

# Findings of the IWSLT 2025 Evaluation Campaign

Idris Abdulmumin<sup>37</sup> Victor Agostinelli<sup>35</sup> Tanel Alumäe<sup>5</sup> Antonios Anastasopoulos<sup>1</sup>  
Luisa Bentivogli<sup>3</sup> Ondřej Bojar<sup>4</sup> Claudia Borg<sup>7</sup> Fethi Bougares<sup>6</sup> Roldano Cattoni<sup>3</sup>  
Mauro Cettolo<sup>3</sup> Lizhong Chen<sup>35</sup> William Chen<sup>9</sup> Raj Dabre<sup>33</sup> Yannick Estève<sup>16</sup>  
Marcello Federico<sup>12</sup> Mark Fishel<sup>39</sup> Marco Gaido<sup>3</sup> Dávid Javorský<sup>4</sup> Marek Kasztelnik<sup>36</sup>  
Fortuné Kponou<sup>16</sup> Mateusz Krubiński<sup>38</sup> Tsz Kin Lam<sup>18</sup> Danni Liu<sup>23</sup> Evgeny Matusov<sup>21</sup>  
Chandresh Kumar Maurya<sup>32</sup> John P. McCrae<sup>22</sup> Salima Mdhaffar<sup>16</sup> Yasmin Moslem<sup>31</sup>  
Kenton Murray<sup>15</sup> Satoshi Nakamura<sup>20</sup> Matteo Negri<sup>3</sup> Jan Niehues<sup>23</sup> Atul Kr. Ojha<sup>22</sup>  
John E. Ortega<sup>24</sup> Sara Papi<sup>3</sup> Pavel Pecina<sup>4</sup> Peter Polák<sup>4</sup> Piotr Połec<sup>36</sup> Ashwin Sankar<sup>33</sup>  
Beatrice Savoldi<sup>3</sup> Nivedita Sethiya<sup>32</sup> Claytone Sikasote<sup>26</sup> Matthias Sperber<sup>27</sup>  
Sebastian Stüker<sup>28</sup> Katsuhito Sudoh<sup>34</sup> Brian Thompson Marco Turchi<sup>28</sup>  
Alex Waibel<sup>9</sup> Patrick Wilken<sup>21</sup> Rodolfo Zevallos<sup>29</sup> Vilém Zouhar<sup>30</sup> Maike Züfle<sup>23</sup>

<sup>1</sup>GMU <sup>3</sup>FBK <sup>4</sup>Charles U. <sup>5</sup>TalTech <sup>6</sup>Elyadata <sup>7</sup>U. Malta <sup>9</sup>CMU <sup>12</sup>Amazon  
<sup>15</sup>JHU <sup>16</sup>Avignon U. <sup>18</sup>U. Edinburgh <sup>20</sup>CUHK Shenzhen <sup>21</sup>AppTek <sup>22</sup>U. Galway  
<sup>23</sup>KIT <sup>24</sup>Northeastern U. <sup>26</sup>U. Cape Town <sup>27</sup>Apple <sup>28</sup>Zoom <sup>29</sup>U. Pompeu Fabra  
<sup>30</sup>ETH Zurich <sup>31</sup>ADAPT Centre <sup>32</sup>IIT Indore <sup>33</sup>IIT Madras <sup>34</sup>Nara Women's U.  
<sup>35</sup>Oregon State U. <sup>36</sup>ACC Cyfronet AGH <sup>37</sup>U. Pretoria <sup>38</sup>Snowflake <sup>39</sup>U. Tartu

## Abstract

This paper presents the outcomes of the shared tasks conducted at the 22nd International Workshop on Spoken Language Translation (IWSLT). The workshop addressed seven critical challenges in spoken language translation: simultaneous and offline translation, automatic subtitling and dubbing, model compression, speech-to-speech translation, dialect and low-resource speech translation, and Indic languages. The shared tasks garnered significant participation, with 32 teams submitting their runs. The field's growing importance is reflected in the increasing diversity of shared task organizers and contributors to this overview paper, representing a balanced mix of industrial and academic institutions. This broad participation demonstrates the rising prominence of spoken language translation in both research and practical applications.

## 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 22-year tradition, the 2025 conference was preceded by a comprehensive evaluation campaign designed to address critical scientific challenges in SLT. This paper presents the outcomes of the 2025 IWSLT Evaluation Campaign, which comprised seven distinct shared tasks organized into three primary research areas:

### • High-resource ST

- **Offline track**, with focus on speech-to-text translation of recorded scientific presentations, TV series, and business news from English to German, Arabic and Chinese.
- **Simultaneous track**, focusing on speech-to-text translation of streamed audio of conferences and interviews from English to German, Japanese and Chinese, and from Czech to English.
- **Subtitling track**, with focus on speech-to-subtitle translation of audio-visual documents from English to German and Spanish and on compression of pre-generated German and Spanish subtitles.
- **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 mul-

tilingual speech-to-text foundation model.

