly collected data is available here, and they are the same as the ones used for the African and Celtic Speech-to-Speech Shared Task. We encourage interested participants to also consider participating in that task as well.

Hausa–English (hau-eng) Hausa is a Chadic language spoken by more than 100 million people in Nigeria and Niger.

Newly collected data is available here, and they are the same as the ones used for the African and Celtic Speech-to-Speech Shared Task. We encourage interested participants to also consider participating in that task as well.

Quechua–Spanish (que-spa) Quechua is an indigenous language spoken by more than 8 million people in South America. It is mainly spoken in Peru, Ecuador, and Bolivia where the official high-resource language is Spanish. It is a highly inflective language based on its suffixes which agglutinate and found to be similar to other languages like Finnish. The average number of morphemes per word (synthesis) is about two times larger than English. English typically has around 1.5 morphemes per word and Quechua has about 3 morphemes per word.

There are two main region divisions of Quechua known as Quechua I and Quechua II. This data set consists of two main types of Quechua spoken in Ayacucho, Peru (Quechua Chanka ISO:quy) and Cusco, Peru (Quechua Collao ISO:quz) which are both part of Quechua II and, thus, considered "southern" languages. We label the data set with que - the ISO code for Quechua II mixtures.

IWSLT participants may obtain the public Quechua-Spanish speech translation dataset along with the additonal parallel (text-only) data for the unconstrained task at no cost here: IWSLT 2026 QUE-SPA Data set. IWSLT particpants should feel free to use any publicly available data for the unconstrained task. This includes a data set of nearly 50 hours of fully transcribed Quechua audio from previous shared tasks along with the introduction of a new data set this year which is about 8 hours of synthetic (post-edited) translations. A new addition this year is a Quechua Collao dataset which contains 15 hours of ASR data with Spanish translation.

Catalan–English (cat-eng) Catalan (català) is a Western Romance language which has approximately 4.1 million L1 speakers and more than 10 million people who can speak the language across its territories. It is spoken primarily in Catalo-

[22] https://lium.univ-lemans.fr/en/corpus-commute-kurdish/

nia, the Valencian Community, the Balearic Islands, and parts of Aragon in Spain, as well as in Andorra, where it is the sole official language. Catalan is also spoken in parts of southern France (Northern Catalonia) and in the city of Alghero in Sardinia, Italy.

In Catalonia, the Valencian Community, and the Balearic Islands, Catalan is co-official with Spanish and is widely used in education, media, and public administration. It is recognized as a regional or minority language in several European regions.

Yoruba–English (yor-eng) The Igbo language is a Niger-Congo language spoken by approximately 50 million people in Nigeria, Benin, and Togo.

Newly collected data is available here, and they are the same as the ones used for the African and Celtic Speech-to-Speech Shared Task. We encourage interested participants to also consider participating in that task as well.

2.1 Metrics

We use standard lowercase BLEU with no punctuation to automatically score all submissions. Additional analyses for some language pairs are provided below. We also report chrF++ (Popović, 2015a).

3 Submissions

QUESPA (Ortega et al., 2026b) submitted three unconstrained systems for the Quechua–Spanish track, marking the team's fourth consecutive participation in the shared task. This year's submission introduced three notable novelties beyond prior work: a machine translation case study using LLM-based prompting, an audio enhancement pipeline using SIDON (Nakata et al., 2025), and the formal incorporation of a Quechua Collao (ISO: quz) speech corpus (Paccotacya-Yanque et al., 2022) into the provided datasets. For the MT case study, the team benchmarked a range of large language models—including GPT-5, GEMINI 3, CLAUDE, DEEPSEEK-V3, and QWEN—using guided prompts in Spanish, finding that none surpassed the fine-tuned NLLB-200 baseline from the previous year (19.5 BLEU / 23.5 ChrF), with the best prompt-based result reaching 10.8 BLEU via Gemini 3 Flash. The team attributed the underperformance of prompting approaches to hallucinations, dialectal confusion between Quechua variants, and a tendency of models to prioritize high-resource language signals over low-resource Quechua input.

The primary and first contrastive systems follow a cascaded ASR+MT architecture, employing ConMamba and Whisper Large V3, respectively, each enhanced with SIDON noise reduction prior to fine-tuning, and decoded through the NLLB-based MT system; these yield 15.0 and 15.4 BLEU, gains of approximately 0.4 BLEU over the equivalent 2025 systems, attributable to SIDON preprocessing. The best-performing system (contrastive 2) is an end-to-end SpeechT5 model (Ao et al., 2022) fine-tuned on the full unconstrained training set augmented with `nlpaug` noise and distortion techniques, 15 hours of Quechua Collao speech (Paccotacya-Yanque et al., 2022), and SIDON-enhanced audio, yielding a BLEU score of 27.2—an improvement of 0.5 BLEU over the team's 2025 best result of 26.7 and the highest score recorded for the Quechua–Spanish task to date. As in 2025, ChrF scores for the SpeechT5 system did not improve alongside BLEU, while the cascaded Mamba and Whisper systems saw gains on both metrics with the addition of SIDON.

ADAPT-MTU HAI (Sonawane and Afli, 2026) presented a cascaded speech translation framework for the Bhojpuri-Hindi and Irish-English language pairs. Their approach combined Whisper-based Automatic Speech Recognition (ASR) with the NLLB-200 multilingual translation model. The team evaluated multiple ASR models and routing strategies, comparing direct and pivot-based translation. For Bhojpuri-Hindi, their best configuration utilised Whisper-large-v3 alongside direct NLLB translation, achieving a BLEU score of 14.7, a Chrf++ of 43. While for Irish-English, the team achieves 2.4 BLEU and 0.16 Chrf++ scores.

LIUM Mohammadamini and Tahon (2026) built a system for the Central Kurdish–English language pair. They focused on different data augmentation methods for low-resource speech-to-text translation, including two main pipelines: pseudo-labeling and speech synthesis. Their goal was to generate parallel speech data in low-resource scenarios without relying on human-annotated speech translation data. The pseudo-labeling approach essentially produces silver

translations for untranscribed audio through automated ASR and MT pipeline. The second approach uses a TTS and MT pipeline over source-side text data.

Their main finding is that using synthetic speech generation for real-world applications such as spontaneous speech translation remains particularly challenging and not fruitful. In contrast, the pseudo-labeling pipeline proved to be more effective, achieving performance comparable to cascaded models while reducing latency.

CATENG Zevallos et al. (2026) submitted to the Catalan-English language pair. Their primary system uses a Mabmba-based ASR (ConMamba) with a fine-tuned NLLB-200 MT model. Their contrastive system replaces the ASR with Whisper-v3. Their approach is similar in some aspects to the QUESPA system and highlights some of the popular approaches taken this year. Additionally, they evaluated an end-to-end SpeechT5 model with data augmentation. Overall, they find that cascaded systems continue to outperform end-to-end speech translation, with performance primarily being constrained by ASR quality over MT.

Arn-Spa Martínez et al. (2026) participated in the Mapudungun-to-Spanish speech translation shared task under the unconstrained condition. The team used the provided Mapudungun corpus and experimented with data augmentation techniques as well as discarding long recordings. For end-to-end translation, they applied parameter-efficient fine-tuning to Canary-1B-v2, a multilingual multi-task ASR and speech translation model. To avoid introducing a new language code for Mapudungun, they used English as the source language code. Their experiments showed that filtering out utterances longer than 15 seconds yielded the best performance.

IIIT-BGP (Singh et al., 2026) developed systems for low-resource Bhojpuri-Hindi speech translation, exploring both end-to-end and cascaded architectures. Their end-to-end model connected a Bhojpuri-fine-tuned Wav2Vec 2 encoder to a pre-trained NLLB-200 decoder via a lightweight interconnection adapter. This adapter utilised a combination of learnable layer aggregation, CNN-based temporal compression, and Transformer refinement, with optional LoRA-based decoder adaptation. For their cascaded system, the team fine-tuned Whisper for Bhojpuri ASR and NLLB-200 for Hindi machine translation. Furthermore, they improved the cascaded pipeline's outputs by applying Quality Estimation (QE) Fusion with COMET-Kiwi, which optimised translation selection directly from beam candidates. They submitted one primary and two contrastive systems. In these, the contrastive 1 system achieved slightly higher results: 12.3 BLEU and 35 Chrf++ scores.

CUHKSZ The Chinese University of Hong Kong, Shenzhen (Sun et al., 2026) participated in the shared task with 6 official languages from 2026, plus an additional 2 from prior iterations (bem, ckb, gle, hau, ibo, yor, aeb, est). The crux of their approach was a new method they proposed called Gradient-Driven Parameter Sharing (GDPS) which looks at inter-language gradient behaviors to automatically group languages. The idea is motivated by the problem of gradient conflicts that can occur when fine-tuning multilingual models on low-resource languages. Unlike other IWSLT low-resource systems, they do not rely on data augmentation. They use SeamlessM4T Medium as their baseline model.

Velo Callañaupa et al. (2026) participated in the Quechua-Spanish translation task. They used a fine-tuning and incremental retraining process of the NLLB-200 model for both Quechua variants (Chanka and Collao).

AURA-ST HB et al. (2026) participated in the XX → English speech-to-text translation (S2TT) task for the African-languages track. Their system, AURA-ST (Acoustic-Unconstrained Residual Architecture for Speech Translation), is a low-resource end-to-end speech translation framework that combines frozen w2v-BERT 2.0 (Baevski et al., 2020) and ResNet34 (He et al., 2015) speech encoders with a frozen Gemma-4-E2B language model. Acoustic representations are fused, compressed through a convolutional subsampler, and projected into the language model embedding space, where they are provided as a prefix prompt rather than through cross-attention. The model is adapted using parameter-efficient LoRA fine-tuning applied only to Gemma's MLP layers, enabling effective speech-to-text learning while minimizing catastrophic forgetting.

4 Results

Model	Hausa			Igbo			Yorùbá		
	spBLEU	chrF++	SSA-COMET	spBLEU	chrF++	SSA-COMET	sBLEU	chrF++	SSA-COMET
Aura-ST	5.2	23.2	33.9	4.6	16.7	20.6	19.5	41.0	57.1
SeamlessM4T Mono FT	**18.6**	41.9	**54.9**	**17.6**	**39.2**	**52.3**	**21.1**	**43.5**	**60.3**
Cascaded	17.3	**42.0**	54.1	11.0	33.8	42.9	17.0	38.1	50.6

Table 2: **S2TT results (XX → English)** across three metrics: spBLEU (↑), chrF++ (↑), and SSA-COMET (↑). **Bold** = best overall per column; underlined = second best. SeamlessM4T Mono FT is a fully supervised fine-tuned upper bound, monolingually fine tuned for each individual language pair. The cascaded model is a combination of Omnilingual-ASR (OmniASR_LLM_1B) and NLLB-200 for machine translation.

Submission	BLEU	Chrf++
IIIT-BGP Contrastive 2	10.1	39.0
IIIT-BGP Primary	12.1	34.0
IIIT-BGP Contrastive 1	12.3	35.0
ADAPT-MTU Contrastive	14.5	38.0
ADAPT-MTU Primary	**14.7**	**43.0**

Table 3: Bhojpuri–Hindi Results. All submissions were to the unconstrained track.

Submission	BLEU	Chrf++
Velo Primary	8.9	39.9
Quespa Primary	15.0	50.7
Quespa Contrastive 1	15.4	52
Quespa Contrastive 2	**27.2**	**51.4**

Table 4: Quechua–Spanish Results. All submissions were to the unconstrained track.

African-Celtic tracks Table 12 presents S2TT results for Hausa, Igbo, and Yorùbá. The supervised *SeamlessM4T Mono FT* system achieved the strongest overall performance, while the cascaded OmniASR+NLLB approach remained competitive for Hausa but degraded more noticeably for Igbo and Yorùbá. Although AURA-ST trailed both baselines on Hausa and Igbo, it performed particularly well on Yorùbá→English, achieving the second-best result across all three evaluation metrics and outperforming the cascaded system. These findings highlight both the challenges of speech translation for African languages and the potential of parameter-efficient speech-to-LLM adaptation methods, especially for relatively higher-resource languages such as Yorùbá.

Bhojpuri-Hindi Results are presented in Table 3, with the ADAPT-MTU primary submission outperforming the others.

Quechua-Spanish Results are presented in Table 4, with the QUESPA submissions outperforming the Velo submission. Interestingly, the second contrastive submission that employs a different architecture and additional data outperforms all others and improves over all previous years' submissions as well.

Irish-English Results are presented in Table 5. Notably, all submissions struggle to produce meaningful outputs, showcasing the difficulty of this task.

Submission	BLEU	Chrf++
MTU Contrastive 1	2.3	15.0
SLC Primary	2.4	10.0
MTU Primary	2.4	16.0

Table 5: Irish–English Results. All submissions were to the unconstrained track.

Mapuzugun-Spanish Results are presented in Table 6. Notably, all submissions struggle to produce meaningful outputs, showcasing the difficulty of this task.

Central Kurdish-English Results are presented in Table 7. All submissions are unconstrained.

5 Data Track Results and Discussion

This track aims to empower language communities to contribute to key datasets. These datasets are essential for expanding the reach of spoken language technology to more languages and varieties.

Progress made in translation quality has largely been directed at high-resource languages. Recently, focus has started to shift to under-served languages, and foundational datasets such as FLORES (Goyal et al., 2022) and NTREX (Federmann

Submission	BLEU	Chrf++
KK Contrastive 1	0.34	6.92
KK Primary	0.73	12.7
KK Contrastive 2	0.82	14.31

Table 6: Mapazugun–Spanish Results.

	Speech Translation		ASR	
Team	BLEU	Chrf++	CER	WER
SLC	0.16	13.6	-	-
LIUM	21.09	49.48	6.98	19.76

Table 7: Central Kurdish–English Results for Translation and Speech Recognition. Note that all submissions were to the unconstrained track.

et al., 2022b) have made it easier to develop and evaluate MT models for an increasing amount of languages. The high impact of these components left some in the research community wondering: how do we add more languages to these existing open-source datasets?

The goal of this shared task track is to expand open datasets to more languages. In particular, we are soliciting contributions to Speech Translation Training and Evaluation Datasets, either on text-to-speech or speech-to-speech formats.

Data Submission Requirements We highly encouraged participants to get creative, however we also wanted to ensure data quality and we asked participants to adhere to some suggestions.

Translations should be performed, wherever possible, by qualified, native speakers of the target language. We strongly encourage verification of the data by at least one additional native speaker. Dataset card: dataset cards should be attached to new data submissions, detailing precise language information and the translation workflow that was employed. In particular, we ask participants to identify the language with both an ISO 639-3 individual language tag and a Glottocode. The script should be identified with an ISO 15924 script code.

License: We highly encourage new contributions to be released under CC BY-SA 4.0 or other similarly permissive licenses. By contributing data to this shared task, participants agree to have this data released under these terms. At a minimum, data should be made available for research use.

Use of automatic translation or LLMs for data generation: while post-editing of automatic output is allowed, we require that any data submitted for the shared task are 100% verified by humans, if not directly created by humans. Raw, unverified machine translated outputs are not allowed. If using MT, you must ensure that the terms of service of the model you use allow re-using its outputs to train other machine translation models (as an example, popular commercial systems such as DeepL, Google Translate and ChatGPT disallow this).

Fleurs-Badini Mohammadamini et al. (2026) extended the FLEURS benchmark to the Badini variant of Northern Kurdish. Badini is a variant of Norther Kurdish spoken widely in the Kurdistan Region of Iraq. It differs from other Kurdish varieties by distinct phonological, lexical, and syntactic features.

The authors begin with 2,000 English sentences from FLORES and translate them into Badini dialect relying on university students from the Departments of English and Translation at the University of Duhok. The translation process includes various efforts to ensure coherence and naturalness. Speech recordings of the translated sentences are then collected from native speakers through an online platform. Both the translations and recordings have been manually reviewed by faculty members.

The resulting dataset contains 5,224 utterances totaling 15 hours and 40 minutes of speech containing recordings from 45 speakers. It is split into a training set of 2,022 utterances (5h47m), a development set of 1,165 utterances (3h36m), and a test set of 2,037 utterances (6h17m). The authors also evaluate ASR and speech translation models on the dataset. A fine-tuned Whisper model achieves 5.24 BLEU and 29.57 chrF++ on the test set.

Track III Model compression track

1 Introduction

The Model Compression task, now in its second edition, addresses a critical challenge in the natural language processing community: reconciling the remarkable capabilities of large foundation models with the strict constraints of practical deployment. Although large-scale text and speech models have fundamentally transformed spoken language translation, their massive parameter size and intensive computational demands pose severe bottlenecks in real-world applications. This is especially critical in resource-constrained environments, such as mobile devices, embedded systems, and edge computing, where low-latency, on-device inference is paramount. Model compression offers a viable path forward, shrinking model size and complexity while striving to minimize performance degradation. By focusing on this critical trade-off, the task aims to establish a rigorous benchmark for monitoring advancements in the development of efficient, accessible, and sustainable speech translation systems.

2 Task Description

Building on the foundation laid during the inaugural 2025 round, the 2026 evaluation challenged participants to effectively reduce the size of a large multilingual speech-to-text foundation model while minimizing performance drops in English→German and English→Chinese speech translation settings. The chosen baseline model remained Qwen2-Audio (Chu et al., 2024), selected due to its substantial size of 8.2 billion parameters (requiring approximately 16 GB of storage), its versatile support for multiple speech processing tasks, and its permissive Apache 2.0 license. Its memory-intensive nature and computational cost made it an ideal candidate for task-oriented model compression.

To ensure a fair comparison, admissible compression techniques had to focus exclusively on modifying or optimizing the model's internal parameters, ensuring that the final compressed system remained strictly derived from the original Qwen2-Audio. Eligible methodologies, which could be deployed either in isolation or in combination, included *pruning* (the removal of less important neurons or entire layers within the model by identifying and eliminating parameters that contribute minimally to the output), *quantization* (the reduction of the numerical precision of the model's weights, such as converting from 32-bit to 16-bit, 8-bit, or lower, to minimize the overall memory footprint), and *distillation* (the creation of a smaller student model derived from Qwen2-Audio and trained to replicate the behavioral characteristics of the original teacher model), as well as any other method that produces a compressed counterpart of the original model.

As in the 2025 round of the task, system submissions were managed via the centralized "SPEECHM" Evaluation Server,[23] allowing teams to upload multiple runs. However, participants were required to explicitly designate one primary system per track and language direction, with all remaining submissions treated as contrastive. In the absence of an explicit designation, a default timestamp-based rule was applied to automatically select the most recent submission as the primary run.

3 Data and Metrics

Test Data In contrast to the inaugural edition, which evaluated systems exclusively on the scientific presentations domain, the 2026 campaign expanded the evaluation scope to cover a heterogeneous multi-domain suite of test sets. This year's framework featured five common domains shared across both target languages (English→German and English→Chinese): TV Series, Scientific Presentations, Call Center two-person conversations, YouTube videos, and Business News. Additionally, the Accented English Conversations dataset was included among the test data for the English→German direction. Since these evaluation materials are identical to those utilized in the Offline task for the respective language pairs, we refer the reader to Section I for a comprehensive description of the datasets, including their origin, acoustic conditions, and specific challenges.

Training and Development Data As in the previous campaign, participants could submit systems developed under two distinct data conditions, which were differentiated by the resources permitted to assist the model compression workflow (such as post-compression fine-tuning after

[23] https://speechm.cloud.cyfronet.pl/0000005

pruning or quantization, or student-model training via knowledge distillation from the larger teacher model). Under the **constrained** setting, participants were strictly limited to utilizing the ACL60/60 dataset.[24] This corpus, domain-consistent with one of the evaluation sets (Scientific Presentations), features identical sizing and source audio content for both language directions. Conversely, the **unconstrained** condition imposed no restrictions on data usage, thereby allowing teams to leverage any external or proprietary resources to optimize their compressed models.

4 Baselines

As baselines for the task, we provide two models. The first baseline is the **full precision** model. The second is a **4-bit** quantization of the model. We adopt a 4-bit quantization strategy implemented through the bitsandbytes[25] library. The model weights are loaded in 4-bit precision (`load_in_4bit=True`). We use the NormalFloat4 (NF4) quantization format (`bnb_4bit_quant_type="nf4"`), a data type specifically designed for normally distributed neural network weights (Dettmers et al., 2023), as NF4 has been shown to provide better reconstruction fidelity than standard uniform INT4 quantization for large language models. To further decrease storage, we enable double quantization (`bnb_4bit_use_double_quant=True`), where the quantization constants themselves are quantized. Computations during inference and adaptation are performed in half precision (`bnb_4bit_compute_dtype=float16`), as operations in lower bit rates are more efficient only on recent GPUs and not in older ones used by the organizers (ie. Ampere GPUs).

In both cases, the inference has been conducted on segments obtained by segmenting the audios with SHAS (Tsiamas et al., 2022) using 18 seconds as maximum value for all datasets but CHALLENGEACCENT, which contains short audio segments.

5 Submissions

The task has received 3 submissions, two for the en-de language pair, one of en-zh:

KIT Two submission have been prepared: contrastive1 and contrastive2, which has been elected as the primary, since it was the last submission. In both cases, the base model has first been fine-tuned for the speech translation task, to start from a stronger model before compressing it. Their submission mostly relies on HQQ quantization, whihc minimizes the error in the reconstructed weights, where the error is defined as the euclidean norm of the difference between the original weight matrix and the reconstructed one after quantization (Badri and Shaji, 2023). 4-bit HQQ is applied to the linear layers, while 2-bit quantization to the embedding table, using a group size of 128, which means that the quantization parameters (weights and scales) are computed separately for blocks of 128 weights). Lastly, they select 4 MLP layers to quantize with 3 bits instead of 4 bits, using sensitivity selection, in which the cross-entropy with respect to the full precision model on the validation set is the sensitivity score. This corresponds to their contrastive1 submission, while the primary submission (contrastive2) includes a further calibration step based on AWQ (Lin et al., 2025).

APG This submission focuses on a selective quantization of the model for en-zh. Specifically, it applies compression strategies only to the weight matrices of linear layers. It validated both compressing all linear layers and selecting only the linear layers correpsonding to the MLP component of the Transformer layers. The compression is done using the quantization provided by the `numcodecs` library with either 2 o 3 bits. In addition, it further reduces the storage size of the resulting weight representation by means of the Zstandard compression algorithm.

TalTech This submission focuses on efficient neural model compression through a combination of low-bit quantization and adaptation techniques for en-de. It applies uniform INT4 quantization to substantially reduce model storage. To mitigate the degradation typically introduced by aggressive quantization, it employs quantization-aware parameter-efficient fine-tuning (PEFT), enabling the model to adapt to low-precision representations with minimal trainable parameters. Lastly, it incorporates self-distillation, where the compressed model is guided by the predictions of the original higher-precision model under controlled optimization constraints.

[24] https://aclanthology.org/attachments/2023.iwslt-1.2.dataset.zip

[25] https://github.com/bitsandbytes-foundation/bitsandbytes

6 Results and Discussion

Appendix B.5 reports the COMET scores obtained by the submitted systems for the two language pairs. Overall, the proposed quantization approach achieves substantial model size reductions while largely preserving translation quality. In particular, the compressed models require 4–6GB of storage, with a 2.7-4X reduction with respect to the full-precision baseline.

For the English–Chinese (en–zh) direction, the 4-bit baseline surprisingly achieves slightly higher scores than the corresponding full-precision baseline. This result suggests that the NF4 quantization can act as a form of regularization, mitigating overfitting effects while preserving the capacity of the original model. In contrast, for English–German (en–de), compression leads to consistent but moderate degradations across all datasets. This different behavior across language pairs may be related to the fact the Qwen has much larger Chinese data in its training than German data and might uncover another source of exacerbation of the performance gap between high-resource and lower-resource languages. However, this finding has to be confirmed with a broader investigation covering more models, language pairs, and test sets.

Looking at the en-de results in Table 29, KIT obtains the strongest overall performance across most datasets. However, its results on the CHALLENGEACCENT subset are noticeably weaker. A likely explanation is that their initial fine-tuning did not adequately cover accented speech conditions, although a more detailed analysis would be required to confirm this hypothesis. Looking at the results of other models, it does not seem likely that the gap is due to a peculiarity of the quantization procedure that makes the model weaker for accented data. TalTech exhibits a markedly different behavior, achieving competitive performance only on CHALLENGEACCENT, where it outperforms all other submissions by a large margin. Notably, this dataset consists primarily of short audio segments. Inspection of the generated outputs suggests that the system does not segment long recordings into smaller chunks, instead producing a single output sequence per audio file. This behavior likely explains the strong degradation observed on datasets containing longer utterances, where the model under-generates.

Moving to en-zh, a similar issue (under-generation) is observed for the APG submission. The total number of generated characters is substantially lower than that of the baselines (39k characters vs 86k–98k), indicating severe under-generation. As a consequence, direct comparison between systems becomes difficult, since some of the lower evaluation scores may reflect incomplete outputs due to suboptimal splitting of the source audio rather than pure translation quality differences.

Overall, the shared task results demonstrate that aggressive quantization can reduce model size by 4 times while retaining performance close to, and in some cases exceeding, that of the full-precision models. Quantization has emerged as the to-go approach for all submission, while pruning has not been equally investigated. These findings highlight the practical viability of quantization-based approaches for deploying large speech translation systems in resource-constrained environments, although they also demonstrate that compressing for very low-resource settings, where a 4GB system is still too big, is a challenging problem yet to be addressed for speech translation.

Track IV Subtitling track

1 Introduction

In recent years, the task of automatically creating subtitles for audiovisual content in another language has gained significant[mn] attention due to the rapid increase in the global distribution and streaming of movies, series, and user-generated videos. Reflecting these trends, the automatic subtitling track[mn] was introduced for the first time in 2023 (Agarwal et al., 2023) and proposed again in 2024 (Ahmad et al., 2024) and 2025 (Abdulmumin et al., 2025) as part of the IWSLT Evaluation Campaigns. For this year's round of the task, participants[mn] were asked to generate subtitles in a target language chosen from a pool of five (Arabic, Chinese, German, Japanese, Spanish) from English speech in audiovisual content[mn].

2 Task Description

The task of automatic subtitling is multifaceted: starting from speech, not only must the translation be generated, but it must also be segmented into subtitles that comply with constraints ensuring a high-quality user experience. These constraints include proper reading speed, synchrony with the voices, the maximum number of subtitle lines, and the maximum number of[mn] characters per line. Most audio-visual companies define their own subtitling guidelines, which can differ slightly depending on the content type, platform requirements, and the cultural or linguistic expectations of the target audience.[mn] We asked IWSLT participants to generate subtitles according to specific guidelines provided by TED[26] and Netflix (for Japanese[27] and Chinese[28]), including:

- Maximum subtitle reading speed:
  - 21 characters per second for Arabic, German and Spanish
  - 4 characters per second for Japanese (half-width characters counted as 0.5)
  - 9 characters per second for Chinese
- Maximum line length:
  - 42 characters per line for Arabic, German and Spanish
  - 13 characters per line for Japanese (half-width characters counted as 0.5)
  - 16 characters per line for Chinese

  including white spaces
- Maximum lines per subtitles: 2 for all the languages[mn]

Participants were expected to use only the audio track from the provided videos (dev and test sets), as the video track could be either of low quality and primarily intended to check the temporal synchronicity and other aspects of displaying subtitles on screen, or not provided at all.

Subtitles had to be generated for three kinds (domains) of audiovisual[mn] documents, all featuring English as the spoken language:

- **ITV**[29] entertainment series, to be subtitled in Chinese, German, Japanese, and/or Spanish (Europe)
- economic news programs from the **Asharq-Bloomberg** platform,[30] to be subtitled in Arabic, Chinese, German, and/or Japanese
- audio recordings from the **YODAS** YouTube dataset,[31] to be subtitled in Chinese, German, and/or Japanese.

Audio-visual documents of development and evaluation sets were provided in MP4 format (asharq-bloomberg and ITV) and WAV format (YODAS); subtitles of development sets were released in SRT (SubRip File Format) UTF-8 encoded files, the same format required for submissions.

3 Data and Metrics

Data. This track proposed two training data conditions:[32]

- **Constrained**: the official training data condition, in which the allowed training data is limited to a medium-sized framework to keep the training time and resource requirements manageable;
- **Unconstrained**: a setup without data restrictions (any resource, pre-trained language models included, can be used) to allow also the par-

[26] https://www.ted.com/participate/translate/subtitling-tips
[27] https://partnerhelp.netflixstudios.com/hc/en-us/articles/215767517-Japanese-Timed-Text-Style-Guide
[28] https://partnerhelp.netflixstudios.com/hc/en-us/articles/215986007-Chinese-Simplified-Timed-Text-Style-Guide
[29] https://www.itvstudios.com
[30] https://asharqbusiness.com
[31] https://huggingface.co/datasets/espnet/yodas
[32] https://iwslt.org/2026/subtitling#training-and-data-conditions

ticipation of teams equipped with high computational power and effective in-house solutions built on additional resources.

For each language and domain, a development set (**dev2026**) and an evaluation set (**test2026**) were released; in addition, where available, test sets of previous evaluations (**tst2023**, **tst2024** and **tst2025**) were provided as well for measuring progress over years. Table 8 shows some statistics on these sets.

task	set	AV docs	hh: mm	#ref subtitles ar	de	es	ja	zh
ITV	dev26	3	2:25	-	1516	1526	1613	1522
	tst23	7	6:08	-	4806	4896	-	-
	tst24	7	5:54	-	4568	4532	-	-
	tst25	3	2:07	-	1845	-	-	-
	tst26	3	2:16	-	1488	1507	1511	1506
YO-Asharq-DAS-Blmbrg	dev26	2	3:02	2974	3676	-	3796	3329
	tst25	2	3:03	2759	3543	-	-	-
	tst26	1	1:34	1476	1686	-	2262	2281
	dev26	6	1:40	-	2344	-	2746	2442
	tst26	5	1:35	-	1795	-	2579	2497

Table 8: Statistics of the dev and evaluation sets for the subtitling task.

Metrics. The evaluation was carried out from three perspectives, subtitle quality, translation quality, and subtitle compliance, through the following automatic measures:

- Subtitle quality vs. reference subtitles:
  - **SubER (↓)^{mn}**, primary metric, used also for ranking (Wilken et al., 2022);[33]
- Translation quality vs. reference translations:
  - **BLEU (↑)^{mn}**[34] and **CHRF (↑)^{mn}**[35] via sacreBLEU (Post, 2018); BLEU scores are computed using the following tokenization schemes: "13a" for Arabic, German, and Spanish; "ja-mecab" for Japanese; and "zh" for Chinese.
  - **BLEURT (↑)^{mn}** (Sellam et al., 2020).