- **Low resource ST**

- **Low-resource SLT**, focusing on the translation of recorded speech from North Levantine Arabic to English, Tunisian Arabic to English, Bemba to English, Fongbe to French, Irish to English, Bhojpuri to Hindi, Estonian to English, Maltese to English, Marathi to Hindi, and Quechua to Spanish. It also included a data track, inviting participants to submit newly collected speech translation datasets of under-resourced language pairs.

- **Indic Languages Track** focuses on English and multiple Indic languages. The speech translations are from English speech to Indic language text and from Indic speech to English language text. Indic languages include Bengali, Hindi, and Tamil.

- **Instruction-following Speech Processing**

- **Speech Recognition, Translation, Question Answering, and Summarization**, with focus on Scientific talk audios from English to German, Italian, and Chinese languages.

The shared tasks drew participation from 32 diverse teams (detailed in Table 1), encompassing both academic institutions and industry leaders. In the following sections, we provide comprehensive coverage of each shared task, including detailed descriptions of the research challenges, specifications of training and testing datasets, evaluation methodologies, and submission analyses. Each task discussion concludes with a thorough results summary, with additional detailed performance metrics available in the corresponding appendices. This structure ensures a systematic presentation of the tasks while maintaining accessibility to both high-level findings and granular technical details.

## 2 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 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.<sup>1</sup> For the IWSLT-2025 Evaluation Campaign, a dedicated instance—SPEECHM-IWSLT2025<sup>2</sup>—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 in order to verify the soundness and completeness of the results. We include most tasks and test sets for human evaluation. We follow Sperber et al. (2024)’s approach to handle the automatically segmented long-form speech in a robust manner. Details are provided in Section A.

## 3 Offline track

The Offline Speech Translation task at IWSLT, a cornerstone of the conference’s tradition, aims to establish a robust evaluation framework for monitoring advancements in spoken language translation. Its core focus lies in unconstrained speech translation, distinguishing it from tasks with inherent temporal and structural limitations such as simultaneous translation or subtitling. While maintaining a consistent task formulation, the emphasis over time has incrementally shifted towards increasing task difficulty to better align with real-world demands, encompassing the translation of

<sup>1</sup>[www.meetween.eu](http://www.meetween.eu)

<sup>2</sup>[iwslt2025.speechm.cloud.cyfronet.pl](http://iwslt2025.speechm.cloud.cyfronet.pl)

Team	Organization	Tracks	Reference
AIB-MARCO			
ALADAN	ALADAN		Kheder et al. (2025)
AppTEK	AppTek		Petrick et al. (2025)
BUINUS	University of Indonesia and Bina Nusantara University		Tjitarianata et al. (2025)
CDAC-SVNIT	Center for Development of Advance Computing & Sardar Vallabhbhai National Institute of Technology		Roy et al. (2025)
CMU	Carnegie Mellon University		Ouyang et al. (2025)
CUNI	Charles University		Macháček and Polák (2025)
CUNI-NL	Charles University		Luu and Bojar (2025)
FFSTC-2			Kponou et al. (2025b)
GMU	George Mason University		Meng and Anastasopoulos (2025)
HITSZ	Harbin Institute of Technology, Shenzhen		Wei et al. (2025)
IIITH-BUT	International Institute of Information Technology Hyderabad (IIITH) and Brno University of Technology (BUT)		Akkiraju et al. (2025)
IITM	SPRING Lab, IIT Madras		Sarkar et al. (2025)
IST	Instituto Superior Tecnico		Attanasio et al. (2025)
JHU	Johns Hopkins University		Robinson et al. (2025)
JU	Jadavpur University		Das et al. (2025)
JU-CSE-NLP	Jadavpur University		Dhar et al. (2025)
KIT	Karlsruhe Institute of Technology		Koneru et al. (2025); Li et al. (2025)
KREASOF-TCD	Kreasof AI, Trinity College Dublin, and African Institute for Mathematical Sciences		Farouq et al. (2025)
KUVOST			Mohammadamini et al. (2025)
LIA	University of Avignon		Chellaf et al. (2025)
MBZAI	Mohamed bin Zayed University of Artificial Intelligence		
MEETWEEN	MeetWeen		
NAIST	Nara Institute of Science and Technology		Widiaputri et al. (2025); Tan et al. (2025)
NLE	NAVER LABS Europe		Lee et al. (2025)
NYA	Netease YiDun AI Lab		Wang et al. (2025)
OSU	Oregon State University		Raffel et al. (2025)
QUESPA	QUESPA		Ortega et al. (2025)
SYSTRAN	company for translation technology		Avila and Crego (2025)
TCD	Trinity College Dublin		Moslem (2025)
UPV	Universitat Politècnica de València		Sanchez et al. (2025)
URDU			Mehmood and Rauf (2025)

Table 1: List of participants to the IWSLT 2025 shared tasks ( Offline track; Simultaneous track; Subtitle track; Compression track; Low-resource track; Indic track; Instruction-following track

new and diverse languages, domains, and speaking styles.

This section provides an overview of this year’s task characteristics, along with a summary of the participating systems and their respective results.