  Before metric computation, automatic subtitles are realigned with the reference subtitles using `mweralign` (Post and Hoang, 2025), which implements a variant of the `AS-WER` algorithm (Matusov et al., 2005),[36] by applying the tokenization methods listed above.
- Subtitle compliance:[37]
  - rate of subtitles with more than 21/4/9 characters per second (**CPS**);
  - rate of lines longer than 42/13/16 characters, whitespace included (**CPL**);
  - rate of subtitles with more than 2 lines (**LPB**).

4 Submissions

The subtitling track saw the participation of three teams: AppTek, the MT unit of Fondazione Bruno Kessler (FBK), and Huawei Translation Service Center (HW-TSC). The details about the participants' systems are provided below.

AppTek: The AppTek[38] system is a cascade of ASR, MT, and ILS components. To a large extent it follows last year's submission (Petrick et al., 2025), but has been extended by LLM-based automatic post-editing.

ILS is AppTek's Intelligent Line Segmentation. It creates the subtitle structure from timed ASR output, extracts full sentences from it, and then inserts the translations of these sentences with proper line breaks as predicted by its neural classifier, at the same time respecting subtitling constraints. Subtitle timings are set from the word timings as predicted by the ASR system, but are extended where possible to fulfill a reading speed limit of 17 characters per second on the English side before translation.

The ASR system is AppTek's hybrid ASR production model. For ITV and YODAS, a variant of this model is used that is additionally trained on media and entertainment content from AppTek's customer Deluxe. Separate punctuation, casing, and inverse text normalization models are used to post-process the raw ASR output.

The MT systems are variants of Transformer Big that support additional metadata inputs, in particular genre, speaker gender, and length class. Genre is set to "dialogs" for the ITV and YODAS domains and "news" for the Asharq domain. For ITV English-to-Spanish, automatic gender detection from speech is used to set the MT gender input. In cases where an initial translation cannot be distributed into the source subtitle structure, the length parameter is used in an iterative way to shorten the translation until ideally all subtitle constraints (CPS, CPL, LPB) are fulfilled.

[33] github.com/apptek/SubER

[34] sacreBLEU signature: `nrefs:1|case:mixed|eff:no|tok:X|smooth:exp|version:2.0.0`

[35] sacreBLEU signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0`

[36] www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

[37] github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/subtitle_compliance.py

[38] www.apptek.ai

For Japanese and some of the German submissions, strict following of the CPS constraint is disabled as it avoids a large drop in MT metrics and also matches the low compliancy of the human references. For YODAS, AppTek's general domain production models are used. For ITV, models fine-tuned on large amounts of subtitle content provided by Deluxe are used. For Asharq, the English-to-Arabic system is fine-tuned on financial news data available to AppTek as part of their cooperation with Asharq business with Bloomberg. For German and Japanese, fine-tuning was done by mining in-domain data from the mixed-domain training data based on embedding similarity to the Asharq dev2026 set. For Chinese, the general domain MT is used as the above fine-tuning did not yield improvements.

For the domains of YODAS (all target languages) and ITV (except Japanese), AppTek's primary submissions also include automatic post-editing of AppTek's NMT output with OpenAI's gpt-4o model. For this, each English SRT file (obtained from ASR+ILS) is first split into several parts using a dynamic programming segmentation algorithm that favors long pause durations between parts but penalizes too few or too many sentences in each part (with a target of 20 sentences per part). Processing parts of approximately this size allows the LLM to exploit enough context to produce high-quality post-edits, while avoiding hallucinations or failure to follow prompt instructions. For each part, full sentences and their NMT translations obtained as described above are sent to OpenAI's API with prompt instructions to fix errors only when absolutely necessary, including errors propagated from ASR, while not altering the order, number, and length of the translated sentences. The post-edits are then processed by ILS, so that the resulting subtitles have proper line segmentation and fulfill CPL and, in most cases, LPB constraints. For the Asharq domain, no such automatic post-editing was performed. Here, contrastive submissions either differ in the MT fine-tuning domain (general for Ja, Deluxe media for Zh), or consideration of the CPS limit for MT length control (disabled for Ar, enabled for De).

FBK: The FBK automatic subtitling system is a two-stage ASR-MT cascade built exclusively from freely available open-source components (Cettolo et al., 2026b).

The first stage constitutes the baseline pipeline used for the `Contrastive 1` submission. Audio is first segmented by SpeechBrain VAD (SB VAD), then transcribed by Whisper large-v3, which natively produces time-aligned SRT output. Each subtitle block is subsequently translated by MADLAD-400-10B-MT using 4-beam decoding, and the resulting translations replace the source transcripts in the SRT file.

Stage 2 introduces a sentence-aware refinement step aimed at improving textual quality; its output was submitted as the FBK `Primary` run. Adjacent SB VAD segments are aggregated into longer units, subject to proximity (< 10 s gap) and duration (≤ 600 s) constraints, and re-transcribed using VOXTRAL, an ASR model whose 32k-token context window yields higher accuracy on long-form audio than Whisper-based models. The resulting text is split into sentence units by punctuation cues and translated by MADLAD-400-10B-MT. The sentence-level translations are then word-level realigned to the Stage 1 subtitles via `mweralign`, and inserted into the original SRT template, preserving subtitle synchronization while leveraging broader linguistic context.

Both stages are followed by post-processing steps that enforce subtitle compliance: end timestamps are extended where reading speed thresholds are exceeded (21/9/4 cps for non-CJK, Chinese, and Japanese, respectively); overly long subtitles are split across lines or blocks; and consecutive n-gram repetitions - a known hallucination pattern of both ASR and MT models - are removed.

FBK also submitted a `Contrastive 2` run generated by a system that follows the same baseline architecture but relies on different models: SHAS for audio segmentation, Faster Whisper for ASR, and a smaller MADLAD-400 (3B) model for MT.

HW-TSC: The HW-TSC system (Lan et al., 2026b) adopts a cascaded pipeline for automatic subtitle generation, articulated into three main stages: streaming speech recognition, machine translation, and subtitle compression.

The speech recognition stage is built upon the Qwen3 model family and integrates several components designed for long-form audio processing. Voice activity detection first filters out silent segments, after which a sliding-window mechanism performs streaming audio segmentation with two-second windows, maintaining cached con-

textual information to ensure output coherence. During inference, audio features extracted by the encoder are fused with textual prompts and decoded efficiently via the vLLM framework, with length-based filtering applied to suppress hallucinations. A forced alignment module then establishes token-level temporal synchronization between the transcribed text and the corresponding audio signal, producing timestamped transcripts of high precision.

A text preprocessing step follows, merging consecutive ASR segments at the semantic level based on sentence-ending punctuation and enforcing word-count constraints informed by the length ratio between English and the target (Chinese) language. This ensures that subtitle blocks are both semantically coherent and compliant with display space limitations before being passed to the translation module. Machine translation is handled by a dedicated Qwen3-based model. Since timestamps are already accurately assigned at the recognition stage, only the textual content is translated while all temporal information is preserved unchanged.

The final stage addresses subtitle compression. Non-compliant segments, those exceeding the CPS or CPL thresholds, are identified automatically and subjected to a two-stage rewriting procedure using Qwen3-32B. A first pass applies greedy decoding at temperature 0, removing only redundant auxiliaries and conjunctions; if constraints are still unsatisfied, a second pass with temperature 0.3 performs deeper compression, always retaining proper nouns and core semantic content.

HW-TSC submitted `Primary` Chinese subtitles generated for each of the three domains.

5 Results

Scores on submitted runs are shown in Tables 31, 32 and 33 for Asharq-Bloomberg, YODAS and ITV domains, respectively.

The YODAS domain was introduced this year, so no comparison with previous editions is possible. In contrast, for Asharq-Bloomberg and especially for ITV, primary runs from systems that participated in past editions of the shared task are available. Specifically, for Asharq-Bloomberg, which was introduced in 2025 (for Arabic and German), results are available from APPTEK, the sole participant in that edition; for ITV, which has been part of the shared task since its inaugural edition in 2023 (for German and Spanish), results are available from all participants across the various test sets and editions. To avoid overloading the tables, we restrict the reported ITV results to those obtained on the tst23 set, omitting those on more recent legacy test sets, namely tst24 and tst25.

Starting from the tst26 results, the most immediately striking observation concerns Japanese subtitles across all three domains, which exhibit comparatively low scores in terms of SubER, translation quality metrics, and CPS. Regarding CPS in particular, the phenomenon may be a direct consequence of the extremely strict threshold (4 cps) imposed for Japanese, especially given that the reference subtitles themselves display an unusually low CPS. For instance, the CPS of the ITV dev26 reference SRT files is 44.89%. A modest relaxation of the reading speed threshold, from 4 to 6, raises that figure to 95.74%, supporting the intuition that low CPS scores on submitted runs may reflect the severity of the threshold rather than actual subtitle non-compliance.

The bad SubER and translation quality scores observed in general for automatically created Japanese subtitles may be attributable to word-level segmentation: Japanese does not natively use whitespace to delimit word boundaries, making explicit segmentation necessary both for subtitle alignment (via `mweralign`) and for score computation. Ambiguities or even errors in automatic word segmentation could negatively affect both operations, and this may account for the observed issue. In support of this hypothesis, we consider ChrF scores, which are sensitive to word-level segmentation during hypothesis re-segmentation against references but not during evaluation itself. The official ChrF scores of APPTEK's primary ITV runs for Japanese and Chinese are 12.19 and 28.14, respectively (Table 33). By bypassing Japanese word-level re-segmentation and evaluating hypotheses against references as single long character-level strings, ChrF rises to 28.18 for Japanese and 35.52 for Chinese. Note that Chinese evaluation is performed directly at the character level and is therefore largely unaffected by word-segmentation ambiguities. The resulting reduction of the gap between the two languages from 16 to 7 points suggests that the underlying translation quality is not as divergent as the official scores would indicate.

Across primary runs on tst26, the following ob-

servations can be made:

- APPTEK systems rank first under virtually all metrics, with only two exceptions:
- for Chinese subtitles in the Asharq-Bloomberg domain, where HW-TSC stands out, particularly in terms of translation quality;
- for German subtitles, again in the Asharq-Bloomberg domain, where FBK achieves notably better overall quality (SubER) and translation scores.

Regarding the comparison with past editions in the Asharq-Bloomberg domain, APPTEK's 2026 and 2025 systems perform on a par on the Arabic tst25 section (SubER of 61.64 and 62.13, respectively), whereas on the German section this year APPTEK seems to have prioritised translation quality over compliance: BLEURT improved from .6020 to .6234, while CPS dropped from 92.44 to 73.50.

For ITV, the numerous runs available on tst23 for both German and Spanish reveal a general upward trend in performance over the years. Translation quality in particular has improved substantially: the best primary BLEURT scores in 2023 were .4438 for German and .4530 for Spanish, rising to .5520 and .5514, respectively, in 2026. SubER and CPS, by contrast, are more volatile, with peak values not necessarily achieved by the most recent systems, a sign that the optimal trade-off between translation quality and compliance with subtitling spatiotemporal constraints has yet to be conclusively resolved.

Track V Simultaneous track

1 Introduction

Simultaneous speech translation (SIMULST) translates source-language audio into target-language text concurrently with the incoming speech. Similar to offline speech translation (*ST*), SIMULST aims to maximize translation quality. At the same time, the *real-time* constraint requires the system to produce output immediately as the speech is being received (Laplante, 1992). The core challenge of SIMULST is to balance translation quality and latency: the longer the system waits to acquire more context, the better the translation quality, but the higher the latency.

Historically, this quality-latency tradeoff was evaluated in a simplified setting where the audio input was pre-segmented into small chunks of a few seconds, typically aligning with sentence boundaries. However, for the second consecutive year, the focus of the IWSLT Simultaneous Speech Translation Shared Task remains on the more realistic setting (Papi et al., 2025) where evaluation is conducted on raw audio streams without any pre-segmentation. A novelty of this year's task is the introduction of an *Extra Context* track, where systems are provided with additional context to improve translation quality.

2 Task Description

The 2026 IWSLT SIMULST shared task is divided into two tracks:

- **Speech-to-Text:** The standard SIMULST task.
- **Speech-to-Text with Extra Context:** A track where systems are additionally provided with relevant context to improve translation quality.

The task continues to focus on the **long-form speech** setting, where systems operate on raw, unsegmented audio streams. Participants are permitted to use publicly available Large Language Models (LLMs), including speech-based foundation models. The task features two data conditions: *constrained with LLMs* allowing the standard constrained data[39] plus any open-weight Large Language Model under a permissive license, and *unconstrained* permitting any external resource except the official evaluation sets. Submissions leveraging closed-source models are excluded from the main ranking.

[39] https://iwslt.org/2026/simultaneous#training-data-and-data-conditions

2.1 Latency Regimes

Participants' systems are evaluated in one of two latency regimes that are *shared across all language pairs*:

- **Low Latency**: 0–2 seconds.
- **High Latency**: 2–4 seconds.

Latency is measured using the non-computation-aware LongYAAL (Polák et al., 2025) on the development set, which is then used to assign systems to their respective latency regimes.

2.2 Submission Format

Participants are allowed to use the SimulStream or SimulEval toolkits, although the use of SimulStream is strongly encouraged. Participants are offered two options for submitting their simultaneous translation systems:

- **Docker Image Submission (Preferred):** Under this setup, the organizers run the participants' systems in a controlled environment equipped with a single NVIDIA H100 GPU (80 GB of memory). This allows for a direct comparison of computation-aware latency metrics.
- **System Log Submission:** Alternatively, participants can submit their system translation outputs and timestamps generated locally using either the SimulStream log format or the legacy SimulEval JSONL format. While computation-aware latency is reported, it cannot be compared directly due to hardware discrepancies.

Regardless of the submission format, participants must submit their translation logs on the development set—MCIF (Papi et al., 2026b) or the dedicated Czech-to-English dev set—to determine the latency regime of their systems.

3 Data and Metrics

3.1 Data

The evaluation features four language directions: English→{German, Chinese, Italian} and Czech→English. The evaluation datasets consist of long-form, unsegmented audio recordings of talks or news broadcasts.

For the English-to-X directions, the evaluation datasets are as follows:

- **Main Evaluation Domain (ACL Talks):** Comprises oral presentations from ACL conferences. For the *Speech-to-Text with Extra Context* sub-

track, the systems are provided with the corresponding ACL paper PDFs.

- **Optional Evaluation Domain (Asharq-Bloomberg News):** A single two-hour recording of business news produced by Asharq Business with Bloomberg.
- **Optional Evaluation Domain (YODAS):**[40] Five audio recordings (ranging from 10 to 30 minutes) from the YouTube-Oriented Dataset for Audio and Speech. The "en003" partition of the dataset is strictly held out from training.

For the Czech-to-English direction, the datasets comprise:

- **Main Evaluation Domain (Political Conference Talks):** The development set consists of recordings of the Chamber of Deputies' meetings from the Czech Parliament (2024–2025). The test set comprises recordings from the *Media and Ukraine* conference held in Prague in June 2025.

3.2 Metrics

Each system is evaluated in terms of quality and latency. The entire evaluation is conducted using the OmniSTEval[41] toolkit on re-segmented outputs. This re-segmentation is based on the alignments between reference translations and system outputs, computed via SoftSegmeter (Polák et al., 2025). For translation quality, the primary metric is COMET-XL (Guerreiro et al., 2024). For consistency with previous years, we also report SacreBLEU. For latency, we report LongYAAL as the primary metric, alongside StreamLAAL (Papi et al., 2024)[42] for comparison.

4 Submissions

CUHKSZ **(Yang and Nakamura, 2026)** participated in the English→{Chinese, German} language directions in the standard and extra-context tracks. Their end-to-end streaming agent is developed with the multimodal Qwen3-Omni-30B-A3B model (Qwen et al., 2025) as the backbone. The agent generates translation hypotheses by 1) feeding each incoming audio window as a new user turn in a multi-turn conversation setup, and 2) applying a strict set of emission controls to ensure every decision step complies with the computation-aware latency budget. During inference time, the system relies on the internal policy generated by the LLM itself—i.e., a dedicated `<wait>` token for the READ action, and WRITE otherwise—which is fine-tuned on syntax-aware chunks (Yang et al., 2026). For the extra context track, along with the audio chunk, the agent extracts per-talk named entities (for low-latency) or full paper abstracts (for high-latency), and injects them to the system prompt.

CPII-HK **(Xing et al., 2026)** participated in all language pairs of the standard track. Their solution adopts a test-time approach derived from the wait-k policy (Ma et al., 2019) applied to offline models (Papi et al., 2022), specifically Qwen3-Omni (Qwen et al., 2025). The inference is conducted by using a multi-turn conversation strategy (Ouyang et al., 2025), where each user message contains a new audio chunk and each assistant response provides the accumulated translation. The long-form processing is achieved by leveraging a VAD, Silero (Silero Team, 2021), in a hybrid manner (Gaido et al., 2021a) with minimum and maximum durations. Since multi-turn conversations leads to growing context, response prefilling and KV caching (Pope et al., 2023) were used.

MLLP-VRAIN UPV **(Iranzo-Sánchez et al., 2026)** participated in all language direction in the standard track, and in English→{German, Italian, Chinese} of the context-augmented track. The submission is based on the last year's system (Iranzo-Sánchez et al., 2025), specifically, on the cascade of Parakeet (Sekoyan et al., 2025a) and Qwen 3.5[43] models. The system adopts a Longest Common Prefix (LCP) policy (Liu et al., 2020; Polák et al., 2022) with relaxed constraints, SoftLCP, leveraging the Ratcliff/Obershelp (RO) pattern recognition algorithm (Ratcliff and Obershelp, 1841). The core idea to identify "anchor" tokens via RO and greedily accept all preceding tokens, propagating committed output more frequently than regular LCP. The long-form processing is based on an audio buffer that accumulates the input chunks, pushing old chunks out of the buffer when the maximum length is reached. At decoding time, the entire content of the buffer is provided to the model, and, to avoid output repetitions, timestamp-based repetition control is adopted by selecting tokens with time overlaps of

[40] https://huggingface.co/datasets/espnet/yodas

[41] https://github.com/pe-trik/OmniSTEval

[42] Notably, StreamLAAL uses a different segmenter, mWERSegmenter, instead of SoftSegmeter.

[43] https://qwen.ai/blog?id=qwen3.5

the preceding decoding steps.

PINCH-AST **(Bentes and Safka, 2026)** participated in all four language pairs in the standard track. Their cascaded system consists of three dedicated components: 1) an ASR model of either Parakeet-TDT-0.6B-v3 (for the Czech→English direction) or Qwen3-ASR-1.7B (for the other three; Shi et al. 2026); 2) a Qwen3.5-4B-based translation model; and 3) the SimulStream emission layer (Gaido et al., 2025) with character-level and CJK-aware LCP re-translation policy. Long-form audio processing is handled by feeding the entire accumulated audio chunk since the last utterance boundary and its full transcription as the current hypothesis to the ASR model, while also applying character-level LCP across consecutive hypotheses to identify the stable transcript prefix.

NEMO **(Grigoryan et al., 2026)** participated in all language directions in the standard track, extra context track, and in the optional tracks for Bloomberg News and YODAs. Their submission is a cascaded system featuring models from the Parakeet family for ASR[44] and Qwen 3.5-based LLMs[37] used for translation. Regarding long-form segmentation, they relied on the NVIDIA NeMo streaming inference framework (Kuchaiev et al., 2019) with translation beginning upon consecutive end-of-utterance tokens being emitted by the transducer. For translation, they leverage the LCP policy. Regarding the extra context track, their submission differs from the baseline in that the context extraction pipeline is more fine-grained and engages in further preprocessing and postprocessing of the extracted text.

CUNI-POCKET **(Ortega and Macháček, 2026)** participated in three of the four language pairs in the standard track (English→Chinese is not included). Their system is particularly notable for how compact and efficient it is: the neural network component is a Canary model composed of 1B parameters,[45] making it particularly suitable for edge devices and other low resource applications. Regarding the structure of their system, they customize the Canary 1B model to incorporate AlignAtt (Papi et al., 2023) for making emission decisions. Similar to some other works, they maintain a sliding window of context on the audio side, accumulating the last 30 seconds of audio for use by their system. This is all executed on an adapted version of the NVIDIA NeMo streaming inference framework (Kuchaiev et al., 2019).

CUNI-ALIGN **(Fuxa and Macháček, 2026)** participated in three of the four language pairs in the standard track (Czech→ English is not included). This system positions itself uniquely as one that leverages decoder-only LLMs but still integrates an adapted version of the AlignAtt policy for emission decisions (Papi et al., 2023). Their system is composed of the Qwen3-ASR-1.7B model (Shi et al., 2026) cascaded with Gemma 4 E4B-it[46] for simultaneous translation. The adaptation of AlignAtt to decoder-only models is done by providing the source transcript, produced by Qwen3-ASR, in the prompt (together with the already accepted target prefix, and a fixed translation instruction), selecting a small set of translation-specific self-attention heads offline (following BINBINLIU et al. 2026), and accept only partial translation hypotheses whose reconstructed attention signal remains within the currently available source frontier. To make the cascaded system fast, vLLM (Kwon et al., 2023) is used as a backbone.

UW **(Pong, 2026)** participated in three of the four language pairs in the standard track (Czech→ English is not included). The system uses dynamic attention masking to constrain the encoder lookahead during training based on the predictions of per-layer schedulers injected into its Conformer encoder. It is built upon SeamlessM4T-medium (Seamless Communication et al., 2023) with full fine-tuning of the additional scheduler parameters and LoRA adapters applied to the self-attention layers. In inference, two different strategies are compared: the sliding window retranslation (Sen et al., 2022) and StreamAtt (Papi et al., 2024) with a modification of its cutoff implementation using the proposed schedulers.

5 Results

The results for English→German are available in Table 35, for English→Chinese in Table 36, for English→Italian in Table 37, and for Czech→English in Table 34. We also provide quality-latency tradeoff figures in Figures 4 to 7.

[44] https://huggingface.co/nvidia/parakeet-unified-en-0.6b

[45] https://huggingface.co/nvidia/canary-1b-v2

[46] https://huggingface.co/google/gemma-4-E4B-it

Main Results Across all language pairs and evaluation sets, the expected quality-latency tradeoff between the low- and high-latency regimes is generally observable, though it is not uniform across all systems. In most cases, the high-latency regime yields superior translation quality compared to the low-latency regime, although the performance gains are often modest for the strongest systems. Due to the aforementioned latency measurement issues, several systems exhibit higher latencies on the test sets than their development set metrics initially suggested at the time of submission, explaining the unusually large latency values reported in the results tables.

Based on the results within individual latency regimes (see Tables 34 to 37), the NEMO submission ranks first, demonstrating highly consistent performance across the various evaluation sets and latency regimes. In terms of absolute performance, MLLP-VRAIN UPV delivers robust results in the high-latency regime. However, it shows a more pronounced quality drop in the low-latency regime compared to all other submissions. The baseline system is consistently outperformed by most submissions, often by substantial margins, validating the efficacy of the proposed methodologies.

Quality-Latency Tradeoff Based on the quality-latency tradeoff (see Figures 4 to 7), we observe that the NEMO submission is part of the Pareto frontier in all language pairs. However, other submissions, such as CUHKSZ in English→German, CUHKSZ, CPII-HK, and MLLP-VRAIN UPV in English→Chinese, and CUNI-ALIGN and MLLP-VRAIN UPV in English→Italian, are also part of the Pareto frontier.

Out-of-domain Latency Spike Interestingly, the YODAS test set consistently induces the highest latencies across all language pairs, averaging approximately 1.0 second higher than the other evaluation sets. This latency increase is likely driven by the domain mismatch and distinct acoustic characteristics of the YouTube-sourced YODAS data compared to the other talk- and news-oriented datasets.

Computation-aware Latency Computation-aware latency metrics are generally not available for all submissions, therefore we cannot draw firm conclusions on the computation-aware latency performance. On average, the computation-aware latency is approximately 0.3 seconds higher than the non-computation-aware baseline. Interestingly, the compact CUNI-POCKET submission, the only end-to-end submission, which relies on a single 1B-parameter model, exhibits the largest difference between computation-aware and non-computation-aware latency (by more than 0.4 seconds), despite expectations that smaller models would incur lower computational overhead.

Ranking and Metric Correlation Rankings on the development set generally mirror those on the test sets, indicating the robustness of the evaluation setup and the absence of obvious overfitting. The only exception is the MLLP-VRAIN UPV system, which ranks higher on the development set than on the test sets. While chrF, BLEU, and COMET scores are generally well-correlated, some exceptions exist. For example, in many cases, the MLLP-VRAIN UPV system achieves higher lexical metric scores (chrF and BLEU) than other systems that are ranked higher according to COMET. This discrepancy is most prominent in the English→Chinese direction, where MLLP-VRAIN UPV outperforms the higher-ranked NEMO submission by an average of 1.6 chrF and 1.7 BLEU points, despite scoring 0.067 points lower in COMET.

Impact of additional context This year's simultaneous translation shared task introduces a subtrack in which participants may use additional context relevant to the source speech. For ACL talks, this context is the corresponding paper PDF. We received three submissions to this subtrack, from CUHKSZ, MLLP-VRAIN UPV, and NEMO. In the low-latency regime, the NEMO submission achieves the highest COMET scores across all three language directions, whereas the MLLP-VRAIN UPV achieves the best COMET scores in the high-latency regime. To isolate the impact of context, we compare each team's submissions with and without context (see Figures 8 to 10). Almost all submissions show a positive impact of additional context, with the exception of the CUHKSZ submission in the English→Chinese direction, which shows a slight decrease in COMET score when using additional context. MLLP-VRAIN UPV shows the largest gain, improving by an average of 2.75 COMET points across all three language directions and la-

tency regimes, whereas the NEMO submission performs comparably to its context-free counterpart. We find that the additional context has a positive impact on entity recognition and translation if used properly, which in turn benefits overall translation quality, as exemplified by the MLLP-VRAIN UPV submission. The NEMO submission on the English→Chinese direction exhibits an unusually high latency of 10.86 seconds and case studies reveal that it often buffers an entire Chinese sentence and emits it well after the corresponding speech has ended.

6 Conclusions

The submissions highlight a clear trend toward incorporating large language models. Furthermore, with the exception of a single end-to-end submission, all participating teams adopted cascaded approaches. Although the quality-latency tradeoff remains generally observable across systems, higher latency does not always yield proportional translation quality improvements, particularly for the top-performing submissions.

A key novelty in this year's task was the introduction of the Extra Context track, enabling participants to leverage external document-level information. The integration of this additional context demonstrated clear benefits—particularly for resolving domain-specific vocabulary and translating specialized entities—though the magnitude of these improvements varied significantly depending on the system's integration strategy.

Finally, the performance degradation and latency spikes observed on the YouTube-sourced YODAS dataset underscore the ongoing challenge of handling out-of-domain conditions.

Track VI Indic S2S

1 Introduction

The Indic S2S shared task aims to establish a benchmark dataset and develop speech-to-speech (S2S) translation models for three Devanagari-family Indian languages. Most of these languages are severely low-resource and remain largely unexplored in the context of S2S translation. This poses a unique challenge, as existing pre-trained models often struggle to generalize effectively to such underrepresented languages. An additional difficulty arises from the fact that many of the target languages are linguistically distant from English, further complicating translation tasks. The two major S2S translation architectures, cascaded and end-to-end (E2E) models, each have their own limitations. Cascaded models are prone to error propagation and higher latency due to their multi-stage pipeline (Jia et al., 2019; Gupta et al., 2025). In contrast, E2E models require large-scale parallel speech datasets for training, which are both costly and time-consuming to collect (Jia et al., 2019). Indic languages pose additional challenges, such as morphological variation across scripts and a lack of standardization of their tokens. These challenges also affect the evaluation process.

The dataset we propose will serve as the first benchmark and gold-standard resource for S2S translation across these languages, covering three major Indian languages. By providing this resource, we aim to catalyze the development of robust systems that not only advance research but can also be deployed in real-world applications to improve accessibility and multilingual communication.

2 Task Description

This year, Indic S2S shared track challenged the participants to develop S2S Translation models for Indian languages. The languages include Hindi (Hi), Marathi (Mr), and Punjabi (Pa) from the Devanagari family, which cover a large portion of India's population. However, these languages are low-resource in the S2S domain, which makes them challenging. The objective was to develop either a cascaded or an end-to-end (direct) speech-to-speech (S2S) model for the language pairs (En $\longleftrightarrow$Hi, Mr, Pa). The monolingual or multilingual model submissions were allowed. The S2S translation models are invited in two setups, *constrained* and *unconstrained*:

Constrained In this setup, participants were allowed to use only the provided dataset, which includes speech and their transcriptions. Any external speech or text, or pretrained components, were not allowed, except for the speech quantization module (HuBERT (Hsu et al., 2021) and K-mean clustering model) and the unit-vocoder in the case of direct S2S models.

Unconstrained In the unconstrained setup, participants were allowed to use any external speech or text data, or any pretrained module, to build the model. Speech and text data may include monolingual as well as parallel speech and text data, in addition to the provided dataset. Participants can pretrain the encoder and decoder in an encoder-decoder architecture. However, the final evaluation is done on the provided test set.

3 Data and Metrics

We provided a parallel speech dataset for the three language pairs from Indic-S2ST (Sethiya et al., 2025) shown in Table 9. The dataset includes parallel speech and its transcriptions. The dataset consists of 45 hours of speech data and has 17297 utterances. We provided 3 language pairs (En $\longleftrightarrow$Hi, Mr, Pa), which is an n-way parallel dataset. The dataset consists of a train, validation, and test set; however, we only provided the source and target for the train and validation sets. For the test set, we provided only the source; participants submitted their predictions, and the metrics were evaluated by the Indic S2S organizing team. "'

Table 9: Provided Indic-S2ST (Sethiya et al., 2025) Dataset for Training and Evaluation

Language	Speech Hours		No. of Sentences	
	Indic-S2ST	FLEURS	Indic-S2ST	FLEURS
Hi	46.34	4.82	17,297	1702
Mr	46.72	6.21	17,297	1992
Pa	43.63	5.13	17,297	1588
En	37.22	4.64	17,297	1938

The submitted systems were evaluated on the provided dataset, and an additional evaluation (optional) was also performed on the FLEURS (Conneau et al., 2023) dataset. The translation quality was assessed using sacreBLEU (Papineni et al., 2002), Translation Edit Rate (TER), and BLASER (Dale and Costa-jussà, 2024; Chen et al., 2022)

metrics under both reference-based and reference-free evaluation settings. We also provided a baseline for direct S2S models from Indic-S2ST (Sethiya et al., 2025) to compare the performances as shown in Table 10.

Table 10: Baseline Results of Direct S2S Model

Language Pair	Indic-S2ST	FLEURS
Hi → En	26.08	24.87
Pa → En	19.35	29.18
Mr → En	15.29	13.98

4 Submissions

In the Indic S2S track, three teams with multiple registered members participated in system development. However, none of the teams submitted a system that achieved significant performance.

Track VII African/Celtic S2S

1 Introduction

The African and Celtic Speech Translation track targets a central limitation of current speech and language technology: progress in speech translation is still strongly shaped by the availability of large-scale parallel data, leaving many low-resource language communities underserved. While recent multilingual speech translation systems have expanded coverage, performance and robustness remain uneven for languages with limited speech resources, particularly when systems must preserve culturally specific content, proper nouns, named entities, and phonetic details that are poorly represented in pretraining data.

This track focuses on speech translation for three major Nigerian languages of Hausa, Igbo, and Yorùbá, paired with English. The task is motivated not only by the need for accurate translation, but also by the broader challenge of linguistic and cultural fidelity in low-resource settings. Translation errors in names, local terminology, and culturally grounded expressions can substantially reduce the usefulness of speech translation systems for the communities they are intended to serve. By providing newly recorded parallel speech and text data, the track aims to support more realistic evaluation of both speech-to-text and speech-to-speech translation systems for African languages.

2 Task Description

The official task consists of two modeling tracks. The primary direction for both tracks is from the three source languages, Hausa, Igbo, and Yorùbá, into English. Participants could submit systems for any subset of the language pairs and were allowed to use pretrained models and additional training data.

The first track is *speech-to-text translation* (S2T), in which systems translate source-language speech into English text. This track serves as the foundational setting for the shared task and evaluates whether systems can produce high-quality English translations directly from low-resource speech input. Systems are ranked separately for each language direction.

The second track is *speech-to-speech translation* (S2S), in which systems translate source-language speech into English speech. This track evaluates the more challenging end-to-end scenario in which the output must be intelligible and natural spoken English while preserving the source utterance's content. The dataset includes English speech recorded by native speakers of the African source languages, enabling analysis of how systems handle accented English and paralinguistic aspects of the translation.

Language	Train split (# hours) # Recorded	After QC	Dev-Test split # Dev	# Test	Gender ratio Female:Male
English (en)	30.09	24.96	2.33	2.32	F:19, M:21
Hausa (ha)	58.34	55.06	3.89	5.11	F:47, M:25
Igbo (ig)	65.57	56.89	4.43	4.38	F:47, M:25
Yorùbá (yo)	63.39	61.38	4.38	4.49	F:46, M:26

Table 11: **NaijaS2ST Speech information** before and after Quality Control (QC).

Although the released data also supports translation from English into Hausa, Igbo, and Yorùbá, these reverse directions are considered exploratory. Submissions in these directions may be analyzed, but are not part of the official ranking.

3 Data and Metrics

The track is based on a newly collected parallel speech translation corpus comprising Hausa, Igbo, and Yorùbá paired with English. The source text was originally authored in English; the training sentences are drawn from three existing world-news resources: MAFAND (Adelani et al., 2022), NTREX (Federmann et al., 2022a), and SSA-MT. The test sets were curated from open-source newspapers, with emphasis on arts and culture, business, and sports articles from Voice of America. The test material is balanced between European-context and African-context articles. These sentences are subsequently translated into Hausa, Igbo, and Yorùbá by native speakers. Note that the three African languages are not mutually parallel with one another.

Using this translated text material, approximately 75 hours of speech were newly recorded by native speakers. Each example includes source-language speech, source-language text, English text, and English speech. For each language, the corpus was partitioned into training, development, and test sets containing 5,000, 500, and 500 unique sentences, respectively, with utterance-level segmentation provided. The source-language recordings include up to 70 speakers per language, with held-out speakers in the development and

test sets. In the training split, each low-resource-language sentence is recorded by three different speakers, producing 15,000 recordings per language from 5,000 unique sentences and approximately 60–70 hours of training data per language pair. English recordings are included to represent accented English speech corresponding to the speakers' first-language background. In the test set, each English sentence is recorded in two Nigerian English accent varieties, Northern Nigerian and Southern Nigerian English. The training and development splits were released for system development, while the test split was reserved for the official evaluation period. All speech recordings are provided as 48 kHz `.wav` files, along with text transcriptions and recording metadata. Participants were instructed to shuffle the data before training or testing and to normalize text by removing punctuation, capitalization, and related surface variation before evaluation.

Model performance is assessed using both lexical and embedding-based metrics to ensure a comprehensive evaluation. For S2T, we use SSA-COMET (Li et al., 2025), which is an extension of COMET for African languages. SpBLEU and Chrf++ (Popović, 2015b) are additionally provided for analysis. For S2S, the official metric is character error rate (CER); generated English speech is first transcribed with Omnilingual ASR (OmniASR_LLM_1B), and the resulting transcription is compared against the text reference.

Baseline systems include a fine-tuned SeamlessM4T (Seamless Communication et al., 2023) for S2T, and a cascaded S2T baseline combining Omnilingual ASR with NLLB-200 (NLLB Team et al., 2022). As Hausa is not supported in SeamlessM4T, each language-specific model is trained with a learning rate (LR) of 1e-5, 16 gradient accumulation (GA) steps and 3 epochs.

4 Submissions

This track received one submission from the KIT-RBG-AI team. The team uses AURA-ST (Acoustic-Unconstrained Residual Architecture for Speech Translation), a low-resource S2TT system that directly connects acoustic representations to a large causal language model without relying on cross-attention. The framework consists of three stages: dual-stream feature extraction, modality alignment, and adapter-based fine-tuning.

In order to capture the acoustic diversity of low-resource languages, two complementary encoders are employed. A frozen w2v-BERT 2.0 model (Baevski et al., 2020) extracts 1024-dimensional phonetic and linguistic features, while a ResNet34 encoder (He et al., 2015) captures paralinguistic information such as speaker traits, tonal variation, and dialectal cues. The fused representations provide both high-level linguistic content and fine-grained acoustic context, which is particularly beneficial for languages such as Hausa, Igbo, and Yorùbá.

A trainable convolutional subsampler reduces sequence length by 4× and projects the fused acoustic features into the 2048-dimensional embedding space of a frozen Gemma-4-E2B language model. Rather than using cross-attention, the projected acoustic sequence is treated as a prefix prompt and concatenated with the task instruction and target text before being processed by the causal decoder.

They also apply Low-Rank Adaptation (LoRA; rank 16, α = 32, dropout = 0.05) exclusively to the Gemma MLP layers (gate_proj, up_proj, and down_proj). Standard attention-based LoRA injection was ineffective due to architectural constraints in Gemma's custom projection modules. By adapting only the MLP layers, they enable the model to learn a mapping from acoustic representations to English text while minimizing catastrophic forgetting.

5 Results

Table 12 presents S2TT results for Hausa, Igbo, and Yorùbá across spBLEU, chrF++, and SSA-COMET. The supervised *SeamlessM4T Mono FT* baseline achieves the strongest overall performance, obtaining the highest score across most languages and metrics. This result is unsurprising given that the model is individually fine-tuned for each language pair and therefore represents a practical upper bound under fully supervised conditions. The cascaded system, consisting of Omni-ASR LLM 1B followed by NLLB-200 translation, performs particularly well for Hausa, where it remains competitive with the monolingually fine-tuned SeamlessM4T system. However, its performance deteriorates more substantially for Igbo and Yorùbá, suggesting that recognition errors propagate through the translation pipeline and become increasingly difficult to recover from in lower-

	Hausa			Igbo			Yorùbá		
Model	spBLEU	chrF++	SSA-COMET	spBLEU	chrF++	SSA-COMET	sBLEU	chrF++	SSA-COMET
Aura-ST	5.2	23.2	33.9	4.6	16.7	20.6	19.5	41.0	57.1
SeamlessM4T Mono FT	**18.6**	41.9	**54.9**	**17.6**	**39.2**	**52.3**	**21.1**	**43.5**	**60.3**
Cascaded	17.3	**42.0**	54.1	11.0	33.8	42.9	17.0	38.1	50.6

Table 12: **S2TT results (XX → English)** across three metrics: spBLEU (↑), chrF++ (↑), and SSA-COMET (↑). **Bold** = best overall per column; underlined = second best. SeamlessM4T Mono FT is a fully supervised fine-tuned upper bound, monolingually fine tuned for each individual language pair. The cascaded model is a combination of Omnilingual-ASR (OmniASR_LLM_1B) and NLLB-200 for machine translation.

resource settings. This may also be due to the fact that Yorúbà is the highest resource out of our three studied languages.

AURA-ST exhibits a different performance profile. While its scores lag behind both baselines for Hausa and Igbo, it remains highly competitive for Yorùbá. In the Yorùbá→English direction, AURA-ST achieves the second-best result across all three evaluation metrics and outperforms the cascaded system by +2.5 spBLEU, +2.9 chrF++, and +6.5 SSA-COMET. This finding is notable because AURA-ST does not rely on language-specific supervised fine-tuning or cross-attention-based speech-language fusion. Instead, the system uses frozen speech encoders and a frozen Gemma language model, learning the speech-to-text mapping through lightweight LoRA adaptation of the decoder's MLP layers.

Overall, the results underscore the continuing difficulty of speech translation for African languages, as the nuances in tone and diacritic, for example, are much more difficult to detect than in text. Additionally, despite recent advances in multilingual speech and language modeling, substantial performance disparities remain across languages. The results further indicate that parameter-efficient speech-to-LLM adaptation is a promising direction, particularly for moderately resourced languages such as Yorùbá, but additional advances in multilingual representation learning and low-resource adaptation are needed.

Track VIII Cross-lingual Voice Cloning

1 Introduction

Voice cloning is a type of speech synthesis task that aims to mimic the characteristics of an original voice for a new input text. Cross-lingual voice cloning extends this to target languages different from the source. Current systems have limited capabilities in diverse areas as they support a limited number of languages, struggle with scientific and domain-specific terminology, and show difficulty with code-switching scenarios. The task of voice cloning is used for diverse purposes, including:

- Speaker standardization: In speech systems, voice cloning can be employed as a speaker standardization step, especially in commercial systems that require the final output to match certain voice characteristics.
- Multilingual content creation: Content creators can convert their audio into other languages in their voices.
- Augmentative and alternative communication: In general, speech synthesis can be used in assistive tools. Voice cloning can make the experience more personalized by mimicking the voice of people with special needs or creating age-appropriate voices for teaching children.

Despite recent advances in zero-shot text-to-speech synthesis, cross-lingual voice cloning remains challenging due to the need to preserve speaker identity across phonetically and prosodically distinct languages. In the following sections, we describe the first edition of the Cross-Lingual Voice Cloning Shared Task at IWSLT 2026, which attracted five participating teams who have submitted their systems for Arabic, Chinese, and French. We describe the task setup, evaluation methodology, system submissions, and results.

2 Task Description

The cross-lingual voice cloning task requires systems to synthesize speech in a target language while preserving the voice characteristics of a source speaker. Given a small set of reference utterances from a source speaker in English and target texts in Arabic, Chinese, and French, systems must generate speech that maintains speaker identity (timbre, prosody, and speaking style) while ensuring naturalness and intelligibility in the target language.

3 Dataset

3.1 Training and Development Data

Participants were allowed to use any suitable training and development data for system construction. In addition, the use of the ACL 60/60 dataset[47] (Salesky et al., 2023), comprising translations of ACL 2022 presentations, was recommended due to its alignment with the evaluation domain. Systems are expected to handle cross-lingual voice cloning, where speech synthesis must preserve speaker identity while adapting to a target language. The task explicitly encourages multilingual modeling, although separate language-specific models are also permitted.

3.2 Evaluation Data

The evaluation data consists of English source speech used as reference speaker material and target-language text in Arabic, Chinese, and French. For each target language, a set of 12 reference audio files is provided, each representing a different source speaker. The corresponding reference text contains 49 lines in Arabic, 112 in Chinese, and 99 in French.

Reference audio files were extracted from 12 ACL 2023 presentations in English with diverse speaker accents. Most reference audios include approximately 5 minutes each. While voice cloning typically requires a few seconds of reference audio, we did not trim the reference audio files in order to give the participants the opportunity to investigate different approaches, which indeed benefited some of the submissions.

Target data was collected from various bilingual journals. Scientific papers were sourced from arXiv (Chinese and French), TAL[48] (French), JOS[49] (Chinese), as well as Najah,[50] PalUniv,[51] and PSUT[52] (Arabic). We then filtered out translation pairs with lower semantic similarity using the Sentence-Transformers library (Reimers and Gurevych, 2019) and the LaBSE model (Feng et al., 2022), although for this edition we only provided the target text for simplicity.

[47] https://hf.co/datasets/ymoslem/acl-6060
[48] Traitement Automatique des Langues
[49] Journal of Software
[50] An-Najah University Journal for Research
[51] Journal of Palestine Ahliya University for Research
[52] Princess Sumaya University for Technology

In contrast to the training setup, the evaluation data is blind, meaning that only reference audio and target text are provided without source transcriptions or ground-truth target synthesized speech. Each system submission must generate speech conditioned on the reference audio while producing the correct target-language content. The resulting evaluation sets contain 588 samples for Arabic, 1,344 for Chinese, and 1,188 for French across all submissions.

4 Evaluation Metrics

The evaluation protocol assesses three key aspects of cross-lingual voice cloning systems: content consistency, speaker similarity, and speech quality. Content consistency evaluates whether the synthesized speech accurately reflects the target text, speaker similarity measures preservation of voice identity, and speech quality assesses the naturalness, intelligibility, and overall acoustic quality of the generated audio.

4.1 Content Consistency

Content consistency is evaluated by first converting the generated speech into text using an Automatic Speech Recognition (ASR) system. We use the faster-whisper[53] (Klein et al., 2020) implementation of the Whisper large-v3 model (Radford et al., 2023),[54] configured with beam search decoding (beam size = 5) and VAD[55] filtering to improve robustness across languages and speaking conditions. The ASR-generated transcripts are then compared against the ground-truth reference texts.

Prior to evaluation, both the hypothesis and reference texts are normalized using Unicode canonicalization (NFKC[56]), lowercasing, and punctuation removal to reduce superficial discrepancies. For Chinese, additional word segmentation is performed using `jieba`[57] to ensure consistent token boundaries. Content consistency is then assessed using Word Error Rate (WER) and Character Error Rate (CER), computed via the Hugging Face `evaluate` library,[58] which quantifies the edit distance between hypothesis and reference transcriptions at the word and character levels, respectively (Morris et al., 2004).

Word Error Rate (WER) measures content consistency by computing the minimum edit distance between the hypothesis and reference texts at the word level. It captures insertions, deletions, and substitutions required for alignment. Due to language differences, tokenization is language-specific, with segmentation applied for Chinese before scoring. Final scores are computed per utterance and averaged across all evaluation samples for each language and system.[59]

Character Error Rate (CER) evaluates transcription accuracy at the character level and provides a more fine-grained measure of textual similarity. This metric is particularly important for languages without explicit word boundaries and for capturing minor transcription variations that WER may not reflect. CER is computed per sample and averaged across all evaluation instances per language and system.[60]

4.2 Speech Similarity

Speech-based evaluation measures similarities between generated and reference speech that are not captured by text-based metrics alone, including speaker identity and prosodic characteristics.

Speaker Similarity is computed using a pretrained ECAPA-TDNN model (Desplanques et al., 2020) implemented in SpeechBrain[61] (Ravanelli et al., 2021). The model extracts speaker embeddings from both reference and generated waveforms and compares them using cosine similarity. In addition to similarity scores, the model also produces binary speaker verification decisions. This metric evaluates whether the generated speech preserves the vocal identity of the reference speaker.

Prosodic Similarity evaluates how closely generated speech matches the reference speech in terms of pitch variation and intensity patterns. Acoustic features are extracted using Praat-Parselmouth[62] (Jadoul et al., 2018), including the

[53] https://github.com/SYSTRAN/faster-whisper
[54] https://github.com/openai/whisper
[55] Voice Activity Detection
[56] Normalization Form Compatibility Composition
[57] https://github.com/fxsjy/jieba
[58] https://github.com/huggingface/evaluate
[59] https://github.com/huggingface/evaluate/tree/main/metrics/wer
[60] https://github.com/huggingface/evaluate/tree/main/metrics/cer
[61] https://speechbrain.github.io/
[62] https://github.com/YannickJadoul/Parselmouth

mean and standard deviation of the fundamental frequency (F0) together with mean energy. The extracted features are aggregated into feature vectors and compared using cosine similarity to measure similarities in rhythm, intonation, and speaking dynamics between generated and reference speech.

5 Submissions

This year, five teams have participated in the cross-lingual voice cloning shared task: *HW-TSC*, *IIT-Patna*, *KIT*, *Langswap*, and *SIT-TCD*. All teams submitted systems for the three target languages, Arabic, Chinese and French, except *HW-TSC*, who submitted system outputs for only Chinese and French. The source language for all submissions is English. Across all submissions, the most common models are Qwen3-TTS[63] and OmniVoice[64] while VoxCPM2,[65] Fish Audio S2 Pro,[66] and Chatterbox[67] are each used by individual teams.

HW-TSC (He et al., 2026) sliced each reference audio into 10-second segments with a 5-second sliding window to account for diverse features in the long reference audio. Then, they generated many versions of the target texts in Chinese and French using Qwen3-TTS 1.7B (Hu et al., 2026). Finally, they computed cosine similarity between timbre feature vectors of three random segments from the original reference audio and each synthesized target audio, and then they selected the one with the highest similarity as the optimal output.

IIT-Patna (Ahtasam et al., 2026) benchmarked four zero-shot TTS models to evaluate their cross-lingual voice cloning capabilities, namely Qwen3-TTS (Hu et al., 2026), CosyVoice3 (Du et al., 2025), VoxCPM2 (Zhou et al., 2025), and MOSS-TTS (Gong et al., 2026). Based on the evaluation results, they decided to use VoxCPM2 for Arabic and Qwen3-TTS 1.7B for Chinese and French.

KIT (Akti and Waibel, 2026) used a multilingual text-to-speech (TTS) model, Fish Audio S2 Pro (Liao et al., 2026), which supports several languages, including Arabic, Chinese and French. At inference time, the authors introduced native-script language tag prompting to improve language control and mitigate accent leakage from the source language to the target. To enhance terminology pronunciation in the target language, they chose reference segments that contain word-level matches with the target text. To this end, they used VibeVoice-ASR (Peng et al., 2026) to generate the transcriptions of the reference English speech and create segmented speech-transcription pairs. Furthermore, they fine-tuned the model with reinforcement learning (GRPO) on the ACL 60/60 dev set to adapt it to the newly introduced language tags and improve the quality of the generated audio.

Langswap (Shigabeev et al., 2026) used OmniVoice (Zhu et al., 2026) for Arabic and Qwen3-TTS 0.6B (Hu et al., 2026) for Chinese and French for zero-shot voice cloning at inference time, without fine-tuning. The cloning process is conditioned on 20-second reference audio clips and their transcriptions via Whisper Large V3 (Radford et al., 2023).

SIT-TCD (Abebe and Moslem, 2026) combined multiple voice cloning models through an ensemble distillation pipeline, and fine-tuned the best-performing model for each target language. For data synthesis, they used three voice cloning models, namely OmniVoice (Zhu et al., 2026), VoxCPM (Zhou et al., 2025), and Chatterbox (Resemble AI, 2025), to synthesize candidate audio for every utterance. Then, they selected the best output using a best-of-N strategy, where each candidate is scored using a combined quality metric: $combined = 0.5 \times (1 - CER) + 0.5 \times SIM$, where CER is measured via automatic transcription and speaker similarity (SIM) is the cosine similarity between the speaker embeddings of the synthesized and reference audio. Finally, they fine-tuned OmniVoice with Low-Rank Adaptation (LoRA) (Hu et al., 2022) with the ensemble-distilled data. For inference, they extracted a 20-second segment of clean speech from each reference audio using energy-based Voice Activity Detection (VAD). They split long target texts into chunks of maximum 200 characters at sentence boundaries to prevent quality degradation. Each chunk is synthesized independently with the fine-tuned model, and then concatenated to form the final output.

[63] https://hf.co/Qwen/Qwen3-TTS-12Hz-1.7B-Base
[64] https://hf.co/k2-fsa/OmniVoice
[65] https://hf.co/openbmb/VoxCPM2
[66] https://hf.co/fishaudio/s2-pro
[67] https://hf.co/ResembleAI/chatterbox

Language	HW-TSC		IIT-Patna		KIT		Langswap		SIT-TCD	
	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓
Arabic	-	-	0.219	0.135	0.157	0.055	0.160	0.063	0.132	0.050
Chinese	0.043	0.047	0.042	0.046	0.113	0.099	0.189	0.171	0.191	0.181
French	0.050	0.010	0.051	0.010	0.063	0.017	0.133	0.052	0.069	0.020

Table 13: Content Consistency with WER and CER

Language	HW-TSC		IIT-Patna		KIT		Langswap		SIT-TCD	
	Speaker ↑	Prosody ↑	Speaker ↑	Prosody ↑	Speaker ↑	Prosody ↑	Speaker ↑	Prosody ↑	Speaker ↑	Prosody ↑
Arabic	-	-	0.669	0.990	0.637	0.982	0.785	0.996	0.786	0.997
Chinese	0.580	0.989	0.499	0.989	0.609	0.981	0.686	0.989	0.789	0.993
French	0.582	0.987	0.479	0.987	0.602	0.980	0.761	0.991	0.813	0.996

Table 14: Speaker Similarity and Prosody Similarity

6 Results

In this first edition of the Cross-Lingual Voice Cloning Shared Task, five systems have been submitted across three target languages, Arabic, Chinese, and French. Three teams operated in a zero-shot setting without fine-tuning (HW-TSC, IIT-Patna, Langswap), while KIT fine-tuned their model using reinforcement learning (GRPO) and SIT-TCD applied LoRA supervised fine-tuning. Table 13 and Table 14 present the content consistency and speaker/prosody similarity scores across all submissions.

Content Consistency. French yields the best (lowest) error scores overall, with most systems achieving WER below 0.07 and CER below 0.02, suggesting it is the most tractable target language for cross-lingual voice cloning among the three languages. Arabic proves the most challenging, resulting in higher WER and CER scores across all four submissions, with SIT-TCD and KIT achieving the best scores. Chinese shows a clear split between systems, with HW-TSC and IIT-Patna achieving the lowest error rates, likely due to their use of Qwen3-TTS, which offers strong Chinese language support.

Speaker and Prosody Similarity. Prosody similarity scores are uniformly high across all systems and languages (0.980–0.997), with little variance between systems. Speaker similarity provides a more informative signal, as it ranges from 0.479 to 0.813 across systems and languages. SIT-TCD achieves the highest speaker similarity scores across all three languages, followed by Langswap as the second-best system overall. By contrast, IIT-Patna and HW-TSC achieve the worst speaker similarity scores in Chinese and French, despite their strong content consistency, pointing to a trade-off between textual accuracy and voice identity preservation.

Overall, the results reveal a clear trade-off between content consistency and speaker similarity, with systems excelling in one dimension often underperforming in the other. Fine-tuning can improve voice cloning performance, as evidenced by KIT and SIT-TCD submissions. Nevertheless, strong zero-shot systems remain competitive, particularly for content consistency. While French is the most accessible target language, Arabic remains the most challenging language across both dimensions.

7 Future Work

Future editions of the shared task could expand the set of target languages to cover a more diverse range of language families and scripts. Additionally, rather than providing pre-translated target text, future runs could require systems to perform translation and voice cloning jointly, bringing the task closer to real-world multilingual content creation scenarios. We would like also to expand evaluation to cover other linguistic and technical aspects such as code-switching and terminology consistency.

Track IX Instruction-Following Track

1 Introduction

Large language models (LLMs) have significantly transformed natural language processing by enabling a single model to perform a broad range of tasks without task-specific training or fine-tuning. Through simple textual instructions, these systems can address diverse applications such as translation, summarization, and question answering within a unified interface (Hendy et al., 2023; Gaido et al., 2024). While originally limited to textual inputs, LLMs are increasingly evolving toward multimodal systems, extending their capabilities to additional modalities such as vision and speech (Li et al., 2024).

In parallel, speech foundation models (SFMs) have emerged as scalable architectures for processing spoken language across a wide range of conditions (Latif et al., 2023). The integration of SFMs with the instruction-following abilities of LLMs (Ouyang et al., 2022) has given rise to a new class of systems, often referred to as SpeechLLMs. This paradigm aims to combine the strengths of both components: the ability to robustly model speech inputs together with the flexible reasoning and generalization capabilities of LLMs, enabling more general-purpose and controllable spoken language interfaces (Rubenstein et al., 2023).

Building on the success of the first edition, this second iteration of the Instruction Following (IF) shared task further explores this emerging paradigm. The goal is to evaluate general-purpose instruction-following models for speech that can handle multiple speech-to-text tasks conditioned on natural language instructions. Systems are tested on both short-form audio segments and long-form spoken content, reflecting realistic usage scenarios and current research trends in multimodal language modeling (Papi et al., 2026b).

2 Task Description

In the Instruction-Following (IF) task, participants had to develop a single instruction-following model that can perform multiple speech-to-text tasks based on natural language prompts. The model receives both a speech input and a task instruction in textual form and is expected to follow the instruction to produce the appropriate output.

Sub-Tracks. The task is divided into two sub-tracks based on the nature of the input: `SHORT`, where the input is represented by automatically segmented speech, and `LONG`, where the input is a long-form speech. Depending on the sub-track, the following tasks have to be supported:

- `SHORT` + `LONG` Common Tasks:
  - **Automatic Speech Recognition (ASR)**: the speech is transcribed into the same language;
  - **Speech-to-text Translation (S2TT)**: the speech is translated into the target language;
  - **Spoken Question Answering (SQA)**: textual questions have to be answered based on the spoken content in the same language and in a language different from the speech (questions and answers are always in the same language);
  - **[Suprisal] Quality Esitmation (QE):** a task that was unknown at submission time but doable through in-context learning abilities of SpeechLLMs; this year, Quality Estimation has been chosen, requiring the model to give select the translation having the best quality among two options (A, B).
- Additional `LONG` Only Tasks:
  - **Speech-to-text Summarization (S2TSUM)**: a summary has to be provided from the spoken content in the same language and in a language different from the speech.
  - **Audio Chaptering (ACHAP)**: the spoken content has to be segmented into coherent sections, each labeled with a concise title summarizing its topic.

All tasks listed for each sub-track were mandatory; that is the model must be capable of handling each task type when prompted appropriately.

Languages. The tasks involve both monolingual and cross-lingual processing. The supported languages are English (en) for ASR, monolingual SQA, S2TSUM, and ACHAP, and English to German (de), Italian (it), and Chinese (zh) for S2TT, multilingual SQA, multilingual S2TSUM, multilingual ACHAP, and surprisal QE. Participants were allowed to submit results for a subset of language directions.

Prompts. For each sample in the test set, there is no information about the specific task to be performed (e.g., ASR) or the language pair to support (e.g., en); rather, the model has to correctly interpret and fulfill diverse instructions across the supported language pairs (e.g., "Traduci questo au-

dio in inglese"[it], "Translate this audio into English"[en]).

3 Data and Metrics

3.1 Training and Development Data

We adopt two evaluation conditions: constrained and unconstrained. In the *constrained* condition, participants are allowed to use the specified Speech Foundation Model, SeamlessM4T v2 large[68], and Large Language Model, Qwen3 4B[69], training their systems[70] on the following datasets:

- EuroParl-ST (Iranzo-Sánchez et al., 2020), CoVoST2 (Wang et al., 2020), and GigaST (Ye et al., 2023) for ASR/S2TT,
- LibriSQA (Zhao et al., 2024) for SQA;
- NUTSHELL (Züfle et al., 2025) for S2TSUM;
- YTSeg (Retkowski and Waibel, 2024) for ACHAP.

Development data are MCIF (Papi et al., 2026b) for all data but ACHAP, and the test split of YTSeg (Retkowski and Waibel, 2024) for ACHAP.

No training data is provided for cross-lingual SQA, S2TSUM, or ACHAP tasks where the output languages differ from the source speech language, which is designed to test the models' zero-shot cross-lingual abilities. The *unconstrained* condition places no limitations on model architectures, pre-trained models, or training data.

The constrained evaluation condition is meant for providing a controlled environment for comparing different approaches without the confounding effects of varying data sources or model scales. On the other hand, the unconstrained condition reflects real-world deployment scenarios where practitioners may leverage cutting-edge models, proprietary datasets, and computational scaling to achieve optimal performance.

3.2 Evaluation Data

We evaluate the submitted models with a novel resource, scraped from the scientific talks from 2023 and 2025 available from the ACL Anthology,[71] with the same process of MCIF (Papi et al., 2026b).

The dataset contains 21 videos, corresponding to 2 hours, covering all tasks but S2TSUM, while the S2TSUM section contains 100 videos, corresponding to 10 hours and a half. Source audio and video content in English (talks of about 6 minutes each on average) are enriched with multilingual annotations and translations to support: *i)* ASR (en→en); *ii)* S2TT (en→de, it, zh), *iii)* S2TSUM (en→en, de, it, zh); *iv)* SQA (en→en, de, it, zh); *v)* ACHAP (en→en, de, it, zh).

For SQA, at least 10 questions were manually created for each video, and, for ACHAP, the titles were manually created based on the video and the transcript/translation content. The SQA task includes unanswerable questions, to which the only correct response is "*Not answerable* or its corresponding translations in the other languages. [72] The audio data are provided as complete audio files in WAV format for the `LONG` sub-track, and as automatically segmented audio (of 15-20 seconds) using SHAS (Tsiamas et al., 2023) for the `SHORT` sub-track.

Additionally, for the surprisal QE, we use a subset of the IWSLT 2025 ACL Talks (Abdulmumin et al., 2025), drawing from the development set of the metrics shared task dataset.[73] The candidates consist of system submissions from last year's instruction-following shared task, paired with their corresponding human scores, which serve as ground truth for determining the better translation.

We release the videos, source audio, and task instructions to participate in the shared task. Also, we provide an example submission for the `LONG` sub-track, which could be used as a 1-shot task demonstration. Participants submit their system outputs and may adjust instructions to suit their models' prompts. The evaluation is conducted via the SPEECHM platform, presented in Section 2.

3.3 Metrics

The evaluation was carried out by computing separate scores for each of the tasks involved, similar to last year. Namely, for ASR, we computed WER using the jiWER library[74] after normalizing the test using the Whisper normalizer[75] (Rad-

[68] hf.co/facebook/seamless-m4t-v2-large

[69] hf.co/meta-llama/Qwen/Qwen3-4B-Instruct-2507

[70] The use of the pre-trained SFM and LLM is not mandatory, and submissions with models trained from scratch on the allowed data are accepted, as are systems using only one of the two pre-trained models.

[71] aclanthology.org

[72] Namely, in Italian "*Non è possibile rispondere*", German "Nicht zu beantworten.", and Chinese 无法回答。.

[73] hf.co/datasets/maikezu/iwslt2026-metrics-shared-train-dev

[74] github.com/jitsi/jiwer

[75] Specifically, we used version `0.0.10`.

ford et al., 2023). For S2TT, we used COMET[76] (Rei et al., 2020) after concatenating all segments belonging to the same talk in the case of the SHORT sub-track and resegmenting the text with mwerSegmenter to pair them with the reference sentences. For SQA and S2TSUM, we computed BERTScore (Zhang* et al., 2020) rescaling the scores with baselines to obtain more interpretable scores in a wider range (typically, in the [0, 1] range).[77] For ACHAP, we follow the protocol of Retkowski et al. (2026) as implemented in the chunkseg package[78] reporting Collar-F1 (±3s) for segmentation and BERTScore (Zhang* et al., 2020) under the Global Concatenation protocol for title quality. Since current multimodal LLMs do not produce reliable timestamps, hypotheses are submitted as Markdown transcripts and boundary timestamps are recovered via CTC forced alignment of the predicted transcript to the source audio. In the cross-lingual sub-track (en→{de, it, zh}), the translated body is first re-aligned to the gold translation with mwerSegmenter and substituted with the source-language reference transcript so that alignment is consistently performed in English. As diagnostics, we additionally report WER (monolingual) and COMET (cross-lingual) on the predicted transcript or translation. Lastly, for the QE surprisal task, we introduce two metrics: QE-format-accuracy, which assesses how many samples (in percentage) follow the format specified in the instruction in the output of a system; QE-accuracy, which is the number of times (in percentage) in which a system correctly identifies the best translation among the two options proposed. Notice that QE-accuracy is computed only on the fraction of samples that follow the correct output format: for this reason, if the QE-format-accuracy is low, the QE-accuracy should be disregarded, as it is computed on a few samples, which makes it unreliable. In addition, since the output options are two, a QE-accuracy of 0.5 is equivalent to a random guess or to a system always predicting the same output.

The code used for the evaluation is available at: github.com/hlt-mt/iwslt2026.

[76] With model Unbabel/wmt22-comet-da.

[77] See github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md

[78] github.com/retkowski/chunkseg

4 Submissions

In total, we received 14 submissions from four different teams. Two teams submitted under the constrained settings, and four submissions were contrastive. Moreover, all teams participated in all language directions. The participants' systems in the SHORT (NLE, FBK, KIT, and BSC) and LONG (KIT, and FBK) sub-tracks are detailed below.

BSC (Pareras et al., 2026) participated in the SHORT unconstrained track, submitting to all language pairs. Their system combines a speech encoder with a translation-tailored LLM backbone, SalamandraTA-7b-instruct-WMT25 (Garcia Gilabert et al., 2025), connected via a linear projection layer. Their primary submission uses SeamlessM4T-v2-Large (Seamless Communication et al., 2023) as speech encoder (frozen during training). A key feature of their approach is a chain-of-thought generation strategy that forces the model to produce an intermediate transcription before generating the final output, so to enable the reuse of text-only supervision. Training data cover ASR, S2TT, T2TT, QA, SQA, and IF. At inference, beam search decoding is used alongside a post-editing stage with Gemma4-31B to correct output formatting errors. They additionally submitted contrastive systems using and mHuBERT-base-25Hz (Hassid et al., 2023) as encoder, either with and without post-editing.

FBK (Xie et al., 2026) participated in both the SHORT and LONG tracks in constrained conditions, submitting to all language pairs. Their system uses the allowed models, SeamlessM4T-v2-large as speech encoder and Qwen3-4B as LLM, and merges them through an MLP with 2 layers. The encoder is kept frozen, while the LLM is finetuned using LoRA to the query, key, and output projection layers of the self-attention modules. Their training data is built by creating synthetic translations when the reference for a target language is missing (filtering samples with low COMET), and generating synthetic question-answer pairs with the LLM. For the LONG track, they create long-form data by concatenating samples in LibriSQA and explore three audio segmentation strategies for chunking long audios to be able to process them with the speech encoder: using fixed windows, leveraging a VAD, and a hybrid approach (Gaido et al., 2021b) combining pause-based segmentation with a duration con-

straint.

NLE (Boito et al., 2026) participated in the SHORT constrained sub-track, submitting to all language pairs. Building on their submission last year (Lee et al., 2025), they train two components in parallel and then merge them: a speech-to-text projector mapping averaged SeamlessM4T-v2-large encoder representations into the frozen Qwen3-4B LLM, and text-only LoRA adapters trained on MT and QA data. The main change from last year is replacing the transformer-based projector with an improved SpeechMapper (Mohapatra et al., 2026), which learns the speech-to-embedding mapping from ASR-only data without LLM forward passes. The two modules are then integrated through a brief multimodal SFT stage. To reduce the domain gap with the scientific-talk evaluation data, they introduce fakACL, a synthetic SQA dataset of generated NLP paper presentations, built by prompting the LLM backbone and synthesizing speech with SeamlessM4T-v2-large TTS.

KIT (Ugan et al., 2026) participated in the SHORT and LONG unconstrained sub-track, submitting to all language pairs. Their primary model is based on Qwen2.5-Omni (Xu et al., 2025a), fine-tuned on task-specific data constructed via a three-stage augmentation framework combining segment concatenation, LLM-based label generation, and cross-lingual reference translation. To improve generation quality, they applied a combined Likelihood and MBR re-ranking strategy over 17 candidates, applied selectively for English and Chinese. They additionally submitted a cascaded contrastive system pairing parakeet-tdt-0.6b-v2 (Sekoyan et al., 2025b) for ASR with Qwen2.5-7B-Instruct (Qwen et al., 2025) for instruction following.

5 Results

5.1 Automatic Evaluation

The complete results for both SHORT and LONG sub-tracks are presented in Tables 38–41. For comparison, we include the results of the Phi4-Multimodal (Abouelenin et al., 2025) and Qwen3-Omni (Xu et al., 2025b). The inference code is available at github.com/hlt-mt/mcif.

ASR. In the ASR SHORT sub-track, the two baselines show a substantial gap, with Phi4-Multimodal clearly outperforming Qwen3-Omni (6.9 vs. 19.7 WER). All participants surpass the Qwen3-Omni baseline but fall short of Phi4-Multimodal: KIT's unconstrained primary comes closest with the lowest participant error rate (7.4), followed by FBK's constrained primary (12.3), while BSC's unconstrained primary and NLE's constrained primary follow closely behind, tying at 13.4 WER. In the LONG sub-track, the Phi4-Multimodal baseline degrades considerably (28.1), while Qwen3-Omni remains far more stable (15.8). The participants' end-to-end primaries follow the Phi4-Multimodal pattern: FBK's constrained primary reaches 19.6 WER and KIT's unconstrained primary 26.9 WER, both improving over Phi4-Multimodal but falling behind the more robust Qwen3-Omni. The degradation from SHORT is substantial, with KIT's WER more than tripling (7.4 → 26.9) and FBK's rising by over 50% (12.3 → 19.6). The exceptions are cascaded models like KIT's unconstrained contrastive1 that obtains the best overall result (6.4 WER).

SQA. Overall, for the SQA task, FBK and NLE's constrained systems perform best in the SHORT track, while KIT's unconstrained primary and Phi4-Multimodal are the strongest systems in the LONG track across all languages. Scores drop consistently from the SHORT to the LONG track, with Qwen3-Omni being the weakest system in the LONG track across all languages. In the SHORT track, FBK's constrained systems lead consistently across German (0.477), Italian (0.527), and Chinese (0.520), with NLE's constrained primary leading for English (0.531) and FBK close behind (0.505–0.507). KIT's unconstrained primary is also competitive across all languages (0.466–0.519), often ranking second or third. In the LONG track, the picture shifts, with KIT's unconstrained primary leading for German (0.405), Italian (0.439), and Chinese (0.452), and Phi4-Multimodal leading for English (0.420). FBK's constrained systems remain competitive in the LONG track, consistently ranking second or third across languages (0.348–0.390), while Qwen3-Omni performs worst across all languages (0.246–0.354).

S2TT. In the S2TT SHORT sub-track, the two baselines differ substantially: Qwen3-Omni is consistently strong (0.836, 0.827, 0.857 for de, it, zh), confirming results of previous work (Papi et al., 2026a), whereas Phi4-Multimodal falls be-

hind (0.802, 0.772, 0.809). Among the participants, KIT's unconstrained primary is the strongest, leading in German (0.840) and Italian (0.841) and trailing only Qwen3-Omni in Chinese (0.852 vs. 0.857). Unlike ASR, it surpasses both baselines in two of the three directions. BSC's unconstrained primary is the next-best participant (0.808, 0.773, 0.782), while the constrained submissions sit lower, with NLE's primary marginally ahead of FBK's (e.g., 0.794 vs. 0.777 in Chinese). These results reflect the difficulty of competitive S2TT without external data. In the LONG sub-track, KIT's cascaded contrastive system obtains the strongest results in all three directions (0.840, 0.841, 0.847 for de, it, zh), matching its short-track scores. This confirms the robustness of cascaded systems to long-form input, as also observed on ASR. In contrast, the end-to-end systems degrade substantially: KIT's primary loses substantially in German and Italian (0.840 → 0.733, 0.841 → 0.732), and FBK's constrained systems cluster around 0.65 to 0.72. The baselines are hit hardest, with Qwen3-Omni collapsing in German (0.836 → 0.408) while staying comparatively stable in Chinese (0.857→0.820), and Phi4-Multimodal dropping to 0.55 to 0.62 overall. Across tracks, translation quality is consistently lower in the LONG than in the SHORT sub-track for every end-to-end system.

SSUM. For the SUM task, KIT's unconstrained systems lead across all languages in the LONG track. While KIT's primary and contrastive systems perform similarly for Italian (0.267–0.269) and Chinese (0.378–0.383), KIT's unconstrained primary leads more clearly for German (0.238 vs. 0.208) and KIT's unconstrained contrastive for English (0.218 vs. 0.212). FBK's constrained systems consistently rank second, clustering closely together across languages (0.149–0.185), and performing similarly to the baselines. Qwen3-Omni is the weakest system across all languages, particularly for Chinese (0.021). Scores are notably higher for Chinese than for the other languages across all systems.

ACHAP. As a LONG-only task, ACHAP received submissions only from KIT and FBK, evaluated alongside the baselines. Phi4-Multimodal fails across all settings (Collar-F1 0.030-0.148). Qwen3-Omni is the leading model in English, obtaining the overall top Collar-F1 (0.609), but struggles cross-lingually: it collapses in German (0.062) and Chinese (0.000), remaining competitive only in Italian (0.446). In contrast, KIT's submissions are outperformed in English (with the contrastive scoring 0.583) but are strong cross-lingually. For KIT's primary, other languages even outperform its own English score (0.436), reaching 0.500 (de), 0.456 (it), and 0.503 (zh). Chinese is the hardest setting overall, where all baselines and submissions are weak except KIT's primary (0.503). All FBK's systems score 0.000 Collar-F1 under strict evaluation, due to a mix of model behaviour and XML formatting issues: space indentation that does not strictly follow the Markdown format, missing newlines between chapters, and occasionally a single chapter spanning the whole document. Under a relaxed evaluation, FBK's systems recover some performance in English, German, and Italian, best in Italian at 0.348, though still behind KIT and Qwen3-Omni. As diagnostics, the WER and COMET show that transcription/translation quality is mainly stable compared to the ASR/S2TT task, with some language-specific characteristics: English and German tend to get slightly worse, Italian slightly better, and Chinese remains stable. The notable exception is KIT contrastive, which had the best S2TT results in Chinese, but its COMET drops catastrophically as part of ACHAP (0.847 to 0.451), which also explains its weak Collar-F1 (0.103). Title quality (BERTScore) broadly tracks segmentation quality, with KIT and Qwen3-Omni leading. The notable exception is Phi4-Multimodal in English, which obtains surprisingly strong titles (0.837) despite near-zero Collar-F1: it appears to output sensible titles but a poor transcript (94.1 WER), so the chapters cannot be grounded in time.

SURPRISAL (QE). Overall, the Qwen3-Omni baseline emerges as the best system across language pairs (en-de/zh) and conditions (SHORT/LONG). This demonstrates the generalization abilities of this model, compared to the systems submitted by the participants that are mainly optimized for the known tasks. This also underscores the importance of surprisal tasks in assessing the ability of the systems to follow instructions. The other baseline system (Phi4-Multimodal) is able to produce properly formatted outputs only for German, failing to do so for Chinese. Even in German, though, its accuracy in performing the task is low (0.658), not far from a

random guess (0.5). Looking at the participants' submissions for German, the only system able to perform the task in the LONG condition is KIT's contrastive, which always outputs the correct format and achieves 70.5% accuracy. KIT's primary, instead, never follows the requested format, and all FBK's systems perform at random guess when respecting the output format (less than 60% of the cases). In the SHORT track, instead, systems perform significantly better. At least one system for all the participants is capable of following the requested output format, with all primary submissions (except for KIT) respecting the format instruction in >95% of the cases. The only system achieving 100% format accuracy (KIT's contrastive1) underperforms in the QE task (70.5%) compared to other submissions. Overall, the best submission is BSC's contrastive1 (81.0%), closely followed by NLE's primary, BSC's primary, and then FBK's primary. Moving to Chinese, the situation is similar, as the participants respecting the format requested in the instruction are the same. In the LONG track, the systems are better than in German: although the best system is always KIT's contrastive1, its score is slightly higher, and FBK's systems respect the output format more frequently and achieve scores better than random choice (65.8%). In the SHORT track, the ranking in terms of the QE ability is instead different, with overall significantly higher scores compared to German. The best system in this case is BSC's primary (92.9%), closely followed by FBK's primary (91.5%) and NLE's primary (89.4%).

5.2 S2TT Human Evaluation

Similar to the other tracks of this year's IWSLT Evaluation Campaign, each participant's primary submission was manually evaluated for the S2TT task. Appendix A reports the manual evaluation process, whereas overall results by language with Direct Assessment (0–100 mean scores) are reported in Tables 18 and 23.

Manual evaluation scores for S2TT broadly reflect those of the automatic assessment. In the SHORT sub-track, KIT leads in both language directions (86.5 en→de, 87.8 en→zh), ranking as the best system after the human reference. NLE and FBK follow closely, with BSC competitive in en→de but weaker in en→zh. Across all systems, en→zh scores are consistently higher than en→de.

Only KIT and FBK participated in the LONG sub-track. KIT again leads (69.7 en→de, 74.3 en→zh), while FBK scores lower (66.1 en→de, 60.0 en→zh), struggling particularly with long-form Chinese output.

Comparing the two sub-tracks, SHORT consistently outperforms LONG for both systems. The degradation is substantial for KIT (−16.3 in en→de, −13.5 in en→zh) and even more pronounced for FBK, especially in Chinese (−9.9 in en→de, −22.3 in en→zh). This suggests that handling long-form audio remains a significant open challenge.

SHORT

Model	*en*	*de*	*it*	*zh*	***avg***
Phi4-Multimodal	2				
Qwen3-Omni		2		2	
NLE_PRIMARY	2	2		1	1
NLE_CONTRASTIVE 🌐	3				
FBK_PRIMARY	1	3	3	2	1
FBK_CONTRASTIVE			3		
KIT_PRIMARY 🌐	2		1		2
KIT_CONTRASTIVE 🌐			2	2	3
BSC_PRIMARY 🌐		1		3	3
BSC_CONTRASTIVE1 🌐					
BSC_CONTRASTIVE2 🌐					

LONG

Model	*en*	*de*	*it*	*zh*	***avg***
Phi4-Multimodal	3	3			3
Qwen3-Omni	2				
FBK_PRIMARY			3		
FBK_CONTRASTIVE1					
FBK_CONTRASTIVE2				3	
KIT_PRIMARY 🌐	2	1	1	1	1
KIT_CONTRASTIVE 🌐	1	2	2	2	2

Table 15: Participant Systems' Ranking by sub-track (SHORT/LONG) and language. 🌐 means unconstrained training settings. Baselines are in *italic*.

6 Discussion and Conclusions

Following last year, ASR emerged as the most accessible task, but, in contrast to results from the first edition, it is now strictly followed by S2TT, with scores up to 0.840–0.857 COMET across languages. Another different trend is seen in SQA that, in contrast to last year where monolingual (English) SQA only was better handled by participants, this year shows comparable performance across languages with the only exception of German (up to 0.477 BERTScore against 0.520–0.531 of other languages), highlighting an increased ma-

turity of multilingual QA abilities.

In trend with last year, instead, we observe a significant degradation when processing long-form inputs, especially in SQA. Long-form only tasks (SSUM and ACHAP), as expected, are the most difficult tasks, with SSUM still being the most challenging across languages and systems.

New this year, we introduced the surprisal task, unknown at submission time, i.e., quality estimation. The results underscore that many systems struggle to generalize to unseen tasks during training/finetuning, and some of them are not even able to follow the correct output structure.

Lastly, this year is harder to establish a clear winner of the shared task, as best results are more scattered across the different submissions, especially for the `SHORT` sub-track. Best results for some tasks are from systems that scored very low results in others. Table 15 shows the top-ranked systems for each sub-track (`SHORT`/`LONG`) by language. On average across tasks and languages (***avg***), FBK's and NLE's primary achieves the best result in the `SHORT` sub-track, strictly followed by KIT's primary, and then KIT's contrastive and BSC's primary submissions. Notably, FBK's primary submission is the only submission always ranking at least third across languages. Moreover, the two top systems are trained in constrained settings, similarly to last year. In the LONG sub-track, instead, the picture is clear, with KIT's primary being the best system, always being first or second (and only in English). This is followed by the contrastive counterpart, similarly ranked always second except for English, where it scores the best result. A baseline, Phi4-Multimodal, ranked third.

All in all, this year's IF evaluation campaign highlights past and new challenges: *i)* long-form processing is still a major limitation of current models, but the gap is closing, especially in ASR and S2TT, *ii)* summarization remains the most challenging task, followed by ACHAP—the new long-form task introduced this year—even if training data are available, *iii)* the surprisal task, quality estimation, pointed out that most current systems stuggle to generalize and, notably, to strictly adhere to the instruction, indicating a big room for improvement in the instruction following abilities of SpeechLLMs.

Track X Speech Translation Metrics Track

1 Introduction

This shared task on Speech Translation Metrics is a dedicated evaluation campaign for Quality Estimation (QE) of speech translation. Speech translation has been at the core of IWSLT for years, and the rapid development of speech technology has further expanded the use of speech translation (ST) applications in daily life. Quality estimation makes it possible to assess translation quality without reference translations, which is essential for practical use cases (Specia et al., 2010; Callison-Burch et al., 2012). However, its evaluation remains underexplored (Han et al., 2024; Züfle et al., 2026).

Evaluation of speech-to-text translation still relies heavily on text-based evaluation approaches, which often overlook speech-specific phenomena (Han et al., 2024; Züfle et al., 2026) and make unrealistic assumptions about access to gold segmentation and error-free source transcriptions during evaluation (Amrhein and Haddow, 2022; Abdulmumin et al., 2025). As a result, these methods can be less reliable for real-world usage, motivating the need for broader investigation into speech translation quality estimation. Progress in this area has also been hindered by the scarcity of speech translation datasets with human quality ratings. While the text translation community benefits from large-scale evaluation resources released through the WMT Metrics shared tasks (Kocmi et al., 2024, 2025), comparable speech-native resources remain limited, with only a handful of datasets released in recent years (Agarwal et al., 2023; Abdulmumin et al., 2025).

In light of this, this shared task focuses on Quality Estimation for speech translation. In the following, we describe the task setup, participating systems, and meta-evaluation results for the IWSLT 2026 Speech Translation Metrics track.

2 Task Description

Given automatically segmented source speech and a candidate translation, the goal is to predict a segment-level quality score without relying on reference translations. The task covers two language pairs: English–German and English–Chinese. To support text-based metrics, an ASR transcript of the source speech is provided alongside the audio. Metrics may also exploit the context of adjacent segments, as the data consists of long-form monologue speech where neighboring segments share topical and discourse context.

3 Data

As described above, participants are provided with source audio, an automatic transcript, a candidate translation, and the context of adjacent segments. All data for this shared task is released publicly.[79]

Training and development data. Training data consists of 33.7k samples with human Direct Assessment (DA) scores: IWSLT 2023 ACL talks (Agarwal et al., 2023), which follow the speech translation setting, and WMT 2024–2025 (Kocmi et al., 2024, 2025), which are text translation annotations paired with the original YouTube audio. Additionally, 7k samples with synthetic DA scores from the SpeechQE CommonVoice dataset (Han et al., 2024) are included. The development set consists of 5.6k samples from IWSLT 2025 (Abdulmumin et al., 2025). Training data is skewed towards English–German and English–Chinese, as these are the language pairs covered by the test set; see Table 16 for a full breakdown.

Test data. The test set is built from submissions to the IWSLT 2026 offline and instruction-following (IF) tracks, using their translated outputs as the candidate translations to be scored. It covers five domains and two language pairs (English–German and English–Chinese): ACL conference talks (64.5%), TVSeries (12.5%), CallCenter (11.3%), YouTube (8.1%), and Business News (3.7%).

For the ACL domain, submissions from 7 systems (1 offline, 4 IF short, 2 IF long) are included alongside a human reference translation; for all other domains, only the offline system is available, as the IF track exclusively covers ACL. In total, the dataset contains 48,044 examples based on 5,252 unique audio segments (10,470 unique segment–language pairs). For each segment, we provide the source speech and the source transcription produced by Whisper large-v3 (Radford et al., 2023); for the ACL domain, we additionally provide a human source transcript. Candidate translations are aligned to the reference segmen-

[79] huggingface.co/datasets/maikezu/iwslt2026-metrics-shared-train-dev

Split	Data	Lang. pairs (count)
TRAIN	IWSLT23 *(ACL)*	**De (4160)**, Ja (3328), **Zh (4992)**
	WMT24 *(YouTube)*	Cs (1757), Es (1554), Hi (1221), Is (1221), Ja (1443), Ru (1554), Uk (1221), **Zh (1443)**
	WMT25 *(YouTube)*	Cs→De (882), Cs→Uk (874), Ar (684), Bho (665), Cs (660), Et (684), Is (684), It (666), Ja (684), Mas (646), Ru (684), Sr (646), Uk (684), **Zh (684)**
	SpeechQE *(CommonV.)*	**De (3500)**, Es→En (3500)
DEV	IWSLT25 *(ACL)*	**De (3635)**, **Zh (1921)**
TEST	IWSLT26 *(Multi-domain)*	**De (24,016)**, **Zh (24,028)**

Table 16: Overview of data splits and language pairs. All pairs are En→X unless noted otherwise. Bold: language pairs included in the test set. Test spans five domains: *ACL* (64.5%), *TVSeries* (12.5%), *CallCenter* (11.3%), *YouTube* (8.1%), *Business News* (3.7%).

tation using the mwer-segmenter. Details for segmentation can be found in Section A. Human quality judgements were collected by evaluating translations against the source speech (without transcripts), with 4 annotators per language pair; see Section A for details.

4 Meta-Evaluation of Metrics

Similarly to the WMT Metrics Shared Task (Lavie et al., 2025), we meta-evaluate automated how automated metrics correlate with human ratings at the segment and system levels. The evaluation code is publicly available.[80]

Segment-level correlation. For each item, we compute the ranking of the models based on the metric's predictions $\{\hat{y}\}_{i\in\mathcal{M}}$ and human-annotated ground truth $\{y\}_{i\in\mathcal{M}}$. These rankings are compared using Kendall's τ_b (Kendall, 1945), a correlation similar to Spearman correlation and averaged:

$$\tau_b = \frac{\sum_{i<j} \operatorname{sgn}(\hat{y}_i - \hat{y}_j)\operatorname{sgn}(y_i - y_j)}{\sqrt{\sum_{i<j}\operatorname{sgn}^2(\hat{y}_i - \hat{y}_j)\sum_{i<j}\operatorname{sgn}^2(y_i - y_j)}} \quad (1)$$

We use y for ground truth and $\hat{y}$ for metric predictions. Item indices are omitted. This meta-evaluation measures the practical ability of automated metrics to choose the best translation given an item.

System-level correlation. To evaluate the metric's ability to rank systems on whole datasets, we turn to Soft Pairwise Accuracy (Thompson et al., 2024), which compares the certainties of automatic metrics and human annotators in ranking one system higher than another one:

$$\text{SPA} = 1 - \frac{\sum_{j,i\in\binom{\mathcal{M}}{2}} |p(\mu_i > \mu_j) - p(\hat{\mu}_i > \hat{\mu}_j)|}{\binom{|\mathcal{M}|}{2}} \quad (2)$$

We use μ_i to denote the mean of model $i \in \mathcal{M}$. The $p(\mu_i > \mu_j)$ is a paired permutation test.

5 Submissions

We received submissions from 5 teams totalling 14 systems. In addition, we evaluate 6 state-of-the-art QE models as organizer baselines.

5.1 Baselines

COMETKiwi22 (Rei et al., 2022) is one of the most popular referenceless automated metric for textual machine translation evaluation. We use it for speech evaluation based on automatic speech recognition of the source. Both the source and the translations are encoded using a textual encoder Roberta-XLM (Conneau et al., 2020) into a shared vector, based on which a multi-layer perceptron produces human judgments of translation quality.

COMET Partial (Zouhar et al., 2026) follows the same training pipeline as COMETKiwi22 but match the distribution of automatically aligned translations by training on prefixes and suffixes of the source and target texts.

SpeechQE Han et al. (2024) formulates the task of quality estimation for direct speech translation and compares cascaded and end-to-end approaches for estimating translation quality directly from speech inputs. Their end-to-end approach is used as a baseline system for this shared task. It consists of a frozen Whisper-large-v2 speech encoder, a lightweight convolutional modality adapter, and a TowerInstruct-7B (Alves et al., 2024) text LLM fine-tuned with LoRA (Hu et al., 2022). Training follows a two-stage strategy. First, the modality adapter is pretrained on ASR and speech translation tasks to better align

[80] github.com/zouharvi/iwslt26-metrics

speech and text representations. Then, the LLM is fine-tuned on the SpeechQE task using xCOMET-XL (Guerreiro et al., 2024) pseudo-labels with varying quality levels. Notably, the model is trained only on English–German data, and the English–Chinese setting is evaluated in a zero-shot setting.

BLASER 2.0 (Seamless Communication et al., 2023) is designed for multimodal evaluation across both speech and text modalities. The quality estimation version of BLASER 2.0 is a reference-free quality estimation framework that assesses translation quality by directly comparing the source input and the generated hypothesis, without relying on human reference translations. The framework leverages SONAR (Duquenne et al., 2023, Sentence-level mOdal-agnostic representAtions), which provides a shared embedding space across languages and modalities for both speech and text. Technically, BLASER 2.0 QE computes quality scores using multimodal SONAR embeddings, enabling direct comparison between source speech and target text representations. In this task, the system takes SONAR embeddings of the source speech and target text as input features for quality estimation.

SpeechCOMET (Züfle et al., 2026) extends the COMETKiwi22 (Rei et al., 2022) architecture with a SONAR speech encoder (Duquenne et al., 2023), enabling quality estimation directly from source speech. The resulting speech and text representations are fused and passed to a multilayer perceptron trained to predict human quality judgments. The model is trained on a combination of IWSLT 2026 shared task training data and WMT human evaluation data extended with TTS-synthesised source speech. We evaluate the speech+text variant, which conditions on both the source audio and source text.

SpeechLLM FT (Züfle et al., 2026) fine-tunes Qwen2.5-Omni-7B (Xu et al., 2025a), a state-of-the-art multimodal large language model, for quality estimation of speech translation. The model is prompted with the source audio, transcript and hypothesis translation and is instructed to output a single scalar quality score between 0 and 1. Fine-tuning is performed on the provided IWSLT 2026 shared task metrics training data.

5.2 Submissions

Zarzu and Zouhar (2026) investigate whether incorporating source audio improves automatic metrics for speech translation quality estimation through two standard metric paradigms. Speech-LLM uses few-shot prompting on an open-source multimodal LLM (Phi-4-multimodal-instruct, Abouelenin et al., 2025), while COMET+audio extends CometKiwi with an audio modality, fusing InfoXLM (Chi et al., 2021) text representations with Whisper (Radford et al., 2023) audio embeddings. They evaluate both methods across input configurations that isolate the contribution of each modality. Surprisingly, they find that incorporating audio yields no reliable gains over text-only baselines, which they attribute to audio–transcript misalignments and the technical, low-prosody-occurrence nature of the evaluation data.

Krahn and Fosler-Lussier (2026) proposes HydraQE, a quality estimation system for speech translation built on a Qwen3-ASR-1.7B (Team, 2026) backbone. The model accepts source audio and a text translation hypothesis as joint input, extracts layer-mixed representations via a learnable sparsemax scalar mix, and re-encodes them with a bidirectional transformer to enable full cross-modal interaction. Three independent prediction heads are each trained on a distinct supervision signal: human direct assessment annotations, MetricX-24-XXL pseudo-labels (Juraska et al., 2024), and xCOMET-XXL pseudo-labels (Guerreiro et al., 2024). Training follows a curriculum that begins on synthetic and silver data and gradually shifts weight toward human-annotated examples. Their primary submission is a weighted ensemble of the DA and MetricX heads (choosing best checkpoint for each), balancing segment-level with system-level scores. Contrastive submissions include each individual head in isolation and an equal-weighted average of all three heads.

Dinh and Niehues (2025) propose Boosted-Prob, an unsupervised quality estimation method that modifies a model's softmax output distribution at each decoding step, boosting the probability mass of high-confidence tokens to better align resulting scores with actual translation quality. Whisper Large V3 (Radford et al., 2023) is used to force-decode the provided translations, obtaining per-token softmax distributions conditioned on

	Segment	System	Segment EnDe	Segment EnZh	System EnDe	System EnZh
Zarzu and Zouhar (2026)						
MLP (speech)	14.1	80.2	12.6	15.7	82.4	78.0
MLP (text)	15.7	80.0	15.4	16.0	81.5	78.5
MLP (text+speech)	15.8	81.0	15.2	16.5	81.5	80.5
Phi4 (speech)	15.5	91.3	13.5	17.6	87.8	94.8
Phi4 (text)	23.2	96.4	24.3	22.1	97.1	95.7
Phi4 (text+speech) ◂	17.6	96.3	19.3	15.8	94.2	98.4
Dinh and Niehues (2025)						
BoostedProb (text) ◂	14.3	43.3	17.8	10.8	59.2	27.4
Gupta (2026)						
Lexilogic (text) ◂	31.4	96.8	31.4	31.4	94.7	98.9
Shah et al. (2026)						
TieCal (text) ◂	33.0	94.1	33.7	32.3	95.9	92.4
Krahn and Fosler-Lussier (2026)						
HydraQE (DA head) (speech)	34.7	97.2	36.1	33.3	98.5	95.9
HydraQE (MetricX head) (speech)	33.7	97.2	34.9	32.5	98.7	95.6
HydraQE (XComet head) (speech)	34.6	95.8	36.7	32.5	98.0	93.6
HydraQE (all heads avg) (speech)	34.6	96.9	36.1	33.0	98.7	95.2
HydraQE (primary) ◂ (speech)	34.5	97.3	35.8	33.2	98.6	96.1
Organizers						
BLASER (speech)	22.6	85.9	25.9	19.4	83.9	87.8
COMETkiwi (text)	32.9	94.2	33.0	32.7	95.9	92.5
COMETpartial (text)	14.7	81.4	15.8	13.5	92.8	70.0
SpeechCOMET (text+audio)	30.0	91.6	32.7	27.4	88.5	94.7
SpeechLLM FT (text+audio)	25.6	95.1	23.3	28.0	96.8	93.5
SpeechQE (speech)	26.1	91.3	29.6	22.5	91.6	91.1
Annotators						
Random human annotator	46.2	98.8	45.8	46.6	98.4	99.1

Table 17: Main results for IWSLT 2026 Speech Translation Metrics Track on segment- and system-level (all values $\times 100$). The ◂ denotes primary submission from a team. Segment scores are Kendall's τ_b correlation (range -100 to 100), system scores are Soft Pairwise Accuracy (range 0 to 100).

the source audio and translation prefix. BoostedProb is applied on top of these distributions, and the sentence-level quality score is computed as the log mean of the per-token scores. The method requires no QE training data, no task-specific fine-tuning, and only a single forward pass through Whisper.

Gupta (2026) fine-tune COMETKiwi-22 (Rei et al., 2022) on the provided transcripts with a pairwise ranking objective that directly optimizes within-document Kendall τ_b. Training pairs are constructed from translations of the same source document with a score difference exceeding 1 point, using a combined loss of adaptive margin ranking and MSE for calibration, where the margin scales with the gold score gap. Training follows a two-phase schedule in which the XLM-RoBERTa-Large (Conneau et al., 2020) encoder is frozen during warmup and then unfrozen with a lower learning rate.

Shah et al. (2026) propose tie calibration as a simple post-processing of the COMETKiwi-22 model, which is text-based. It maps continuous output scores into discrete bins to obtain ties on very similar inputs, because Kendall's τ_b, the primary meta-evaluation objective in this task, favors ties over wrong rankings. The mapping is controlled by threshold that maximizes the mean per-document Kendall's τ_b on the training data.

6 Results

General Trends. Table 17 reports the primary meta-evaluation results for all submissions based on averaged human scores. System-level scores are consistently high across all metrics; however, this is partly because for 35.5% of the data (all non-ACL domains), only two systems are evaluated, the offline submission and the reference, making pairwise ranking trivially easy. Even for the ACL domain where 8 systems are evaluated, top metrics achieve SPA scores of 93–97, only slightly below the perfect 100 seen on two-system domains for the top metrics. In contrast, segment-level evaluation is low compared to the agreement between human annotators.

The best performing metrics at the segment level are HydraQE (speech-based, 34.5), TieCal

(text-based, 33.0), and Lexilogic (text-based, 31.4), compared to a human annotator agreement of 45.8 for English–German and 46.6 for English–Chinese, highlighting the large remaining gap between automatic metrics and human judgment. The two language pairs show broadly comparable performance, with a slight English–German advantage for several metrics: for the top system, HydraQE scores 35.8 on English–German versus 33.2 on English–Chinese. TieCal (33.7 vs. 32.3) and BLASER (25.9 vs. 19.4) show a similar pattern, while Lexilogic performs identically on both (31.4).

Automatic Transcript vs. Human Transcript as Source. For the ACL domain, where human source transcripts are available, comparing results on human transcripts versus Whisper-generated automatic transcripts reveals surprisingly little difference in segment-level performance across text-based metrics (Table 42 in the appendix). For instance, Phi4 (text) scores 34.8 on human transcripts versus 34.2 on Whisper transcripts, and top-performing systems such as TieCal (47.7 vs. 46.5) and Lexilogic (44.3 vs. 43.0) show similarly small gaps. In some cases, metrics even perform slightly better on Whisper transcripts, e.g. COMETpartial (24.1 vs. 24.2).

Domain-Level Results. Domain-specific results are reported in Tables 43 to 45 in the appendix. For comparability across domains, these tables include only the offline system and the reference translation, making system-level comparison trivial, most metrics achieve SPA of 100. A notable exception is the CallCenter domain, where text-based metrics such as TieCal (82.7) and COMETkiwi (84.0) fail to consistently rank the reference above the offline system. Segment-level performance varies considerably across domains: metrics score highest on TVSeries (e.g. COMETkiwi 39.1, TieCal 38.0, human agreement 47.7) and lowest on CallCenter (e.g. TieCal 20.9, COMETkiwi 20.7, human agreement 42.4), with ACL in between (e.g. TieCal 28.2, COMETkiwi 27.1, human agreement 44.6).

Text vs. Speech Models. The overall top-performing submitted system, HydraQE (Krahn and Fosler-Lussier, 2026), is speech-based, though the margin over the best text-based system is small (34.5 vs. 33.0 for TieCal). Text-based models generally perform competitively with or better than speech-based and multimodal counterparts at the segment level, at least for the high-resource language pairs that we evaluate. This is confirmed by the direct ablation of Zarzu and Zouhar (2026), whose text-only Phi4 model outperforms the speech-only and text+speech variants (23.2 vs. 15.5 and 17.6), a gap that is preserved even when using ASR transcripts (at most 1.3 points difference between human and Whisper transcripts across metrics, see Table 42). Similarly, the text-only COMETkiwi baseline outperforms audio-based organizer baselines such as BLASER (22.6) and SpeechQE (26.1).

7 Discussion and Conclusion

Overall, the results from 5 participating teams with 14 submitted systems show that current automatic metrics for speech translation quality estimation still fall considerably short of human judgment at the segment level, despite performing well at ranking systems. Text-based and speech-based metrics perform surprisingly similarly, suggesting that the test data does not expose speech-specific phenomena, such as prosody, speaking style, or speaker gender, that would require audio input to assess translation quality (Züfle et al., 2026). Similarly, using ASR transcripts instead of human transcripts has little impact on metric performance, indicating that ASR quality is not a bottleneck for current approaches on English source speech, for the relatively high-resource language pairs tested here. Nevertheless, submitted metrics advance the state of the art over existing baselines, demonstrating that the shared task has successfully stimulated progress in this underexplored area.

Acknowledgments

Atul Kr. Ojha and John P. McCrae would like to thank Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, 13/RC/2106_P2 ADAPT SFI Research Centre, and thank RTÉ/TG4 for sharing the Irish speech data.
This work was supported by Czech Operational Program OP JAK, the MSCA CZ project MSCA Fellowships – UK 4, CZ.02.01.01/00/22_010/0013392, "LCT". This work has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People).

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Amanuel Gizachew Abebe and Yasmin Moslem. 2026. One Voice, Many Tongues: Cross-Lingual Voice Cloning for Scientific Speech. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

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*.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, and 43 others. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

Mo Ahtasam, Jamal Uddin, and Mohammad Nadeem. 2026. Balancing Linguistic Intelligibility and Speaker Identity in Zero-Shot Cross-Lingual Voice Cloning. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Seymanur Akti and Alexander Waibel. 2026. KIT's Submission to Cross-Lingual Voice Cloning in IWSLT 2026. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (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, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual

large language model for translation-related tasks. In *First Conference on Language Modeling*.

Chantal Amrhein and Barry Haddow. 2022. Don't discard fixed-window audio segmentation in speech-to-text translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 203–219, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Lrec*.

Hicham Badri and Appu Shaji. 2023. Half-quadratic quantization of large machine learning models.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Carlos Bentes and Christian Safka. 2026. Pinch-AST: Robust Cascaded Speech Translation System for the IWSLT 2026 Simultaneous Speech Translation Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

BINBINLIU, Wenhan Han, Feng Chen, Yifan Zhang, Ping Guo, Haobin Lin, Bingni Zhang, Taifeng Wang, and Yin Zheng. 2026. Token alignment heads: Unveiling attention's role in LLM multilingual translation. In *The Fourteenth International Conference on Learning Representations*.

Marcely Zanon Boito, Hemant Yadav, Jean-Luc Meunier, and Ioan Calapodescu. 2026. NAVER LABS Europe Submission to the Instruction-following 2026 Short Track. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Mitzuko Davis Quispe Callañaupa, Max Erixon Toledo Bernal, Ronil Nilo Torres Bautista, and Patrick Michael Pumaccahua Huallpa. 2026. Optimization of Voice Translation Systems for Indigenous Languages: Retraining the NLLB-200 Model for the Quechua–Spanish Pair. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Mauro Cettolo, Roldano Cattoni, Matteo Negri, and Luisa Bentivogli. 2026a. The FBK Sentence-Aware Subtitling System at the IWSLT 2026 Subtitling Track. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Mauro Cettolo, Roldano Cattoni, Matteo Negri, and Luisa Bentivogli. 2026b. The FBK Sentence-Aware Subtitling System at the IWSLT 2026 Subtitling Track. In *Proc. of IWSLT*, San Diego, US-CA.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, K. Sudoh, K. Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT 2017)*, pages 2–14, Tokyo, Japan.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2022. Blaser: A text-free speech-to-speech translation evaluation metric. *arXiv preprint*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. Ieee.

David Dale and Marta R. Costa-jussà. 2024. BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085, Miami, Florida, USA. Association for Computational Linguistics.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tu Anh Dinh and Jan Niehues. 2025. Are generative models underconfident? better quality estimation with boosted model probability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3364–3382, Suzhou, China. Association for Computational Linguistics.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, and 3 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022a. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022b. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the first workshop on scaling up multilingual evaluation*, pages 21–24.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Stroudsburg, PA, USA. Association for Computational Linguistics.

Quentin Fuxa and Dominik Macháček. 2026. AlignAtt4LLM: Fast AlignAtt for Decoder-Only LLMs at IWSLT 2026 Simultaneous Speech Translation Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021a. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 55–62, Trento, Italy. Association for Computational Linguistics.

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021b. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. *Preprint*, arXiv:2104.11710.

Marco Gaido, Sara Papi, Mauro Cettolo, Matteo Negri, and Luisa Bentivogli. 2025. Simulstream: Open-source toolkit for evaluation and demonstration of streaming speech-to-text translation systems. *Preprint*, arXiv:2512.17648.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech translation with speech foundation models and large language models: What is there and what is missing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14778, Bangkok, Thailand. Association for Computational Linguistics.

Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt, Carlos Escolano, and Maite Melero. 2025. From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 614–637, Suzhou, China. Association for Computational Linguistics.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, Yiyang Zhang, Yang Gao, Hanfu Chen, Ke Chen, Songlin Wang, Xiaogui Yang, Yuqian Zhang, Kexin Huang, ZhengYuan Lin, and 7 others. 2026. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Lilit Grigoryan, Vladimir Bataev, Andrei Andrusenko, Oleksii Hrinchuk, Davit Karamyan, Enas Albasiri, Vitaly Lavrukhin, Nikolay Karpov, and Boris Ginsburg. 2026. NeMo@IWSLT 2026: Cascaded System for Simultaneous Speech Translation. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Mahendra Gupta, Maitreyee Dutta, and Chandresh Kumar Maurya. 2025. Quantitative analysis of error propagation in hindi-english cascaded speech-to-speech translation models. In *2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, pages 751–756.

Pranav Gupta. 2026. Lexilogic@IWSLT 2026: Pairwise Ranking Fine-tuning of CometKiwi for Speech Translation Quality Estimation. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the quality of direct speech translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21852–21867, Miami, Florida, USA. Association for Computational Linguistics.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, and 1 others. 2023. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36:63483–63501.

Barathi Ganesh HB, Michal Ptaszynski, Jairam R, and Reshma Unnikrishnan. 2026. AURA-ST: Acoustic-Unconstrained Residual Architecture for Speech Translation. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Yu He, Daimeng Wei, Jiaxin Guo, Yuanchang Luo, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Boqi Huang, and Xiaoqing Lan. 2026. HW-TSC's Submission to the IWSLT 2026 Cross-Lingual Voice Cloning Track. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*.

Boqi Huang, Daimeng Wei, Jiaxin Guo, Yuanchang Luo, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Yu He, and Xiaoqing Lan. 2026a. HW-TSC's submission to the IWSLT 2026 Offline Speech Translation track. In *Proc. of IWSLT*, San Diego, US-CA.

Boqi Huang, Daimeng Wei, Jiaxin Guo, Yuanchang Luo, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Yu He, and Xiaoqing Lan. 2026b. HW-TSC's Submissions to the IWSLT 2026 Offline Speech Translation Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Jorge Iranzo-Sánchez, Javier Iranzo-Sanchez, Adrià Giménez Pastor, Jorge Civera Saiz, and Alfons Juan. 2025. MLLP-VRAIN UPV system for the IWSLT 2025 simultaneous speech translation translation task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 340–346, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Jorge Iranzo-Sánchez, Gerard Mas-Mollà, Adrià Gimenez, Jorge Civera Saiz, Albert Sanchis, and Alfons Juan. 2026. MLLP-VRAIN UPV System for the IWSLT 2026 Simultaneous Speech Translation Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

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.

Maurice George Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

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.

Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep Crego, and Jean Senellart. 2020. Efficient and high-quality neural machine translation with OpenNMT. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 211–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Kevin Krahn and Eric Fosler-Lussier. 2026. HydraQE: OSU's Submission for the IWSLT 2026 Speech Translation Metrics Shared Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules. *Preprint*, arXiv:1909.09577.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Xiaoqing Lan, Daimeng Wei, Jiaxin Guo, Yuanchang Luo, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Boqi Huang, and Yu He. 2026a. HW-TSC's Submission to the IWSLT 2026 Subtitling Track. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Xiaoqing Lan, Daimeng Wei, Jiaxin Guo, Yuanchang Luo, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Boqi Huang, and Yu He. 2026b. HW-TSC's submission to the IWSLT 2026 Subtitling track. In *Proc. of IWSLT*, San Diego, US-CA.

Phillip A Laplante. 1992. *Real-time systems design and analysis: an engineer's handbook*. IEEE press.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of Large Audio Models: A Survey and Outlook. *arXiv preprint arXiv:2308.12792*.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Beomseok Lee, Marcely Zanon Boito, Laurent Besacier, and Ioan Calapodescu. 2025. NAVER LABS Europe submission to the instruction-following track. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 186–200, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2024. Multimodal foundation models: From specialists to general-purpose assistants. *Found. Trends. Comput. Graph. Vis.*, 16(1–2):1–214.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12979–12998, Suzhou, China. Association for Computational Linguistics.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, Zhizhuo Zhou, Jiahua Liu, Xin Chen, and Dawei Han. 2026. Fish audio s2 technical report. *arXiv preprint arXiv:2603.08823*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. 2025. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Comp. and Comm.*, 28(4):12–17.

Danni Liu, Sai Koneru, and Jan Niehues. 2026. DietKIT: Post-Training Quantization for Speech LLMs. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In *Interspeech 2020*, pages 3620–3624.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Diego Alberto Barriga Martínez, Amilkar Gazque, Mikel Segura Elizalde, Carlos Daniel Hernandez Mena, Ximena Gutierrez-Vasques, and Ivan Vladimir Meza Ruiz. 2026. Mapudungun-Spanish Speech Translation: A Low-Resource End-to-End System for the IWSLT 2026 Shared Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, pages 138–144.

Mohammad Mohammadamini and Marie Tahon. 2026. LIUM Submission for IWSLT 2026 Low-resource Speech Translation Track. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Mohammad Mohammadamini, Dilgash Mohammed Salih Tayib, Dezheen H. Abdulazeez, Barzan Hussein Mohammed, Imad Saeed Sadeeq, Aveen Jalal Mohammed, Amera Ismail Melhum, and Abuobaida Abdullah Dheyab. 2026. Fleurs-Badini: Translation and Recording Fleurs Dataset for Badini Variant of Northern Kurdish. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Biswesh Mohapatra, Marcely Zanon Boito, and Ioan Calapodescu. 2026. Speechmapper: Speech-to-text embedding projector for llms. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16277–16281.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768.

Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 265–273, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Wataru Nakata, Yuki Saito, Yota Ueda, and Hiroshi Saruwatari. 2025. Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing. *arXiv preprint arXiv:2509.17052*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint*.

Atul Kr. Ojha. 2019. English-Bhojpuri SMT System: Insights from the Kāraka Model. *arXiv preprint arXiv:1905.02239*.

Atul Kr. Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37, Suzhou, China. Association for Computational Linguistics.

Atul Kr. Ojha and Daniel Zeman. 2020. Universal Dependency treebanks for low-resource Indian languages: The case of Bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38, Marseille, France. European Language Resources Association (ELRA).

Aziz Sharipov Ortega and Dominik Macháček. 2026. A Pocket Offline Model for Simultaneous Speech Translation as CUNI Submission to IWSLT 2026. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

John E. Ortega, Rodolfo Joel Zevallos, Fabrício Carraro, Stephanny Gabriela Sánchez Bautista, and Chad Howe. 2026a. Team QUESPA System Submission for the IWSLT 2026 Dialectal and Low-resource Speech Translation Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

John E. Ortega, Rodolfo Joel Zevallos, Fabrício Carraro, Stephanny Sánchez, and Lewis C. Howe. 2026b. Team quespa system submission for the iwslt 2026 dialectal and low-resource speech translation task. In *Proceedings of the 23nd International Conference on Spoken Language Translation (IWSLT 2026)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Siqi Ouyang, Xi Xu, and Lei Li. 2025. InfiniSST: Simultaneous translation of unbounded speech with large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3032–3046, Vienna, Austria. Association for Computational Linguistics.

Rosa YG Paccotacya-Yanque, Candy A Huanca-Anquise, Judith Escalante-Calcina, Wilber R Ramos-Lovón, and Álvaro E Cuno-Parari. 2022. A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.

Alonso Palomino. 2026. Selected-Layer Codec Compression for Compact Speech Translation Models: An IWSLT 2026 English-to-Chinese Submission. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. StreamAtt: Direct streaming speech-to-text translation with attention-based audio history selection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3692–3707, Bangkok, Thailand. Association for Computational Linguistics.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Does simultaneous speech translation need simultaneous models? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sara Papi, Javier Garcia Gilabert, Zachary Hopton, Vilém Zouhar, Carlos Escolano, Gerard I. Gállego, Jorge Iranzo-Sánchez, Ahrii Kim, Dominik Macháček, Patricia Schmidtova, and Maike Züfle. 2026a. Hearing to translate: The effectiveness of speech modality integration into llms. *Preprint*, arXiv:2512.16378.

Sara Papi, Peter Polák, Dominik Macháček, and Ondřej Bojar. 2025. How "real" is your real-time simultaneous speech-to-text translation system? *Transactions of the Association for Computational Linguistics*, 13:281–313.

Sara Papi, Marco Turchi, Matteo Negri, and 1 others. 2023. AlignAtt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. In *Proceedings of Interspeech 2023*. Isca.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2026b. MCIF: Multimodal crosslingual instruction-following benchmark from scientific talks. In *The Fourteenth International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics.

Oriol Pareras, Joan Llado, Pol Buitrago, Marc Casals-Salvador, Federico Costa, and Cristina Espana-Bonet. 2026. BSC's Submission to the Instruction Following Track of IWSLT 2026. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, Yutao Sun, Hangbo Bao, Weijiang Xu, Yi Zhu, Zehua Wang, Ting Song, Yan Xia, Zewen Chi, Shaohan Huang, and 5 others. 2026. Vibevoice-asr technical report. *arXiv preprint arXiv:2601.18184*.

Frithjof Petrick, Patrick Wilken, Evgeny Matusov, Nahuel Unai Roselló Beneitez, and Sarah Beranek. 2025. AppTek's Automatic Speech Translation: Generating Accurate and Well-Readable Subtitles. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.

Peter Polák, Sara Papi, Luisa Bentivogli, and Ondřej Bojar. 2025. Better late than never: Meta-evaluation of latency metrics for simultaneous speech-to-text translation. *arXiv preprint arXiv:2509.17349*.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. CUNI-KIT system for simultaneous speech translation task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Benjamin Pong. 2026. Towards Dynamic Attention Masking for Simultaneous Speech Translation. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624.

Maja Popović. 2015a. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Maja Popović. 2015b. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and Hieu Hoang. 2025. Effects of Automatic Alignment on Speech Translation Metrics. In *Proc. of IWSLT*, Vienna, Austria.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

John W. Ratcliff and John A. Obershelp. 1841. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–51.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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. 2022. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Resemble AI. 2025. Chatterbox-TTS. `https://github.com/resemble-ai/chatterbox`. GitHub repository.

Fabian Retkowski and Alexander Waibel. 2024. From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 406–419, St. Julian's, Malta. Association for Computational Linguistics.

Fabian Retkowski, Maike Züfle, Thai Binh Nguyen, Jan Niehues, and Alexander Waibel. 2026. Beyond transcripts: A renewed perspective on audio chaptering. *Preprint*, arXiv:2602.08979.

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, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

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. Association for Computational Linguistics.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. Ieee.

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.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025a. Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025b. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sukanta Sen, Ondřej Bojar, and Barry Haddow. 2022. Simultaneous translation for unsegmented input: A sliding window approach. *Preprint*, arXiv:2210.09754.

Nivedita Sethiya, Puneet Walia, and Chandresh Kumar Maurya. 2025. Indic-S2ST: a multilingual and multimodal many-to-many Indic speech-to-speech translation dataset. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3766–3775, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Mubashir Hussain Shah, Aymen Fatima, Kiho Choi, and Daehee Jang. 2026. Tie-Calibrated COMETKiwi for Speech Translation Quality Estimation: IWSLT2026 Metrics Track. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-asr technical report. *Preprint*, arXiv:2601.21337.

Ilya Shigabeev, Ilia Latyshev, Nikolay Pakhtusov, and Milana Shkhanukova. 2026. LangSwap: Dub any

video into another language. *GitHub*. Accessed: 2026-04-15.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Claytone Sikasote, Eunice Mukonde, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023a. BIG-C: a multimodal multi-purpose dataset for Bemba. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2062–2078, Toronto, Canada. Association for Computational Linguistics.

Claytone Sikasote, Kalinda Siaminwe, Stanly Mwape, Bangiwe Zulu, Mofya Phiri, Martin Phiri, David Zulu, Mayumbo Nyirenda, and Antonios Anastasopoulos. 2023b. Zambezi Voice: A Multilingual Speech Corpus for Zambian Languages. In *Proc. INTERSPEECH 2023*, pages 3984–3988.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier.

Kaustuk Pratap Singh, Dipanshu ., Vedant Singh, and Kumar Rishu. 2026. IIIT-BGP IWSLT 2026 Systems for Low-resource ST. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Supriti Sinhamahapatra, Thai-Binh Nguyen, Yiğit Oğuz, Enes Yavuz Ugan, Jan Niehues, and Alexander Waibel. 2026. Muscat: Multilingual, scientific conversation benchmark. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 5926–5937, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Pournima Sonawane and Haithem Afli. 2026. ADAPT–MTU HAI at IWSLT2026: Robust Cascaded Speech Translation for Bhojpuri–Hindi and Irish–English. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Matthias Sperber, Ondřej Bojar, Barry Haddow, Dávid Javorský, Xutai Ma, Matteo Negri, Jan Niehues, Peter Polák, Elizabeth Salesky, Katsuhito Sudoh, and Marco Turchi. 2024. Evaluating the IWSLT2023 speech translation tasks: Human annotations, automatic metrics, and segmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6484–6495, Torino, Italia. ELRA and ICCL.

Ruiyan Sun, Qingming Li, and Satoshi Nakamura. 2026. The CUHKSZ System for the IWSLT 2026 Low-Resource Speech-to-Text Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Qwen Team. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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

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.

Enes Yavuz Ugan, Maike Züfle, Yuka Ko, Supriti Sinhamahapatra, Fabian Retkowski, Seymanur Akti, Jan Niehues, and Alexander Waibel. 2026. Multilingual Long-Form Speech Instruction Following: KIT's Submission to IWSLT 2026. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. SubER - a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Zhihang Xie, Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2026. FBK's Long-form SpeechLLMs for IWSLT 2026 Instruction Following. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Yi Xing, Manli Yu, Pengfei Liu, and Helen Meng. 2026. Test-Time Adaptation of an Offline Multimodal Foundation Model for Simultaneous Speech Translation. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Zeyu Yang and Satoshi Nakamura. 2026. CUHKSZ Simultaneous Speech Translation System for IWSLT 2026. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Zeyu Yang, Lai Wei, Roman Koshkin, Xi Chen, and Satoshi Nakamura. 2026. Sasst: Leveraging syntax-aware chunking and llms for simultaneous speech translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):34358–34367.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2023. GigaST: A 10,000-hour Pseudo Speech Translation Corpus. In *Interspeech 2023*, pages 2168–2172.

ZamStats. 2012. 2010 census of population and housing - national analytical report.

Victor Eugen Zarzu and Vilem Zouhar. 2026. Hurdles of Automatic Metric for Speech Translation Evaluation. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.

Rodolfo Joel Zevallos, Marc Casals, John E. Ortega, Fabrício Carraro, Pol Buitrago, and Guillermo Cámbara. 2026. CATENG Submission for the IWSLT 2026: Dialectal and Low-resource Speech Translation Task. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (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 *International Conference on Learning Representations*.

Zihan Zhao, Yiyang Jiang, Heyang Liu, Yu Wang, and Yanfeng Wang. 2024. Librisqa: A novel dataset and framework for spoken question answering with large language models. *IEEE Transactions on Artificial Intelligence*.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. 2025. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. 2026. Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*.

Vilém Zouhar, Maike Züfle, Beni Egressy, Julius Cheng, Mrinmaya Sachan, and Jan Niehues. 2026. Early-exit and instant confidence translation quality estimation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 55–76, Rabat, Morocco. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models? *Transactions of the Association for Computational Linguistics*, 13:1789–1811.

Maike Züfle, Sara Papi, Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, and Jan Niehues. 2025. NUTSHELL: A dataset for abstract generation from scientific talks. *CoRR*, abs/2502.16942.

Maike Züfle, Danni Liu, Vilém Zouhar, and Jan Niehues. 2026. Why we need speech to evaluate speech translation. *Preprint*, arXiv:2605.28227.

Appendix A. Human Evaluation

A Human Evaluation

Human evaluation includes direct assessment for the offline, compression, and instruction following tasks (A.1).

A.1 Direct Assessment

For the offline translation track (Section I), compression track (Section III), and instruction following track (Section IX), we conduct a human evaluation of primary submissions. Human graders are asked for direct assessment (DA) (Graham et al., 2013; Cettolo et al., 2017; Akhbardeh et al., 2021) of the system output given the input audio, expressed as scores ranging from 0 to 100. We include the Business, CallCenter, TVSeries, ACL, and Accent sets in English to German and English to Chinese directions for human evaluation. Human eval is done over the entire set, with no subsampling performed.

Since many tasks have standardized their test sets, we evaluate all outputs for a given test set, across any task that used the respective test set. This gives us the opportunity to compare across tasks and get a general sense of the relative progress across tasks. Caution should be exercised when comparing systems across tasks, as the tasks may have somewhat different objectives.

A.1.1 Automatic Segmentation

We collect segment-level annotations based on the re-segmented test data, generating automatic resegmentations of the hypothesis based on the reference translation by mwerSegmenter.[81] Because we do not want issues from the segmentation to influence scores negatively, we follow Sperber et al. (2024) and provide translators not only with the source audio and system translation but also with the system translation of the previous and following segments. Segments are shuffled and randomly assigned to annotators to avoid bias related to the presentation order. Annotation is conducted by professional translators fluent in the source language and native in the target language.

A.1.2 Computing System Rankings

System rankings are produced from average DA scores, normalized according to each individual annotator's mean and standard deviation, following the procedure of Akhbardeh et al. (2021). Statistical significance of pairwise differences between (system, task) cells is established via the Wilcoxon rank-sum (Mann-Whitney U) test with $p < 0.05$, applied to per-segment averaged z-scores over the common source segments of each pair. Within a given test set, every task uses the same source audio segments, so cross-task cell pairs are testable on the full segment set. Rank ranges are derived from win/loss counts in the clustering procedure.

We present one ranking table per (language pair, test set) combination, with each (system, task) combination as a separate row. Annotation items were shuffled and distributed randomly across annotators, so that each annotator saw a mix of test sets and tasks within a vendor batch. This allows per-annotator score normalization to be applied uniformly.

To increase the robustness of the evaluation, annotations were collected from two independent vendor pools, which produced 3 and 1 annotations per source segment, respectively. Scores from both pools are pooled prior to ranking and per-annotator normalization. Because ranks are derived from a rank-based test, two systems with a sizeable mean-score gap may still fall in the same cluster when their per-segment score distributions overlap heavily; the median score column is included to help interpret such cases.

A.2 Towards Efficient Evaluation

The previous IWSLT evaluation campaign (Abdulmumin et al., 2025) reported negative results in using informative subset selection. This was attributed to the lower quality of human evaluation in terms of mismatched segmentation. This year's human evaluation does not suffer from the same problems. This

[81] www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

enables using smaller number of human annotations, such as based on item-level diversity of translation outputs (Zouhar et al., 2025), to arrive at the same evaluation outcome with only a portion of the evaluation cost (Figure 1).

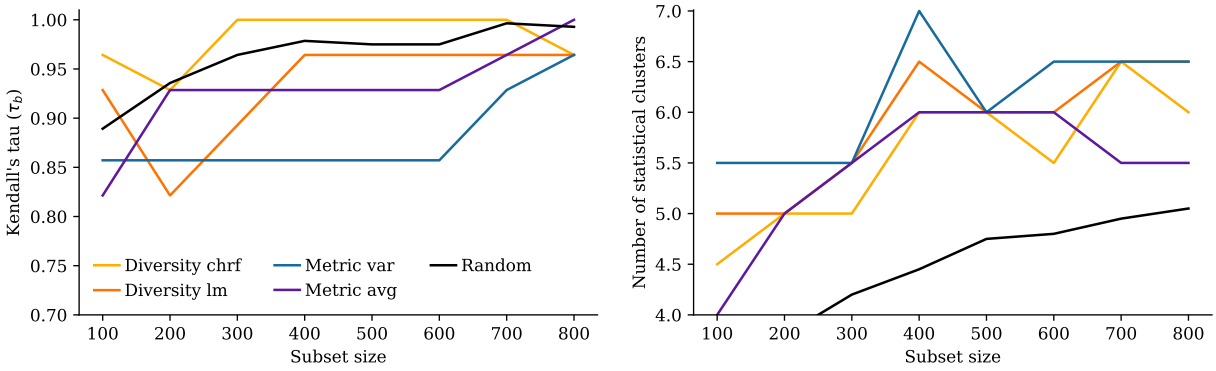

Figure 1: System-level ranking correctness and number of statistical clusters with respect to the proportion of evaluated subset size. Selecting based on character-level F-score diversity in model outputs leads to both correct correlation and ranking that is statistically stable.

Table 18: Human evaluation scores for en→de on the ACL test set (968 source segments; 3,872 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	90.6	+0.418	92.5
if_short	KIT	2-3	86.5	+0.258	90.0
offline	HW-TSC	2-3	82.3	+0.098	91.0
if_short	BSC	4	83.0	+0.120	88.0
if_short	NLE	5-6	79.5	-0.016	83.0
if_short	FBK	5-6	78.4	-0.054	82.0
if_long	KIT	7	70.2	-0.381	82.0
if_long	FBK	8	68.5	-0.442	78.8

Table 19: Human evaluation scores for en→de on the Accent test set (1,448 source segments; 5,792 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	80.1	+0.115	88.0
offline	HW-TSC	2	73.6	-0.115	83.8

Table 20: Human evaluation scores for en→de on the Business test set (441 source segments; 1,764 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	89.8	+0.135	91.8
offline	HW-TSC	2	84.9	-0.135	88.5

Table 21: Human evaluation scores for en→de on the CallCenter test set (1,357 source segments; 5,428 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	87.8	+0.108	91.8
offline	HW-TSC	2	82.9	-0.108	90.0

Table 22: Human evaluation scores for en→de on the TVSeries test set (1,488 source segments; 5,952 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1-2	76.6	+0.079	84.0
offline	HW-TSC	1-2	71.2	-0.079	84.8

Table 23: Human evaluation scores for en→zh on the ACL test set (968 source segments; 3,872 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	89.3	+0.394	91.0
if_short	KIT	2-3	87.8	+0.327	90.0
offline	HW-TSC	2-3	81.3	+0.060	89.8
if_short	NLE	4	83.0	+0.140	86.5
if_short	FBK	5-7	82.3	+0.107	85.1
if_short	BSC	5-7	79.7	+0.003	85.2
if_long	KIT	5-7	74.3	-0.224	86.5
if_long	FBK	8	60.0	-0.806	75.8

Table 24: Human evaluation scores for en→zh on the Business test set (441 source segments; 1,764 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	88.1	+0.076	89.5
offline	HW-TSC	2	86.1	-0.076	89.2

Table 25: Human evaluation scores for en→zh on the CallCenter test set (1,357 source segments; 5,428 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	89.4	+0.145	91.5
offline	HW-TSC	2	83.9	-0.145	90.0

Table 26: Human evaluation scores for en→zh on the TVSeries test set (1,506 source segments; 6,024 annotations; 4 annotations per source segment).

Task	System	Rank ↑	Mean Score	Mean z	Median Score
offline	ref	1	82.8	+0.155	90.2
offline	HW-TSC	2	73.6	-0.155	86.2

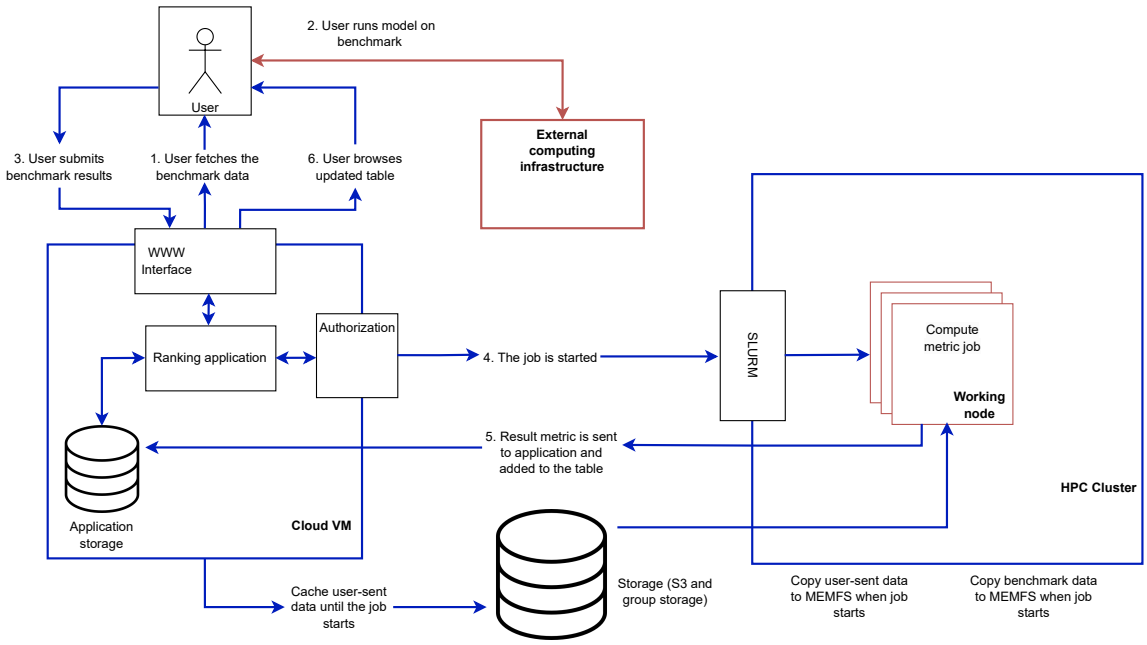

Figure 2: SPEECHM architecture. The platform is composed of the WebUI for managing user submissions and showing evaluation results, produced by the evaluation scripts executed in the scope of Slurm jobs on the HPC Ares (for CPU-based calculations) and Athena (for GPU-based calculations) HPC clusters.

Appendix B. Automatic Evaluation Results and Details

B.1 Evaluation Server

B.1.1 Introduction

The Evaluation Server is a collection of benchmarking resources and tools to evaluate the capability of user systems with respect to a set of tasks. It is part of the SPEECHM platform, released by the Meetween European Project[82], which consists of (a) ten downstream tasks, (b) a set of task-dependent evaluation metrics and (c) a WebUI for submissions and performance tracking by means of a leaderboard.

For the IWSLT-2025 Evaluation Campaign a dedicate instance of the SPEECHM has been developed, named SPEECHM-IWSLT2025[83]. It supports three of the IWSLT-2025 shared tasks, namely the *Offline*, the *Model Compression* and the *Instruction Following* tasks.

B.1.2 User operations

Given a task testset (e.g. the TvSeries English-German testset for the Offline SLT task), users typically perform the following operations:

1. download the source data (i.e. the English audios archive);
2. run their system and produce the hypothesis output (i.e. the German translations)
3. submitt their system output (i.e. the German translations);
4. wait for the evaluation process and read the evaluation scores (e.g. the COMET, and BLEU scores).

The SPEECH-IWSLT2025 allows the users to perform the above operations except the 2. one (users are expected to run their systems outside the Evaluation Server). In addition users can also delete and replace a submission with another one.

Submissions are managed trough the concept of user *models*, a user-defined entity that describes the main features of a given user system. By means of models, users can submitt multiples hypothesis

[82] www.meetween.eu
[83] iwslt2025.speechm.cloud.cyfronet.pl

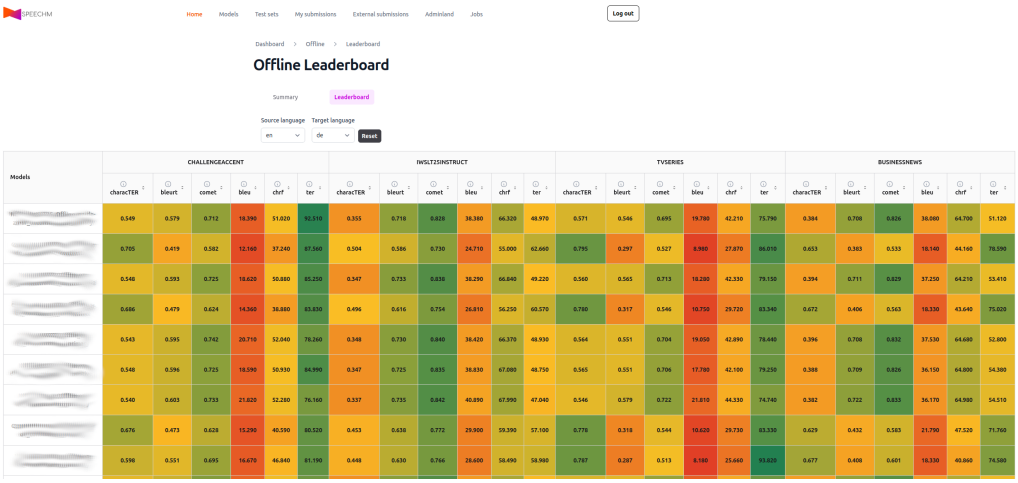

Figure 3: SPEECHM leaderboard.

outputs for the same task testset, one for each different developed system.

B.1.3 The Web UI

The Web UI facilitates the submission process, manages evaluation submissions, and monitors interactions with the external HPC cluster. This workflow is illustrated in Figure 2. Initially, users must create an account in the SPEECHM system, a straightforward process due to its integration with PLGrid, GitHub, and Google identity providers. Once registered, users can download the challenge input files (Step 1). These files serve as input for the participant's model inference (Step 2), which must currently be performed outside the SPEECHM system. In future iterations, SPEECHM aims to integrate this step as well.

After generating the outputs, users can conveniently upload them to the SPEECHM portal (Step 3). At this stage, challenge owners initiate the hypothesis evaluation process (Step 4). This step is restricted to challenge owners since they alone have access to the HPC computational resources required for evaluation. SPEECHM employs *slurmrestd*[84] to submit SLURM jobs to HPC clusters and to monitor job execution status.

Upon completion of the evaluations, the scores are stored in the SPEECHM database (Step 5). These scores contribute to generating various leaderboards, such as those specific to a task, testset, or model. An example leaderboard is shown in Figure 3.

B.1.4 The evaluation scripts

The evaluation metrics are computed through a set of scripts that run on the PLGRID clusters[85]. Scripts to compute the metrics that benefit from usage of GPU cards (such as COMET, BLEURT and BERT scores) run on the *Athena*[86] cluster while the other scripts (computing ASR, BLEU and CharacTER scores) are executed on the *Ares* cluster[87].

It is worth noticing here that while the references of the Offline and Model Compression task testsets are typically unstructured plain files, those of the Instruction Following task are structured as XML files. Therefore, the evaluation script for the Instruction Following task testsets has been developed specifically in order to manage the XML input structure.

B.2 Offline Speech Translation

B.3 Offline SLT

- Systems are ordered according to the COMET score (denoted by COMET, the first column).

[84] slurm.schedmd.com/slurmrestd.html
[85] portal.plgrid.pl
[86] www.cyfronet.pl/en/19073,artykul,athena.html
[87] www.cyfronet.pl/en/computers/18827,artykul,ares_supercomputer.html

- The "Joint" table is computed by averaging the scores of all the test sets, aka macro-averaging.
- The "D" column indicates the data condition in which each submitted run was trained, namely: Constrained (C), Constrained^{+LLM} (C^{+}), Unconstrained (U).
- This year, we have submissions of both cascade and end-to-end architectures.

System	D	Joint					
		COMET (↑)	BLEU (↑)	BLEURT (↑)	chrF (↑)	CharacTER (↓)	TER (↓)
HW-TSC	U	0.796	38.2	0.590	66.64	0.450	54.58
		TV Series					
HW-TSC	U	0.69	17.27	0.520	43.75	0.582	88.35
		Scientific Presentations					
HW-TSC	U	0.874	48.22	0.601	79.13	0.392	43.4
		Call Center					
HW-TSC	U	0.797	35.07	0.668	62.23	0.421	50.78
		YouTube videos					
HW-TSC	U	0.67	31.12	0.496	57.16	0.577	55.56
		Business News					
HW-TSC	U	0.803	36.09	0.669	64.93	0.386	56.92
		Accented English Conversations					
HW-TSC	U	0.698	22.28	0.559	52.53	0.564	75.24
		Synthetic TTS Audio data 1					
HW-TSC	U	0.877	51.3	0.607	80.29	0.356	39.65
		Synthetic TTS Audio data 2					
HW-TSC	U	0.879	50.62	0.61	80.06	0.380	40.23
		Synthetic TTS Audio data 3					
HW-TSC	U	0.874	50.19	0.578	79.68	0.392	41.21

Table 27: Official results of the automatic evaluation for the Offline Speech Translation Task on official test sets, **English to German**.

B.4 Low-resource Speech Translation

System	D	Joint					
		COMET (↑)	BLEU (↑)	BLEURT (↑)	chrF (↑)	CharacTER (↓)	TER (↓)
HW-TSC	U	0.847	44.55	0.585	39.10	0.563	55.55
		TV Series					
HW-TSC	U	0.738	26.16	0.494	22.81	0.673	66.42
		Scientific Presentations					
HW-TSC	U	0.911	55.39	0.605	48.63	0.496	55.98
		Call Center					
HW-TSC	U	0.83	35.57	0.645	30.69	0.57	49.31
		YouTube videos					
HW-TSC	U	0.715	28.48	0.475	25.45	0.746	57.24
		Business News					
HW-TSC	U	0.854	40.09	0.649	34.75	0.582	55.59
		Synthetic TTS Audio data 1					
HW-TSC	U	0.909	57.62	0.607	50.77	0.465	51.72
		Synthetic TTS Audio data 2					
HW-TSC	U	0.908	56.86	0.606	50.27	0.474	54.38
		Synthetic TTS Audio data 3					
HW-TSC	U	0.907	56.22	0.598	49.39	0.495	53.73

Table 28: Official results of the automatic evaluation for the Offline Speech Translation Task on official test set, **English to Chinese**. When computing the TER scores via sacreBLEU, we provide these two additional arguments: "–ter-normalized" and "–ter-asian-support"

B.5 Compression Task

Table 29: COMET scores across datasets and settings for English–German (en–de).

Setting	Storage (GB)	ACL6060	BUSINESSNEWS	CALLCENTER	CHALLENGEACCENT	TVSERIES	YOUTUBE
Full Precision	16.8	0.665	0.405	0.507	0.546	0.387	0.415
4-bit	5.9	0.591	0.357	0.457	0.504	0.378	0.379
TalTech_constrained_primary	5.1	0.319	0.122	0.155	0.668	0.169	0.150
KIT_unconstrained_primary	4.0	0.767	0.630	0.617	0.363	0.474	0.504
KIT_unconstrained_contrastive1	4.0	0.756	0.616	0.627	0.364	0.490	0.519

Table 30: COMET scores across datasets and settings for English-Chinese (en-zh).

Setting	Storage (GB)	ACL6060
Full Precision	16.8	0.538
4-bit	5.9	0.575
APG_constrained_primary	10.3	0.339

B.6 Subtitling Task

Trg lang.	Set	Team	System	Sub. qual. SubER	Translation quality Bleu	ChrF	Bleurt	Subtitle compliance CPS	CPL	LPB
ar	tst26	AppTek	prmry	67.31	17.85	48.80	.5688	99.81	100.00	100.00
		AppTek	cntrs1	67.60	18.11	49.09	.5699	93.61	100.00	100.00
		FBK	prmry	70.22	16.05	47.18	.5583	91.15	94.98	86.36
		FBK	cntrs1	73.03	14.11	44.57	.5206	91.69	100.00	100.00
		FBK	cntrs2	77.27	14.08	44.41	.5244	81.40	100.00	99.94
	tst25	AppTek	prmry	61.64	22.28	53.38	.6054	99.91	100.00	99.97
		AppTek	cntrs1	61.72	22.60	53.75	.6078	92.46	100.00	99.97
		AppTek 25	prmry	62.13	21.55	52.75	.5945	99.81	100.00	100.00
		FBK	prmry	65.60	19.29	50.93	.5994	91.43	94.41	88.05
		FBK	cntrs1	67.92	17.66	49.13	.5680	91.62	100.00	100.00
		FBK	cntrs2	76.63	16.42	47.93	.5564	74.68	100.00	99.95
de	tst26	FBK	prmry	52.22	37.80	63.26	.6539	72.16	92.83	75.55
		AppTek	prmry	53.63	32.86	59.10	.5979	77.25	100.00	99.01
		FBK	cntrs2	54.58	32.50	59.21	.5980	71.42	100.00	99.95
		AppTek	cntrs1	54.59	30.95	57.01	.5826	93.55	100.00	99.01
		FBK	cntrs1	56.51	33.55	59.59	.6093	73.23	100.00	100.00
	tst25	FBK	prmry	48.72	41.51	65.36	.6657	71.46	92.74	79.11
		AppTek	prmry	49.84	37.94	62.20	.6234	73.50	100.00	98.92
		AppTek 25	prmry	50.87	35.25	59.17	.6020	92.44	100.00	100.00
		AppTek	cntrs1	51.00	35.03	59.26	.6037	92.72	100.00	98.92
		FBK	cntrs1	52.69	37.48	62.82	.6333	71.69	100.00	100.00
		FBK	cntrs2	53.05	36.34	61.99	.6215	69.62	100.00	99.97
ja	tst26	AppTek	prmry	76.05	11.57	16.49	.2545	26.10	100.00	96.22
		AppTek	cntrs1	76.94	11.25	16.33	.2575	25.17	100.00	95.97
		FBK	cntrs2	84.85	8.74	13.54	.2309	9.30	58.96	100.00
		FBK	prmry	86.64	10.39	15.72	.2488	19.18	69.73	100.00
		FBK	cntrs1	107.20	4.72	7.96	.0913	14.07	100.00	100.00
zh	tst26	AppTek	cntrs1	58.78	29.82	25.66	.5094	100.00	100.00	100.00
		AppTek	prmry	59.39	30.44	25.82	.5129	99.88	100.00	99.94
		HW-TSC	prmry	59.72	33.64	29.23	.5316	98.82	99.26	100.00
		FBK	cntrs2	61.55	28.34	24.29	.4852	96.89	50.76	100.00
		FBK	prmry	78.15	30.13	25.74	.4762	95.58	89.17	100.00
		FBK	cntrs1	89.15	28.43	24.40	.4376	96.50	100.00	100.00

Table 31: Subtitling Task: automatic evaluation scores on Asharq-Bloomberg domain. Systems are sorted by SubER score within each group. *prmry*/*cntrstv* stands for *primary*/*contrastive* systems. Results obtained on the legacy test set (tst25) by re-evaluating runs submitted in previous shared task editions are shown in gray.

Trg lang.	Set	Team	System	Sub. qual. SubER	Translation quality Bleu	ChrF	Bleurt	Subtitle compliance CPS	CPL	LPB
de	tst26	AppTek	prmry	54.86	31.16	56.92	.6079	75.26	99.97	97.91
		AppTek	cntrs1	55.87	29.78	55.45	.5752	76.11	100.00	99.95
		AppTek	cntrs2	56.95	26.71	52.24	.5572	95.28	100.00	99.95
		FBK	prmry	58.14	31.21	57.13	.5823	69.08	95.29	97.49
		FBK	cntrs1	60.07	29.74	56.24	.5877	71.08	100.00	100.00
		FBK	cntrs2	61.25	27.29	52.41	.5334	69.10	100.00	100.00
ja	tst26	AppTek	prmry	73.59	16.42	22.08	.3455	23.88	99.97	91.29
		AppTek	cntrs1	76.92	10.36	15.83	.2640	34.63	100.00	98.31
		FBK	cntrs2	78.07	11.41	17.14	.2994	21.21	52.20	100.00
		FBK	prmry	83.67	11.10	16.53	.2709	40.81	78.23	100.00
		FBK	cntrs1	85.69	8.88	15.37	.2614	32.90	100.00	100.00
zh	tst26	AppTek	prmry	61.69	26.54	23.19	.5132	98.61	100.00	99.75
		AppTek	cntrs1	63.44	23.37	20.42	.4744	99.85	100.00	99.90
		HW-TSC	prmry	66.99	24.78	21.06	.4827	91.06	99.79	99.94
		FBK	cntrs1	68.79	23.47	20.31	.4823	92.23	100.00	100.00
		FBK	prmry	69.32	22.28	19.44	.4545	92.69	90.91	100.00
		FBK	cntrs2	73.65	18.97	16.61	.4032	73.42	83.01	100.00

Table 32: Subtitling Task: automatic evaluation scores on YODAS domain. Systems are sorted by SubER score within each group. *prmry* and *cntrstv* stands for *primary* and *contrastive* systems, respectively.

Trg lang.	Set	Team	System	Sub. qual. SubER	Translation quality Bleu	ChrF	Bleurt	Subtitle compliance CPS	CPL	LPB
de	tst26	AppTek	cntrs1	67.30	21.36	46.00	.5179	99.26	100.00	100.00
		Fbk	cntrs1	71.57	18.93	43.69	.4941	81.80	100.00	100.00
		AppTek	prmry	74.01	18.62	47.83	.5408	90.48	100.00	94.67
		Fbk	prmry	75.51	18.41	44.68	.5252	79.96	96.34	98.51
		Fbk	cntrs2	78.40	16.04	36.16	.3853	76.30	100.00	100.00
	tst25	Fbk	cntrs1	62.91	19.97	41.61	.5108	71.23	100.00	100.00
		AppTek	cntrs1	64.16	17.83	40.06	.4839	98.51	100.00	100.00
		Fbk	prmry	64.76	21.73	44.02	.5529	67.77	96.27	96.65
		AppTek	prmry	64.89	20.50	44.68	.5356	83.67	100.00	93.53
		Fbk	cntrs2	68.88	17.10	37.91	.4389	67.68	100.00	100.00
	tst24	AppTek	cntrs1	66.28	18.87	41.68	.4966	98.13	100.00	99.96
		Fbk	cntrs1	70.03	18.40	42.12	.5088	71.83	100.00	100.00
		AppTek	prmry	71.44	18.29	44.49	.5257	83.20	100.00	93.42
		Fbk	prmry	71.78	18.74	43.61	.5380	70.29	95.46	95.85
		Fbk	cntrs2	75.90	16.13	37.34	.4288	67.77	100.00	100.00
	tst23	AppTek	cntrs1	64.27	19.01	41.91	.5068	98.38	100.00	100.00
		AppTek 25	prmry	65.26	18.80	41.83	.5012	93.32	100.00	100.00
		Fbk	cntrs1	66.98	19.95	42.51	.5091	70.58	100.00	100.00
		AppTek	prmry	67.91	20.09	46.26	.5520	81.72	100.00	93.67
		AppTek 24	prmry	68.70	17.96	41.40	.4720	67.64	100.00	99.96
		AppTek 23	prmry	69.15	14.42	35.51	.4023	86.01	100.00	100.00
		Fbk	prmry	69.50	20.25	44.81	.5492	67.84	95.65	95.92
		Hw-Tsc 24	prmry	70.97	18.33	42.97	.5057	60.15	62.37	100.00
		Tlt 23	prmry	71.35	14.90	37.26	.4438	80.21	99.47	100.00
		Fbk	cntrs2	73.19	17.30	38.00	.4301	68.01	100.00	100.00
		Fbk 24	prmry	74.25	16.16	36.10	.3928	54.70	92.97	100.00
		Fbk 23	prmry	81.45	8.03	26.42	.2283	67.75	85.12	100.00
es	tst26	AppTek	cntrs1	60.49	23.36	47.81	.5630	99.73	100.00	100.00
		AppTek	prmry	62.86	23.34	48.86	.5735	94.06	100.00	97.84
		Fbk	cntrs1	63.28	21.17	43.84	.5152	86.37	100.00	100.00
		Fbk	prmry	65.65	22.53	45.49	.5488	82.96	96.23	99.46
		Fbk	cntrs2	70.32	17.11	37.19	.4030	82.16	100.00	100.00
	tst24	AppTek	cntrs1	62.56	23.04	46.29	.5115	99.30	100.00	100.00
		AppTek	prmry	64.87	23.60	48.56	.5378	88.64	100.00	97.27
		Fbk	cntrs1	66.41	21.71	44.54	.4954	76.80	100.00	100.00
		Fbk	prmry	67.75	22.15	46.07	.5251	75.67	96.39	96.91
		Fbk	cntrs2	71.59	18.23	39.87	.4220	73.93	100.00	100.00
	tst23	AppTek	cntrs1	62.42	23.85	47.32	.5266	99.34	100.00	100.00
		AppTek	prmry	64.67	23.65	49.70	.5514	88.12	100.00	97.05
		Fbk	cntrs1	65.59	22.83	45.61	.4962	75.89	100.00	100.00
		AppTek 24	prmry	66.55	22.05	45.49	.4782	77.61	100.00	100.00
		Hw-Tsc 24	prmry	66.78	22.44	46.67	.5098	68.95	67.58	100.00
		Fbk	prmry	67.28	23.16	47.69	.5356	73.84	96.22	97.20
		Tlt 23	prmry	69.34	18.52	41.41	.4530	81.93	99.51	100.00
		Fbk 24	prmry	70.35	19.15	40.08	.3959	62.11	94.22	100.00
		Fbk	cntrs2	73.08	19.58	40.49	.4122	70.60	100.00	100.00
		AppTek 23	prmry	80.33	11.23	29.87	.2478	94.67	100.00	100.00
		Fbk 23	prmry	81.41	9.23	27.44	.2083	74.67	92.94	100.00
ja	tst26	AppTek	prmry	78.81	7.61	12.19	.2521	91.88	100.00	99.86
		AppTek	cntrs1	82.55	8.50	14.26	.2872	75.78	100.00	99.59
		Fbk	prmry	91.44	7.21	14.06	.2845	61.60	73.89	100.00
		Fbk	cntrs1	96.88	7.37	14.00	.2661	56.63	100.00	100.00
		Fbk	cntrs2	97.94	5.39	10.41	.1831	23.16	68.88	100.00
zh	tst26	AppTek	cntrs1	56.43	32.85	27.89	.5419	99.93	100.00	99.93
		AppTek	prmry	57.13	33.27	28.14	.5681	99.53	100.00	99.26
		Fbk	cntrs1	63.84	24.85	21.51	.4703	97.25	100.00	100.00
		Fbk	prmry	64.78	24.94	21.47	.4931	96.57	93.38	100.00
		Hw-Tsc	prmry	70.87	24.18	21.77	.4978	84.22	99.89	100.00
		Fbk	cntrs2	79.72	17.98	15.81	.3426	68.81	94.02	100.00

Table 33: Subtitling Task: automatic evaluation scores on ITV domain. Systems are sorted by SubER score within each group. *prmry*/*cntrstv* stands for *primary*/*contrastive* systems. Results obtained on the legacy test sets (tst23, tst24 and tst25) by re-evaluating runs submitted in previous shared task editions are shown in gray.

B.7 Simultaneous Task

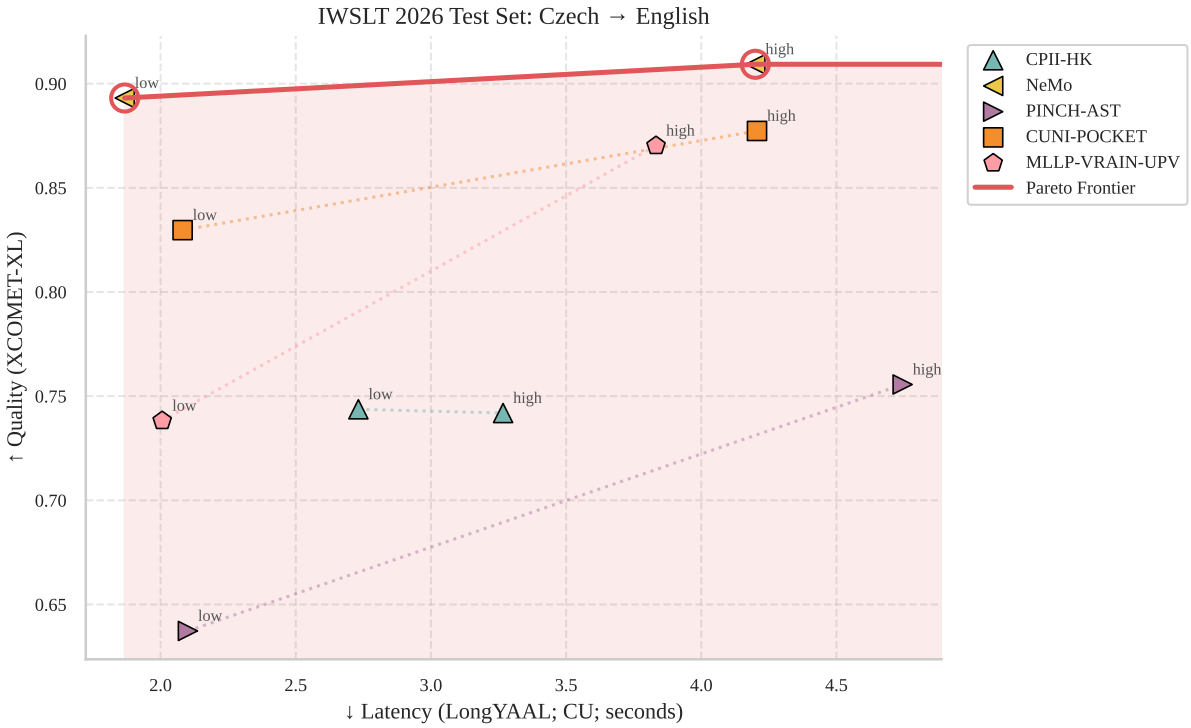

Figure 4: Quality-latency tradeoff curves for Czech→English Test Set. Latency is measured by computation-unaware LongYAAL in seconds, quality is measured by XCOMET-XL. Pareto frontier (red line) represents optimal systems where neither quality or latency can be improved without sacrificing the other.

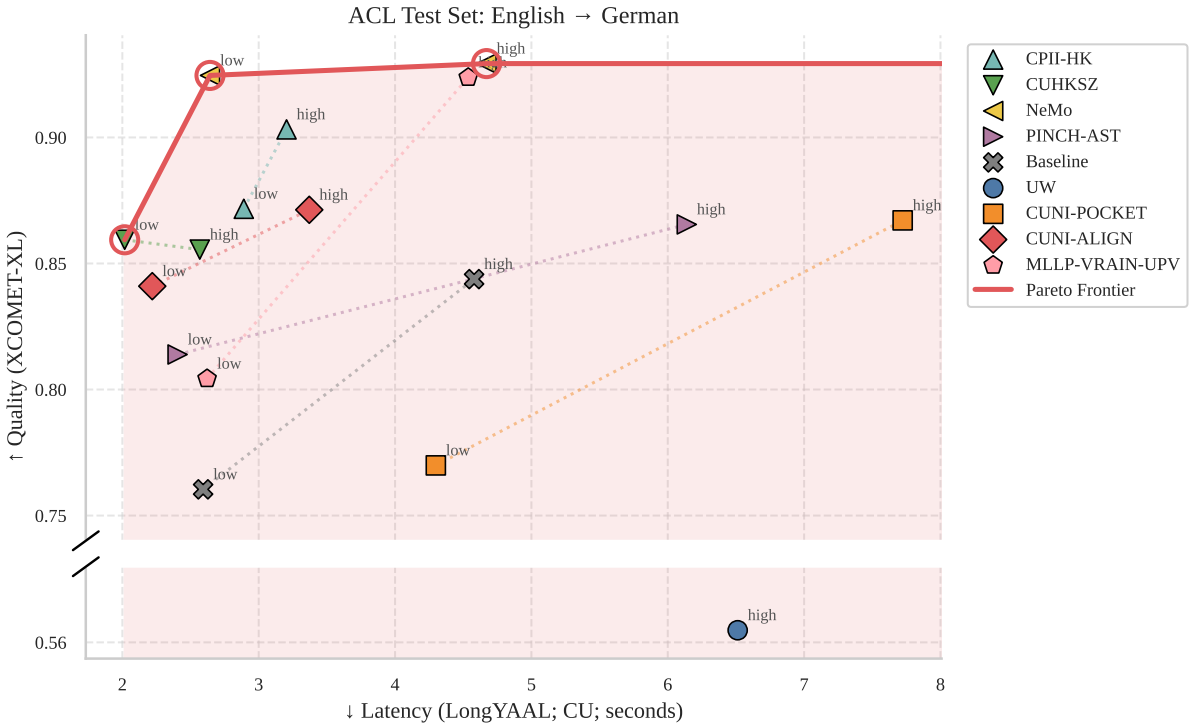

Figure 5: Quality-latency tradeoff curves for English→German ACL Test Set. Latency is measured by computation-unaware LongYAAL in seconds, quality is measured by XCOMET-XL. Pareto frontier (red line) represents optimal systems where neither quality or latency can be improved without sacrificing the other.

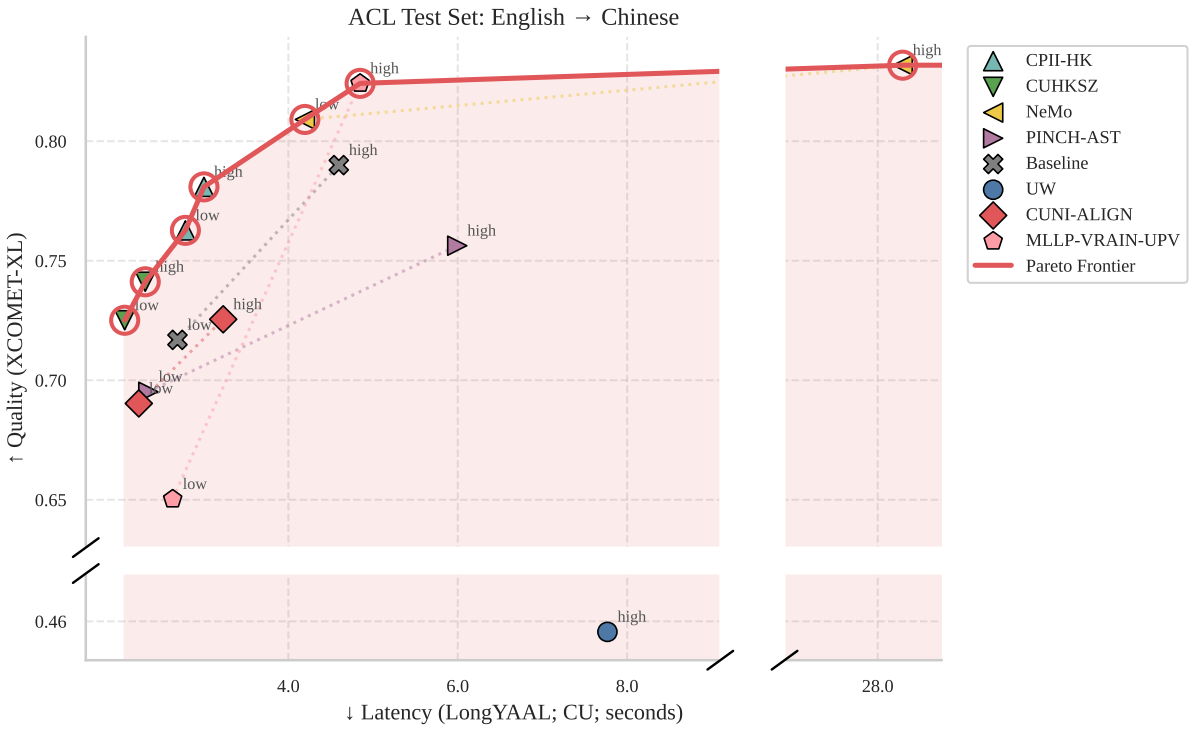

Figure 6: Quality-latency tradeoff curves for English→Chinese ACL Test Set. Latency is measured by computation-unaware LongYAAL in seconds, quality is measured by XCOMET-XL. Pareto frontier (red line) represents optimal systems where neither quality or latency can be improved without sacrificing the other.

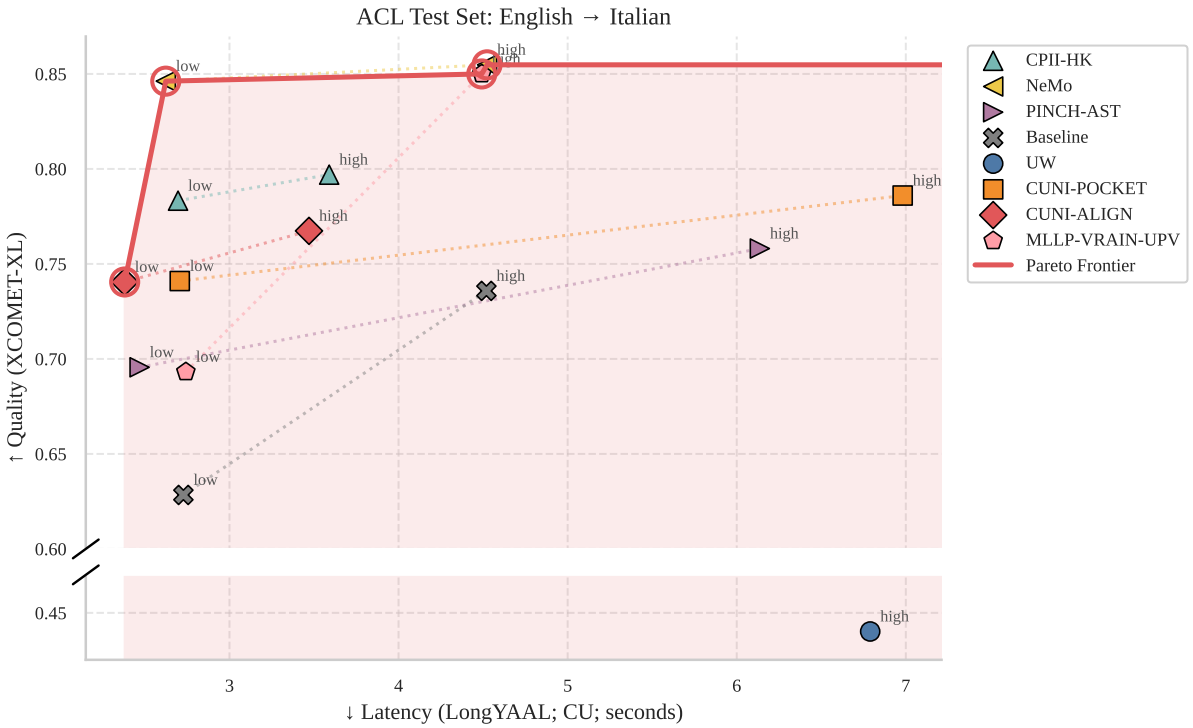

Figure 7: Quality-latency tradeoff curves for English→Italian ACL Test Set. Latency is measured by computation-unaware LongYAAL in seconds, quality is measured by XCOMET-XL. Pareto frontier (red line) represents optimal systems where neither quality or latency can be improved without sacrificing the other.

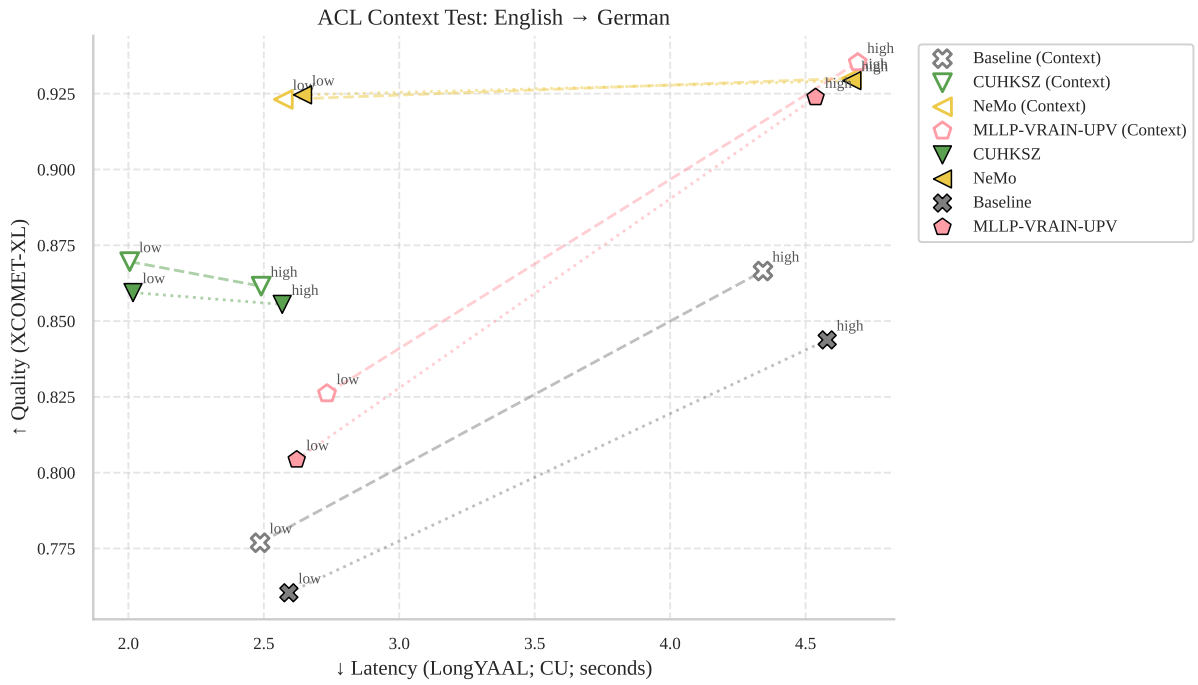

Figure 8: Quality-latency tradeoff curves for English→German ACL Test Set. Latency is measured by computation-unaware LongYAAL in seconds, quality is measured by XCOMET-XL.

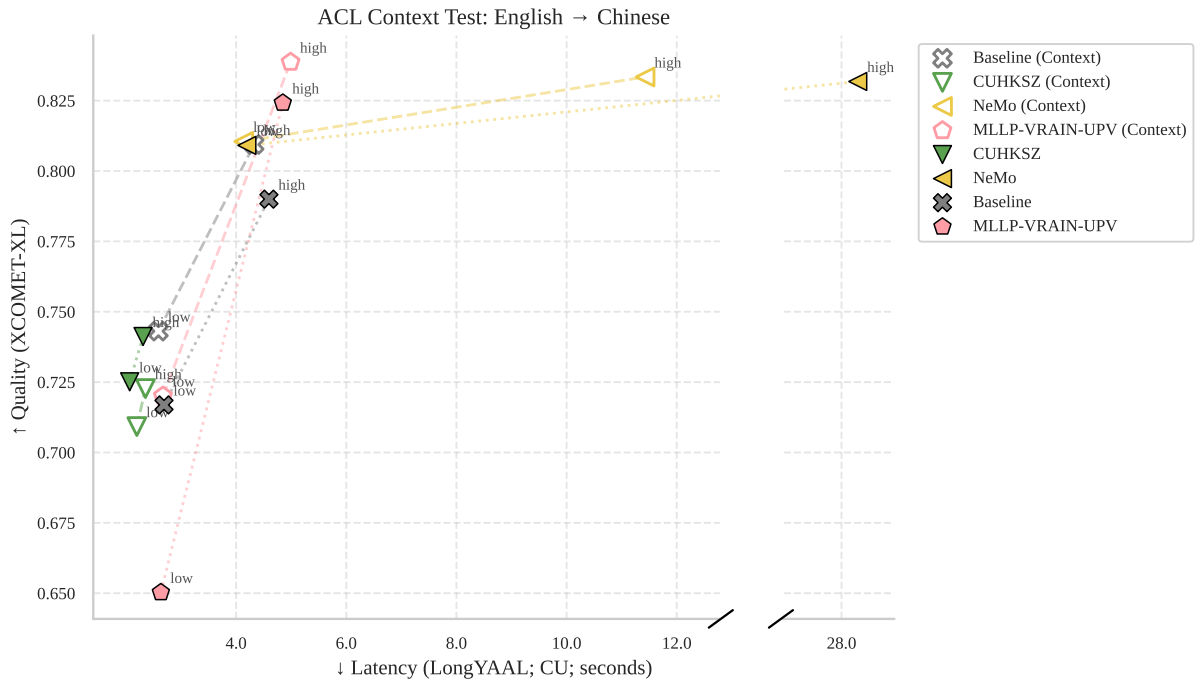

Figure 9: Quality-latency tradeoff curves for English→Chinese ACL Test Set. Latency is measured by computation-unaware LongYAAL in seconds, quality is measured by XCOMET-XL.

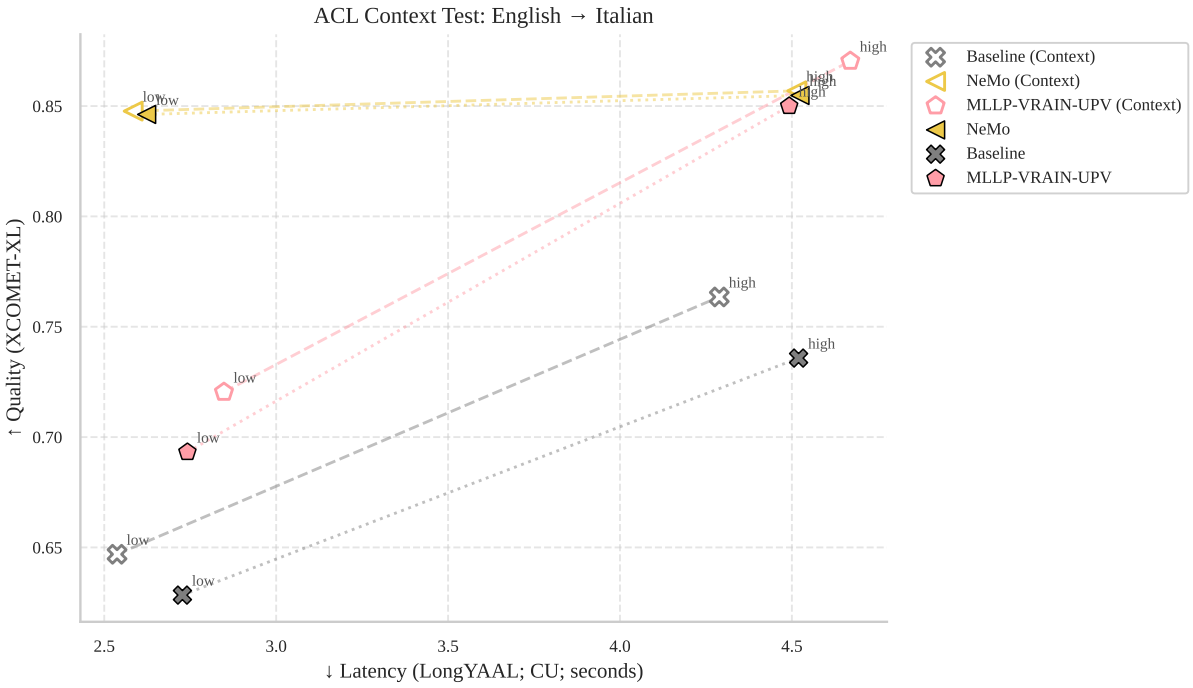

Figure 10: Quality-latency tradeoff curves for English→Italian ACL Context Test Set. Latency is measured by computation-unaware LongYAAL in seconds, quality is measured by XCOMET-XL.

Regime	Pl.	System	COMET	BLEU	chrF	LongYAAL	LongLAAL	LongDAL	StreamLAAL
IWSLT 2026 Test Set									
high	**1**	NeMo (logs)	**0.91**	**48.72**	**70.07**	4.2	4.0	5.9	4.1
	2	CUNI-Pocket	0.88	44.77	66.69	4.2 [4.6]	4.1 [4.5]	5.7 [6.1]	–
	2	MLLP-VRAIN UPV	0.87	47.06	69.78	3.8 [4.2]	3.9 [4.3]	5.0 [5.4]	**4.0 [4.3]**
	3	pinch-ast (logs)	0.76	31.37	59.21	4.7	4.6	7.0	4.7
	3	CPII-HK (logs)	0.74	31.65	56.58	**3.3**	**3.2**	**4.2**	–
low	**1**	NeMo (logs)	**0.89**	**49.21**	**70.62**	**1.9**	**1.8**	3.1	**1.8**
	2	CUNI-Pocket	0.83	39.74	64.34	2.1 [2.3]	2.1 [2.3]	3.2 [3.5]	–
	3	CPII-HK (logs)	0.74	32.22	58.72	2.7	2.7	3.9	–
	3	MLLP-VRAIN UPV	0.74	39.14	67.16	2.0	2.0	**2.9**	2.1
	4	pinch-ast (logs)	0.64	24.36	55.17	2.1	2.1	3.7	2.1
IWSLT 2026 Development Set									
high	**1**	NeMo (logs)	**0.87**	**38.68**	**65.29**	4.5	4.3	6.4	4.0
	2	MLLP-VRAIN UPV	0.83	34.81	63.07	**3.4**	**3.4**	4.7	**3.6**
	3	CUNI-Pocket	0.81	32.01	59.14	3.6	3.5	5.3	–
	4	CPII-HK (logs)	0.67	23.85	51.36	3.5	3.4	**4.6**	–
	4	pinch-ast (logs)	0.65	24.12	53.64	4.6	4.4	7.2	4.5
low	**1**	NeMo (logs)	**0.86**	**35.91**	**64.23**	2.3	2.1	3.6	1.9
	2	MLLP-VRAIN UPV	0.78	30.46	60.66	2.2	2.1	**3.0**	**1.8**
	2	CUNI-Pocket	0.76	27.78	56.68	**2.0**	**1.9**	3.3	–
	3	CPII-HK (logs)	0.67	23.14	53.30	2.7	2.7	4.1	–
	4	pinch-ast (logs)	0.54	17.78	49.37	2.5	2.5	4.2	2.3

Table 34: Results for Czech→English. Pl. = place within latency regime; shared with multiple systems if the scores are deemed not significantly different by a paired t-test with bootstrap resampling ($p > 0.05$; Koehn, 2004). Bold = best in the latency regime. Latency (in seconds): compute-unaware value [compute-aware value in brackets]. We do not report compute-aware latency for systems which were submitted as logs.

Regime	Pl.	System	COMET	BLEU	chrF	LongYAAL	LongLAAL	LongDAL	StreamLAAL
ACL Test Set									
high	**1**	NeMo (logs)	**0.93**	**45.01**	71.44	4.7	4.4	6.1	4.4
	2	MLLP-VRAIN UPV	0.92	44.88	**71.68**	4.5 [4.9]	4.4 [4.8]	5.5 [5.9]	4.4 [4.8]
	3	CPII-HK (logs)	0.90	36.56	67.19	3.2	3.2	4.2	–
	4	CUNI-Align	0.87	34.11	67.47	3.4 [3.6]	3.0 [3.2]	4.4 [4.6]	3.0 [3.3]
	4	CUNI-Pocket	0.87	36.98	63.60	7.7 [8.3]	7.7 [8.3]	9.4 [10.0]	–
	4	pinch-ast (logs)	0.87	34.23	65.39	6.1	5.8	7.8	5.8
	5	CUHKSZ	0.86	35.36	64.82	**2.6**	**2.5**	**3.4**	**2.5**
	5	*Baseline*	0.84	29.97	62.50	4.6 [4.8]	4.7 [4.9]	6.0 [6.2]	4.7 [4.9]
	6	UW	0.56	16.29	47.06	6.5 [6.8]	6.7 [7.0]	8.2 [8.5]	6.6 [6.9]
low	**1**	NeMo (logs)	**0.92**	**44.94**	**71.45**	2.6	2.5	3.6	2.6
	2	CPII-HK (logs)	0.87	32.93	63.64	2.9	2.7	3.5	–
	3	CUHKSZ	0.86	33.24	63.99	**2.0**	**2.0**	**2.6**	**2.0**
	4	CUNI-Align	0.84	30.61	65.23	2.2 [2.4]	2.1 [2.2]	3.0 [3.2]	2.1 [2.2]
	5	pinch-ast (logs)	0.81	30.07	63.38	2.4	2.3	3.2	2.4
	5	MLLP-VRAIN UPV	0.80	38.06	68.98	2.6	2.4	3.3	2.7
	6	CUNI-Pocket	0.77	19.47	55.18	4.3 [4.7]	4.0 [4.5]	5.7 [6.2]	–
	6	*Baseline*	0.76	23.34	59.32	2.6 [2.7]	2.5 [2.7]	3.5 [3.6]	2.6 [2.7]
ACL Context Test Set									
high	**1**	MLLP-VRAIN UPV	**0.94**	**45.04**	71.35	4.7	4.6	5.7	4.7
	2	NeMo (logs)	0.93	44.84	**71.44**	4.7	4.3	6.1	4.4
	3	*Baseline*	0.87	31.08	63.56	4.3	4.5	5.7	4.5
	3	CUHKSZ	0.86	34.91	64.83	**2.5**	**2.4**	**3.4**	**2.5**
low	**1**	NeMo (logs)	**0.92**	**45.42**	**71.49**	2.6	2.4	3.5	2.5
	2	CUHKSZ	0.87	34.04	65.02	**2.0**	**2.0**	**2.6**	**1.9**
	3	MLLP-VRAIN UPV	0.83	39.26	69.36	2.7	2.6	3.5	2.7
	4	*Baseline*	0.78	24.43	60.61	2.5	2.5	3.4	2.6
Bloomberg Test Set									
high	**1**	NeMo (logs)	**0.88**	36.32	64.92	5.8	5.0	7.9	5.0
	2	MLLP-VRAIN UPV	0.83	**38.91**	**65.98**	5.0 [5.5]	4.6 [5.1]	6.4 [6.9]	4.5 [4.9]
	3	CUNI-Pocket	0.80	34.53	60.51	9.7	8.9	12.0	–
	3	CPII-HK (logs)	0.79	32.91	62.21	**3.9**	**3.6**	**4.9**	–
	4	pinch-ast (logs)	0.76	28.53	59.07	5.6	4.8	7.8	4.7
	5	*Baseline*	0.73	25.78	56.14	4.5 [4.7]	4.3 [4.5]	6.1 [6.4]	**4.3 [4.5]**
	6	UW	0.40	16.11	46.49	6.6 [6.9]	6.2 [6.6]	8.8 [9.1]	5.6 [5.9]
low	**1**	NeMo (logs)	**0.87**	**39.03**	**66.40**	2.7	2.4	4.0	2.2
	2	CUNI-Align	0.74	29.04	61.49	2.1 [2.3]	**1.8 [2.0]**	3.3 [3.5]	**1.6 [1.8]**
	3	CUNI-Pocket	0.67	22.70	55.99	**2.1**	2.0	3.4	–
	3	CPII-HK (logs)	0.66	25.88	55.85	3.7	3.3	4.0	–
	4	MLLP-VRAIN UPV	0.65	33.31	64.33	3.5	3.1	4.8	3.1
	4	pinch-ast (logs)	0.64	26.00	57.74	2.2	2.0	**3.2**	1.9
Yodas Test Set									
high	**1**	NeMo (logs)	**0.79**	**29.57**	**56.73**	6.0	5.8	7.3	**6.1**
	2	CUNI-Pocket	0.77	26.69	52.61	8.3	8.3	9.5	–
	3	pinch-ast (logs)	0.75	24.18	52.00	7.2	7.1	8.7	7.3
	3	CPII-HK (logs)	0.74	24.98	51.71	**4.0**	**4.0**	**4.8**	–
	4	MLLP-VRAIN UPV	0.72	26.92	55.91	6.8 [7.4]	7.0 [7.6]	8.1 [8.6]	7.2 [7.8]
	5	CUNI-Align	0.70	18.67	51.20	4.9	19.0	20.5	12.9
	6	UW	0.52	9.17	34.60	8.1 [8.3]	8.1 [8.3]	9.1 [9.4]	8.7 [9.0]
low	**1**	NeMo (logs)	**0.79**	**29.47**	56.53	3.4	3.4	4.3	3.7
	2	CUNI-Align	0.72	27.21	54.43	3.1 [3.3]	3.1 [3.3]	3.9 [4.1]	3.2 [3.4]
	3	pinch-ast (logs)	0.69	22.30	50.70	**2.7**	**2.8**	**3.5**	**3.1**
	4	*Baseline*	0.67	21.15	49.04	3.2 [3.3]	3.3 [3.4]	4.0 [4.1]	3.5 [3.6]
	5	MLLP-VRAIN UPV	0.63	27.30	**56.77**	4.1	4.3	5.2	4.5
MCIF Development Set									
high	**1**	NeMo (logs)	**0.93**	**37.41**	**66.77**	4.1	3.9	5.6	3.9
	2	MLLP-VRAIN UPV	0.93	36.83	66.45	3.9	3.8	4.9	3.9
	3	CPII-HK (logs)	0.91	30.61	63.11	2.6	2.6	3.7	–
	4	CUNI-Pocket	0.88	31.73	60.83	3.8	3.8	5.2	–
	4	pinch-ast (logs)	0.87	29.44	61.80	5.4	5.1	7.1	5.2
	5	CUHKSZ	0.87	30.54	60.92	**2.0**	**2.0**	**2.9**	**2.0**
	6	UW	0.60	14.82	44.03	6.4	7.0	8.5	6.7
low	**1**	NeMo (logs)	**0.93**	**37.91**	**67.20**	2.0	1.9	2.9	1.9
	2	MLLP-VRAIN UPV	0.90	32.48	63.63	2.2	2.0	2.6	2.1
	3	CPII-HK (logs)	0.89	27.94	60.47	2.2	2.1	2.8	–
	4	CUHKSZ	0.85	27.72	59.58	**1.5**	**1.5**	**2.1**	**1.5**
	4	pinch-ast (logs)	0.84	25.62	59.77	1.8	1.8	2.6	1.8
	5	CUNI-Pocket	0.77	20.70	52.60	1.7	1.6	2.7	–

Table 35: Results for English→German. Pl. = place within latency regime; shared with multiple systems if the scores are deemed not significantly different by a paired t-test with bootstrap resampling ($p > 0.05$; Koehn, 2004). Bold = best in the latency regime. Latency (in seconds): compute-unaware value [compute-aware value in brackets]. We do not report compute-aware latency for systems which were submitted as logs.

Regime	Pl.	System	COMET	BLEU	chrF	LongYAAL	LongLAAL	LongDAL	StreamLAAL
ACL Test Set									
high	**1**	NeMo (logs)	**0.83**	47.64	40.30	28.3	27.4	30.2	27.4
	2	MLLP-VRAIN UPV	0.82	**50.03**	**42.75**	4.8 [5.2]	4.5 [4.9]	5.9 [6.3]	4.6 [4.9]
	3	*Baseline*	0.79	45.39	39.93	4.6 [4.8]	4.6 [4.8]	5.9 [6.1]	4.6 [4.8]
	3	CPII-HK (logs)	0.78	45.04	40.46	3.0	2.9	4.0	3.0
	4	PINCH-AST (logs)	0.76	41.70	34.98	6.0	5.5	7.7	5.6
	5	CUHKSZ	0.74	42.57	35.82	**2.3**	**2.2**	**2.8**	**2.3**
	6	CUNI-Align	0.73	34.04	33.98	3.2 [3.4]	2.8 [3.0]	4.3 [4.5]	2.9 [3.1]
	7	UW	0.46	15.96	14.39	7.8 [8.0]	7.3 [7.5]	8.8 [9.1]	7.0 [7.2]
low	**1**	NeMo (logs)	**0.81**	**46.31**	**39.44**	4.2	4.0	5.6	4.0
	2	CPII-HK (logs)	0.76	44.06	39.23	2.8	2.6	3.6	2.7
	3	CUHKSZ	0.73	41.47	34.58	**2.1**	**1.9**	**2.4**	**2.0**
	3	*Baseline*	0.72	40.68	35.17	2.7 [2.8]	2.6 [2.7]	3.6 [3.7]	2.6 [2.8]
	4	PINCH-AST (logs)	0.70	36.95	31.02	2.3	2.2	3.1	2.2
	4	CUNI-Align	0.69	32.40	31.70	2.2 [2.4]	2.0 [2.2]	3.1 [3.3]	2.1 [2.3]
	5	MLLP-VRAIN UPV	0.65	43.03	37.53	2.6	2.4	3.4	2.5
ACL Context Test Set									
high	**1**	MLLP-VRAIN UPV	**0.84**	50.24	**43.93**	5.0	4.6	6.1	4.7
	1	NeMo (logs)	0.83	**50.73**	43.80	11.4	11.0	13.3	11.0
	2	*Baseline*	0.81	46.47	42.32	4.3	4.4	5.7	4.5
	3	CUHKSZ	0.72	41.18	35.67	**2.4**	**2.2**	**2.9**	**2.4**
low	**1**	NeMo (logs)	**0.81**	**46.42**	39.60	4.1	3.9	5.6	3.9
	2	*Baseline*	0.74	40.62	37.28	2.6	2.5	3.5	2.5
	3	MLLP-VRAIN UPV	0.72	43.49	**40.10**	2.7	2.4	3.6	2.6
	3	CUHKSZ	0.71	39.41	34.13	**2.2**	**2.0**	**2.5**	**2.2**
Bloomberg Test Set									
high	**1**	NeMo (logs)	**0.75**	30.46	27.05	36.3	34.7	40.5	34.6
	2	MLLP-VRAIN UPV	0.70	**37.35**	**32.63**	5.7 [6.2]	5.1 [5.6]	6.9 [7.4]	5.2 [5.7]
	3	*Baseline*	0.67	35.43	30.58	5.0 [5.2]	4.6 [4.8]	6.4 [6.6]	4.7 [4.9]
	4	PINCH-AST (logs)	0.62	28.97	25.37	5.6	4.7	7.6	5.0
	5	CPII-HK (logs)	0.60	35.01	30.21	3.5	3.2	**4.3**	3.2
	5	CUNI-Align	0.59	31.30	28.13	**3.0 [3.3]**	**2.6 [2.8]**	4.5 [4.8]	**2.8 [3.1]**
	6	UW	0.33	11.49	12.47	8.2 [8.5]	7.1 [7.3]	9.3 [9.6]	8.4 [8.6]
low	**1**	NeMo (logs)	**0.72**	32.14	28.19	9.5	8.6	12.0	8.6
	2	*Baseline*	0.58	32.34	27.33	2.6 [2.8]	2.4 [2.5]	3.7 [3.8]	2.5 [2.6]
	2	CPII-HK (logs)	0.56	33.48	28.89	2.9	2.6	3.5	2.7
	3	CUNI-Align	0.55	27.88	25.77	**2.1 [2.3]**	**1.8 [1.9]**	3.3 [3.5]	2.0 [2.2]
	3	MLLP-VRAIN UPV	0.54	**35.67**	**30.36**	5.3	4.7	6.6	4.8
	4	PINCH-AST (logs)	0.51	27.67	24.05	2.3	2.0	**3.1**	**1.9**
Yodas Test Set									
high	**1**	NeMo (logs)	**0.68**	22.50	21.06	52.7	52.1	54.3	52.0
	2	MLLP-VRAIN UPV	0.66	**28.32**	**25.41**	6.5 [7.0]	6.5 [7.0]	7.4 [8.0]	7.5 [8.0]
	2	*Baseline*	0.65	22.57	21.25	**1.0 [1.2]**	**1.1 [1.3]**	**1.9 [2.2]**	**1.7 [1.9]**
	3	PINCH-AST (logs)	0.64	21.13	19.51	6.9	6.8	8.3	7.0
	4	CPII-HK (logs)	0.59	23.08	21.14	3.8	3.8	4.5	4.8
	4	CUNI-Align	0.59	21.91	19.61	5.1	5.0	6.1	4.0
	5	UW	0.49	5.33	8.51	8.4 [8.7]	7.9 [8.1]	8.9 [9.2]	9.0 [9.2]
low	**1**	NeMo (logs)	**0.66**	23.74	21.76	6.2	6.1	7.4	6.3
	2	*Baseline*	0.61	23.44	21.20	3.4 [3.5]	3.5 [3.6]	4.1 [4.3]	3.7 [3.8]
	2	CUNI-Align	0.60	24.40	21.35	3.2 [3.4]	3.3 [3.5]	4.1 [4.3]	3.4 [3.6]
	3	PINCH-AST (logs)	0.60	19.83	18.34	**2.8**	**2.9**	**3.5**	**3.0**
	4	CPII-HK (logs)	0.57	20.48	18.81	10.5	10.5	11.1	–
	5	MLLP-VRAIN UPV	0.54	**25.41**	**22.27**	3.9	3.9	4.8	4.2
MCIF Development Set									
high	**1**	NeMo (logs)	**0.84**	47.48	41.01	15.4	15.4	18.0	15.4
	2	MLLP-VRAIN UPV	0.83	**50.21**	**43.09**	4.2	3.9	5.3	4.0
	3	CPII-HK (logs)	0.79	46.17	42.65	2.4	2.4	3.4	2.4
	4	PINCH-AST (logs)	0.77	42.09	36.04	5.4	5.0	7.1	5.0
	4	CUHKSZ	0.77	46.67	40.12	**1.7**	**1.6**	**2.3**	**1.7**
	5	UW	0.50	16.03	15.43	6.9	6.9	8.3	7.5
low	**1**	NeMo (logs)	**0.82**	47.08	40.84	3.5	3.3	4.9	3.3
	2	MLLP-VRAIN UPV	0.78	**48.09**	**40.95**	2.1	1.9	2.7	2.0
	3	CPII-HK (logs)	0.77	44.44	40.44	2.2	2.2	3.1	2.2
	4	CUHKSZ	0.74	44.57	37.76	**1.5**	**1.5**	**2.0**	–
	5	PINCH-AST (logs)	0.72	38.28	32.81	1.8	1.7	2.5	**1.7**

Table 36: Results for English→Chinese. Pl. = place within latency regime; shared with multiple systems if the scores are deemed not significantly different by a paired t-test with bootstrap resampling ($p > 0.05$; Koehn, 2004). Bold = best in the latency regime. Latency (in seconds): compute-unaware value [compute-aware value in brackets]. We do not report compute-aware latency for systems which were submitted as logs.

Regime	Pl.	System	COMET	BLEU	chrF	LongYAAL	LongLAAL	LongDAL	StreamLAAL
ACL Test Set									
high	**1**	NeMo (logs)	**0.85**	**37.98**	**66.30**	4.5	4.3	5.9	4.3
	1	MLLP-VRAIN UPV	0.85	36.39	65.20	4.5 [4.9]	4.4 [4.8]	5.4 [5.8]	4.5 [4.8]
	2	CPII-HK (logs)	0.80	30.11	60.48	3.6	3.5	4.6	–
	2	CUNI-Pocket	0.79	31.95	60.44	7.0 [7.5]	7.0 [7.5]	8.6 [9.1]	**2.9 [3.4]**
	3	CUNI-Align	0.77	31.04	62.58	**3.5 [3.7]**	**3.2 [3.4]**	**4.5 [4.7]**	3.0 [3.3]
	3	pinch-ast (logs)	0.76	30.79	61.55	6.1	5.8	7.8	5.8
	4	*Baseline*	0.74	26.53	58.98	4.5 [4.7]	4.6 [4.8]	5.9 [6.1]	4.6 [4.8]
	5	UW	0.44	11.37	38.57	6.8 [7.0]	6.8 [7.1]	8.3 [8.6]	6.9 [7.2]
low	**1**	NeMo (logs)	**0.85**	**37.59**	**66.03**	2.6	2.5	3.5	2.5
	2	CPII-HK (logs)	0.78	29.25	61.01	2.7	2.6	3.4	–
	3	CUNI-Pocket	0.74	28.50	58.60	2.7	2.6	3.5	–
	3	CUNI-Align	0.74	27.97	60.82	**2.4 [2.6]**	**2.2 [2.4]**	**3.2 [3.4]**	**2.2 [2.4]**
	4	pinch-ast (logs)	0.70	27.90	59.68	2.5	2.4	3.2	2.4
	4	MLLP-VRAIN UPV	0.69	33.49	64.10	2.7	2.6	3.4	2.7
	5	*Baseline*	0.63	21.57	56.29	2.7 [2.9]	2.7 [2.8]	3.6 [3.8]	2.6 [2.8]
ACL Context Test Set									
high	**1**	MLLP-VRAIN UPV	**0.87**	37.38	65.72	4.7	4.5	**5.6**	4.6
	2	NeMo (logs)	0.86	**38.07**	**66.18**	4.5	**4.2**	5.9	**4.3**
	3	*Baseline*	0.76	27.57	59.91	**4.3**	4.4	5.7	4.4
low	**1**	NeMo (logs)	**0.85**	**37.30**	**66.01**	2.6	**2.4**	3.4	**2.5**
	2	MLLP-VRAIN UPV	0.72	34.67	64.66	2.8	2.6	3.6	2.7
	3	*Baseline*	0.65	22.12	56.54	**2.5**	2.5	**3.4**	2.6
MCIF Development Set									
high	**1**	NeMo (logs)	**0.88**	**50.58**	**73.66**	3.8	3.6	5.2	**3.7**
	1	MLLP-VRAIN UPV	0.88	49.55	72.90	3.8	3.8	4.8	3.8
	2	CPII-HK (logs)	0.85	43.13	69.85	**2.5**	**2.5**	**3.6**	–
	3	CUNI-Pocket	0.82	43.56	68.32	3.3	3.3	4.7	–
	4	pinch-ast (logs)	0.80	41.55	68.64	5.4	5.2	7.1	5.2
	5	UW	0.50	15.61	40.96	6.4	6.8	8.2	8.5
low	**1**	NeMo (logs)	**0.88**	**50.62**	**73.98**	1.8	1.7	2.8	1.8
	2	MLLP-VRAIN UPV	0.85	47.16	71.70	2.1	2.0	2.6	2.0
	3	CPII-HK (logs)	0.82	37.66	67.89	2.0	2.0	2.8	–
	4	CUNI-Pocket	0.76	34.79	62.21	2.0	1.9	2.9	–
	4	pinch-ast (logs)	0.75	38.29	67.07	**1.8**	**1.7**	**2.6**	**1.8**

Table 37: Results for English→Italian. Pl. = place within latency regime; shared with multiple systems if the scores are deemed not significantly different by a paired t-test with bootstrap resampling ($p > 0.05$; Koehn, 2004). Bold = best in the latency regime. Latency (in seconds): compute-unaware value [compute-aware value in brackets]. We do not report compute-aware latency for systems which were submitted as logs.

B.8 Indic S2S

B.9 African/Celtic S2S

B.10 Cross-lingual Voice Cloning

B.11 Instruction Following

Name	Constrained	Primary	SHORT					
			ASR WER↓	*SQA* BERTScore↑	*SSUM* BERTScore↑	*ACHAP* WER↓	*ACHAP* CollarF1↑	*ACHAP* BERTScore↑
Phi4-Multimodal	✗	—	**0.069**	0.463	—	—	—	—
Qwen3-Omni	✗	—	0.197	0.325	—	—	—	—
NLE_PRIMARY	✓	✓	0.136	**0.531**	—	—	—	—
NLE_CONTRASTIVE	✗	✗	0.134	0.501	—	—	—	—
FBK_PRIMARY	✓	✓	0.123	0.507	—	—	—	—
FBK_CONTRASTIVE	✓	✗	0.145	0.505	—	—	—	—
KIT_PRIMARY	✗	✓	0.074	0.495	—	—	—	—
KIT_CONTRASTIVE	✗	✗	0.170	0.450	—	—	—	—
BSC_PRIMARY	✗	✓	0.134	0.425	—	—	—	—
BSC_CONTRASTIVE1	✗	✗	0.127	0.383	—	—	—	—
BSC_CONTRASTIVE2	✗	✗	0.127	**0.420**	—	—	—	—
Name	*Constrained*	*Primary*	**LONG**					
Phi4-Multimodal	✗	—	0.281	0.420	0.183	0.941	0.084	0.837
Qwen3-Omni	✗	—	0.158	0.327	0.187	0.208	**0.609**	**0.877**
FBK_PRIMARY	✓	✓	0.196	0.390	0.152	0.359	0.000 (0.183)†	0.804 (0.823)†
FBK_CONTRASTIVE1	✓	✗	0.126	0.377	0.156	0.200	0.000 (0.271)†	0.803 (0.842)†
FBK_CONTRASTIVE2	✓	✗	0.175	0.377	0.160	0.200	0.000 (0.271)†	0.803 (0.842)†
KIT_PRIMARY	✗	✓	0.269	0.412	0.212	0.311	0.436	0.869
KIT_CONTRASTIVE	✗	✗	**0.064**	0.385	**0.218**	**0.093**	0.583	**0.877**

† Values in parentheses are obtained under a relaxed Markdown-format evaluation.

Table 38: ***English Official Results.*** **Bold** indicates the best track-wise (SHORT and LONG) result per language direction, and underline indicates the overall best result among tracks.

Name	Constrained	Primary	SHORT							
			ST COMET↑	*SQA* BERTScore↑	*QE* Accuracy↑	*QE* Format↑	*SSUM* BERTScore↑	*ACHAP* COMET↑	*ACHAP* CollarF1↑	*ACHAP* BERTScore↑
Phi4-Multimodal	✗	—	0.802	0.363	0.658	0.927	—	—	—	—
Qwen3-Omni	✗	—	0.836	0.330	**0.858**	**1.000**	—	—	—	—
NLE_PRIMARY	✓	✓	0.765	0.470	0.786	0.997	—	—	—	—
NLE_CONTRASTIVE	✗	✗	0.749	0.462	0.333	0.005	—	—	—	—
FBK_PRIMARY	✓	✓	0.762	**0.477**	0.762	0.974	—	—	—	—
FBK_CONTRASTIVE	✓	✗	0.739	0.471	0.501	0.584	—	—	—	—
KIT_PRIMARY	✗	✓	**0.840**	0.466	0.000	0.000	—	—	—	—
KIT_CONTRASTIVE	✗	✗	0.830	0.423	0.705	1.000	—	—	—	—
BSC_PRIMARY	✗	✓	0.808	0.467	0.785	0.953	—	—	—	—
BSC_CONTRASTIVE1	✗	✗	0.798	0.425	0.810	0.997	—	—	—	—
BSC_CONTRASTIVE2	✗	✗	0.799	0.383	0.796	0.712	—	—	—	—
			LONG							
Phi4-Multimodal	✗	—	0.617	0.370	0.658	0.927	0.162	0.375	0.053	0.543
Qwen3-Omni	✗	—	0.408	0.247	**0.858**	**1.000**	0.092	0.381	0.062	0.664
FBK_PRIMARY	✓	✓	0.694	0.348	0.501	0.584	0.149	0.690	0.000 (0.195)†	0.584 (0.618)†
FBK_CONTRASTIVE1	✓	✗	0.723	0.350	0.501	0.584	0.153	0.647	0.000 (0.165)†	0.616 (0.650)†
FBK_CONTRASTIVE2	✓	✗	0.722	0.351	0.501	0.584	0.153	0.635	0.000 (0.136)†	0.616 (0.640)†
KIT_PRIMARY	✗	✓	0.733	**0.405**	0.000	0.000	**0.238**	0.740	0.500	**0.690**
KIT_CONTRASTIVE	✗	✗	**0.840**	0.308	0.705	1.000	0.208	**0.836**	**0.508**	**0.676**

† Values in parentheses are obtained under a relaxed Markdown-format evaluation.

Table 39: ***German Official Results.*** **Bold** indicates the best track-wise (SHORT and LONG) result per language direction, and underline indicates the overall best result among tracks.

Name	Constrained	Primary	SHORT ST COMET↑	SHORT SQA BERTScore↑	SHORT SSUM BERTScore↑	SHORT ACHAP COMET↑	SHORT ACHAP CollarF1↑	SHORT ACHAP BERTScore↑
Phi4-Multimodal	✗	—	0.772	0.414	—	—	—	—
Qwen3-Omni	✗	—	0.827	0.373	—	—	—	—
NLE_PRIMARY	✓	✓	0.763	0.456	—	—	—	—
NLE_CONTRASTIVE	✗	✗	0.733	0.514	—	—	—	—
FBK_PRIMARY	✓	✓	0.742	0.524	—	—	—	—
FBK_CONTRASTIVE	✓	✗	0.733	**0.527**	—	—	—	—
KIT_PRIMARY	✗	✓	**0.841**	0.519	—	—	—	—
KIT_CONTRASTIVE	✗	✗	0.830	0.449	—	—	—	—
BSC_PRIMARY	✗	✓	0.773	0.454	—	—	—	—
BSC_CONTRASTIVE1	✗	✗	0.787	0.395	—	—	—	—
BSC_CONTRASTIVE2	✗	✗	0.787	0.395	—	—	—	—
LONG								
Phi4-Multimodal	✗	—	0.562	0.404	0.175	0.413	0.148	0.668
Qwen3-Omni	✗	—	0.681	0.354	0.110	0.503	0.446	**0.716**
FBK_PRIMARY	✓	✓	0.707	0.385	0.185	0.723	0.000 (0.335)†	0.532 (0.597)†
FBK_CONTRASTIVE1	✓	✗	0.695	0.373	0.174	0.735	0.000 (0.315)†	0.528 (0.665)†
FBK_CONTRASTIVE2	✓	✗	0.702	0.376	0.171	0.719	0.000 (0.348)†	0.586 (0.660)†
KIT_PRIMARY	✗	✓	0.732	**0.439**	0.267	0.737	0.456	0.703
KIT_CONTRASTIVE	✗	✗	**0.841**	0.322	**0.269**	**0.842**	**0.489**	0.709

† Values in parentheses are obtained under a relaxed Markdown-format evaluation.

Table 40: ***Italian Official Results.*** **Bold** indicates the best track-wise (SHORT and LONG) result per language direction, and underline indicates the overall best result among tracks.

Name	Constrained	Primary	SHORT ST COMET↑	SHORT SQA BERTScore↑	SHORT QE Accuracy↑	SHORT QE Format↑	SHORT SSUM BERTScore↑	SHORT ACHAP COMET↑	SHORT ACHAP CollarF1↑	SHORT ACHAP BERTScore↑
Phi4-Multimodal	✗	—	0.809	0.443	0.917	0.085	—	—	—	—
Qwen3-Omni	✗	—	**0.857**	0.273	**0.957**	**1.000**	—	—	—	—
NLE_PRIMARY	✓	✓	0.794	0.487	0.894	1.000	—	—	—	—
NLE_CONTRASTIVE	✗	✗	0.755	0.466	0.500	0.014	—	—	—	—
FBK_PRIMARY	✓	✓	0.777	**0.520**	0.915	0.961	—	—	—	—
FBK_CONTRASTIVE	✓	✗	0.734	0.482	0.658	0.819	—	—	—	—
KIT_PRIMARY	✗	✓	0.852	0.456	0.000	0.000	—	—	—	—
KIT_CONTRASTIVE	✗	✗	0.845	0.471	0.739	0.993	—	—	—	—
BSC_PRIMARY	✗	✓	0.782	0.413	0.929	0.950	—	—	—	—
BSC_CONTRASTIVE1	✗	✗	0.750	0.398	0.872	0.663	—	—	—	—
LONG										
Phi4-Multimodal	✗	—	0.554	0.412	0.917	0.085	0.160	0.385	0.030	0.617
Qwen3-Omni	✗	—	0.820	0.246	**0.957**	**1.000**	0.021	0.290	0.000	0.571
FBK_PRIMARY	✓	✓	0.655	0.380	0.658	0.819	0.319	0.681	0.000 (0.073)†	0.496 (0.518)†
FBK_CONTRASTIVE1	✓	✗	0.685	0.373	0.658	0.819	0.326	0.683	0.000 (0.021)†	0.496 (0.506)†
FBK_CONTRASTIVE2	✓	✗	0.699	0.374	0.658	0.819	0.325	0.698	0.000 (0.071)†	0.496 (0.522)†
KIT_PRIMARY	✗	✓	0.789	**0.452**	1.000	0.004	**0.383**	**0.785**	**0.503**	**0.685**
KIT_CONTRASTIVE	✗	✗	**0.847**	0.360	0.739	0.993	0.378	0.451	0.103	**0.511**

† Values in parentheses are obtained under a relaxed Markdown-format evaluation.

Table 41: ***Chinese Official Results.*** **Bold** indicates the best track-wise (SHORT and LONG) result per language direction, and underline indicates the overall best result among tracks.

B.12 Metrics task

ACL/Human transcript	Segment	System	Segment EnDe	Segment EnZh	System EnDe	System EnZh
Zarzu and Zouhar (2026)						
MLP (speech)	25.8	80.2	24.8	26.8	82.4	78.0
MLP (text)	26.9 +0.2	80.0 +0.0	25.1 +0.1	28.7 +0.4	81.5 +0.0	78.5 +0.0
MLP (text+speech)	27.6 +0.0	81.0 +0.0	25.8 +0.0	29.4 -0.0	81.5 +0.0	80.6 +0.1
Phi4 (speech)	23.0	93.6	20.7	25.3	91.9	95.3
Phi4 (text)	34.8 +0.6	95.4 -0.9	34.0 +1.1	35.5 +0.1	98.6 +1.5	92.3 -3.4
Phi4 (text+speech) ◂	25.3 +0.1	96.5 +0.3	24.3 -0.4	26.3 +0.6	97.4 +3.2	95.7 -2.6
Dinh and Niehues (2025)						
BoostedProb (text) ◂	-1.1 +0.0	43.3 +0.0	1.6 +0.0	-3.7 +0.0	59.2 +0.0	27.4 +0.0
Gupta (2026)						
Lexilogic (text) ◂	44.3 +1.3	95.4 -1.4	44.8 +1.6	43.8 +1.0	96.2 +1.5	94.5 -4.4
Shah et al. (2026)						
TieCal (text) ◂	47.7 +1.2	92.7 -1.4	47.9 +1.3	47.6 +1.0	94.5 -1.4	91.0 -1.4
Krahn and Fosler-Lussier (2026)						
HydraQE (DA head) (speech)	45.1	97.2	44.7	45.6	98.5	95.9
HydraQE (MetricX head) (speech)	45.0	97.2	44.7	45.2	98.7	95.6
HydraQE (XComet head) (speech)	45.5	95.8	45.5	45.6	98.0	93.6
HydraQE (all heads avg) (speech)	44.8	96.9	44.6	45.1	98.7	95.2
HydraQE (primary) ◂ (speech)	44.7	97.3	44.4	45.0	98.6	96.1
Organizers						
BLASER (speech)	32.4	85.9	31.0	33.9	83.9	87.8
COMETkiwi (text)	45.1 +1.2	92.7 -1.4	45.2 +1.4	45.0 +1.0	94.4 -1.5	91.1 -1.4
COMETpartial (text)	24.1 -0.1	81.3 -0.0	24.7 +0.1	23.6 -0.3	92.9 +0.1	69.8 -0.2
SpeechCOMET (text+audio)	43.0 +1.2	92.3 +0.6	42.3 +1.7	43.7 +0.7	88.9 +0.4	95.6 +0.9
SpeechLLM FT (text+audio)	42.1 +0.5	95.7 +0.6	36.1 +0.2	48.2 +0.8	98.5 +1.7	93.0 -0.5
SpeechQE (speech)	36.8	91.3	38.0	35.5	91.6	91.1
Annotators						
Random human annotator	53.3 +1.0	98.8 +0.1	51.6 +1.8	55.0 +0.3	98.1 -0.3	99.6 +0.4

Table 42: Results on the ACL testset using human transcripts (Kendall's $\tau \times 100$). Values in small text indicate the difference over Whisper-based transcript scores. The ◂ denotes the primary submission from a team.

ACL/Whisper transcript	Segment	System	Segment EnDe	Segment EnZh	System EnDe	System EnZh
Zarzu and Zouhar (2026)						
MLP (speech)	12.8	100.0	14.8	10.8	100.0	100.0
MLP (text)	13.4	100.0	14.4	12.5	100.0	100.0
MLP (text+speech)	16.6	100.0	18.3	15.0	100.0	100.0
Phi4 (speech)	18.1	100.0	18.5	17.7	100.0	100.0
Phi4 (text)	18.5	100.0	17.5	19.5	100.0	100.0
Phi4 (text+speech) ◂	15.8	100.0	12.8	18.8	100.0	100.0
Dinh and Niehues (2025)						
BoostedProb (text) ◂	-1.9	0.0	0.6	-4.3	0.0	0.0
Gupta (2026)						
Lexilogic (text) ◂	27.7	100.0	27.9	27.6	100.0	100.0
Shah et al. (2026)						
TieCal (text) ◂	28.2	100.0	26.8	29.6	100.0	100.0
Krahn and Fosler-Lussier (2026)						
HydraQE (DA head) (speech)	26.3	100.0	26.5	26.0	100.0	100.0
HydraQE (MetricX head) (speech)	27.0	100.0	28.1	25.8	100.0	100.0
HydraQE (XComet head) (speech)	25.9	100.0	27.6	24.2	100.0	100.0
HydraQE (all heads avg) (speech)	26.4	100.0	26.9	26.0	100.0	100.0
HydraQE (primary) ◂ (speech)	26.3	100.0	26.4	26.1	100.0	100.0
Organizers						
BLASER (speech)	19.1	100.0	21.0	17.3	100.0	100.0
COMETkiwi (text)	27.1	100.0	26.7	27.6	100.0	100.0
COMETpartial (text)	14.3	100.0	15.2	13.3	100.0	100.0
SpeechCOMET (text+audio)	23.7	100.0	21.4	26.0	100.0	100.0
SpeechLLM FT (text+audio)	16.9	100.0	14.9	19.0	100.0	100.0
SpeechQE (speech)	20.9	100.0	23.9	17.9	100.0	100.0
Annotators						
Random human annotator	44.6	100.0	43.7	45.6	100.0	100.0

Table 43: Results on the ACL testset using Whisper transcripts (Kendall's $\tau \times 100$, all values $\times 100$). The ◂ denotes the primary submission from a team.

AppTek Call Center	Segment	System	Segment EnDe	Segment EnZh	System EnDe	System EnZh
Zarzu and Zouhar (2026)						
MLP (speech)	9.9	97.6	7.5	12.4	95.3	100.0
MLP (text)	13.7	100.0	13.0	14.4	100.0	100.0
MLP (text+speech)	13.3	100.0	11.2	15.3	100.0	100.0
Phi4 (speech)	9.1	100.0	8.4	9.8	100.0	100.0
Phi4 (text)	13.7	100.0	12.9	14.6	100.0	100.0
Phi4 (text+speech) ◂	10.8	100.0	11.0	10.7	100.0	100.0
Dinh and Niehues (2025)						
BoostedProb (text) ◂	11.1	30.7	12.2	10.1	0.2	61.2
Gupta (2026)						
Lexilogic (text) ◂	19.0	100.0	16.6	21.4	99.9	100.0
Shah et al. (2026)						
TieCal (text) ◂	20.9	82.7	19.2	22.5	65.3	100.0
Krahn and Fosler-Lussier (2026)						
HydraQE (DA head) (speech)	24.1	100.0	22.6	25.6	100.0	100.0
HydraQE (MetricX head) (speech)	22.7	100.0	20.2	25.2	100.0	100.0
HydraQE (XComet head) (speech)	24.3	100.0	22.9	25.7	100.0	100.0
HydraQE (all heads avg) (speech)	23.7	100.0	22.1	25.4	100.0	100.0
HydraQE (primary) ◂ (speech)	24.1	100.0	22.4	25.7	100.0	100.0
Organizers						
BLASER (speech)	17.7	100.0	16.7	18.7	100.0	100.0
COMETkiwi (text)	20.7	84.0	17.8	23.6	68.0	100.0
COMETpartial (text)	8.0	0.5	9.1	6.9	0.0	0.9
SpeechCOMET (text+audio)	17.1	100.0	17.4	16.7	100.0	100.0
SpeechLLM FT (text+audio)	17.9	100.0	15.8	20.0	100.0	100.0
SpeechQE (speech)	16.8	0.0	19.7	13.9	0.0	0.0
Annotators						
Random human annotator	42.4	100.0	40.8	44.1	100.0	100.0

Table 44: Results on the AppTek Call Center testset (all values ×100). The ◂ denotes the primary submission from a team.

ITV (TV Series)	Segment	System	Segment EnDe	Segment EnZh	System EnDe	System EnZh
Zarzu and Zouhar (2026)						
MLP (speech)	9.2	47.9	6.2	12.1	95.9	0.0
MLP (text)	11.4	50.0	11.2	11.6	100.0	0.0
MLP (text+speech)	12.5	56.9	13.2	11.8	100.0	13.7
Phi4 (speech)	16.5	100.0	12.6	20.4	100.0	100.0
Phi4 (text)	19.7	100.0	20.4	19.1	100.0	100.0
Phi4 (text+speech) ◂	15.5	100.0	16.4	14.6	100.0	100.0
Dinh and Niehues (2025)						
BoostedProb (text) ◂	18.2	55.2	17.6	18.8	10.4	100.0
Gupta (2026)						
Lexilogic (text) ◂	30.2	100.0	26.1	34.3	100.0	100.0
Shah et al. (2026)						
TieCal (text) ◂	31.4	100.0	29.3	33.5	100.0	100.0
Krahn and Fosler-Lussier (2026)						
HydraQE (DA head) (speech)	29.4	100.0	26.5	32.3	100.0	100.0
HydraQE (MetricX head) (speech)	29.0	100.0	27.4	30.6	100.0	100.0
HydraQE (XComet head) (speech)	28.6	100.0	26.7	30.5	100.0	100.0
HydraQE (all heads avg) (speech)	29.3	100.0	26.4	32.1	100.0	100.0
HydraQE (primary) ◂ (speech)	29.0	100.0	25.9	32.0	100.0	100.0
Organizers						
BLASER (speech)	15.6	79.0	19.7	11.4	58.0	100.0
COMETkiwi (text)	31.7	100.0	28.0	35.5	100.0	100.0
COMETpartial (text)	12.4	100.0	12.0	12.7	100.0	100.0
SpeechCOMET (text+audio)	27.1	100.0	25.5	28.8	100.0	100.0
SpeechLLM FT (text+audio)	24.1	100.0	22.6	25.5	100.0	100.0
SpeechQE (speech)	23.5	100.0	23.1	24.0	100.0	100.0
Annotators						
Random human annotator	74.9	100.0	100.0	49.7	100.0	100.0

Table 45: Results on the ITV TV Series testset (all values ×100). The ◂ denotes the primary submission from a team.