### 3.1 Challenge

In line with the track’s emphasis on the challenges posed by diverse and increasingly complex evaluation scenarios, this year’s round focused on incorporating a new language, Arabic, into an evaluation setting designed to capture the complexity of real-world speech. This scenario encompassed diverse language settings (English → Arabic/Chinese/German) and domains (scientific presentations, TV series, and business news), alongside varied speaking styles and challenging recording conditions (e.g., single speakers, multiple overlapping speakers, background noise, and accent data).

Within this framework, participants were tasked with developing their system(s) for any of the three language combinations, selecting one from three distinct training data conditions (i.e., constrained, constrained with large language models, unconstrained), which differed in terms of allowed training resources. Consistent with previous rounds, the task welcomed participation with both cascade and end-to-end models, the latter being defined as solutions that eschew intermediate discrete representations (e.g., source language transcripts), instead employing joint training of all parameters and components used during decoding. Multiple submissions to the “SPEECHM” centralized evaluation server<sup>3</sup> were permitted, with the requirement of designating one as the *primary* submission and any others as *contrastive*.

<sup>3</sup>[iwslt2025.speechm.cloud.cyfronet.pl](https://iwslt2025.speechm.cloud.cyfronet.pl)

### 3.2 Data and Metrics

**Test Data** Also this year, participants were provided with test data representative of diverse domains and conditions, namely:

- **Scientific Presentations** – This dataset, derived from the Instruction Following task (Section 9), comprises 21 recordings, each lasting approximately 5.5 minutes, featuring transcripts of scientific oral presentations and their corresponding translations into several languages. The talks encompass a variety of technical content delivered by speakers from around the world.
- **TV Series** from ITV Studios<sup>4</sup> – This dataset includes 3 recordings, each approximately 40 minutes in length, featuring multiple individuals interacting in various scenarios. The speech translation system needs to handle challenges such as overlapping speakers, different accents, and background noise.
- **Business News** from Asharq Business with Bloomberg<sup>5</sup> – This dataset comprises two recordings, each approximately 2.5 hours in duration, and specifically focuses on the economy domain. The content is derived from a TV channel and distributed through various digital and social media platforms.
- **Accented English Conversations** sampled from the Edinburgh International Accents of English Corpus (EdAcc, Sanabria et al., 2023) – This dataset provides approximately 3.5 hours of conversations, each featuring two friends interacting on daily topics such like hobbies and vacation. The speakers were selected to cover a wide range of English accents from around the globe. In addition to the variety of accents (33 in total), another major challenge presented is the presence of spontaneous speech.

Contingent on data availability, each language direction was evaluated across distinct scenarios, specifically:

- English → German: TV series, scientific presentations, business news, and accent challenge.
- English → Arabic: business news.
- English → Chinese: scientific presentations.

Continuing the practice of previous years, the test sets were either entirely or partially shared with other tasks. This included the subtitling track (for TV series and business news data), the simultaneous, instruction-following, and model compres-

sion tracks (for scientific presentations). This collaborative approach significantly fosters broader integration and comparability across the various components of the evaluation campaign.

**Training and Development Data** As in the last two rounds of the challenge, participants were offered the possibility to submit systems built under three training data conditions:

1. **Constrained:** In this condition, permitted training data is limited to a medium-sized framework to ensure manageable training time and resource requirements. The comprehensive list<sup>6</sup> of allowed training resources (speech, speech-to-text-parallel, text-parallel, text-monolingual) explicitly excludes any pre-trained language models.
2. **Constrained with large language models** (constrained<sup>+LLM</sup>): This condition allows all training data permitted in the constrained setup, with the addition of any other LLMs, provided they are freely accessible and released under a permissive license. This setup aims to enable participants to leverage accessible LLMs in a standardized evaluation scenario.
3. **Unconstrained:** Under this condition, any resource, including pre-trained language models, may be utilized, with the sole exception of the evaluation sets. This setup is designed to allow the participation of teams equipped with high computational power and capable of developing effective solutions leveraging additional in-house resources.

Development data were supplied only for English-German and English-Chinese. For English-German, they comprise the development set from IWSLT 2010, along with the test sets released for the 2010, 2013-2015, and 2018-2022 IWSLT campaigns. For English-Chinese, they consist of the test set used for the 2022 round.

**Evaluation Metrics** Systems were evaluated based on their ability to produce translations similar to the target-language references. This similarity was quantified using multiple automatic metrics: COMET<sup>7</sup> (Rei et al., 2020), BLEU<sup>8</sup> (Papineni et al., 2002), BLEURT (Sellam et al., 2020), Char-

<sup>4</sup>[www.itvstudios.com](http://www.itvstudios.com)

<sup>5</sup>[asharqbusiness.com](http://asharqbusiness.com)

<sup>6</sup>See the IWSLT 2025 offline track: [iwslt.org/2025/offline](https://iwslt.org/2025/offline)

<sup>7</sup>[huggingface.co/Unbabel/wmt22-comet-da](https://huggingface.co/Unbabel/wmt22-comet-da)

<sup>8</sup>BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.14

acTER (Wang et al., 2016), chrF<sup>9</sup> (Popović, 2015), and TER<sup>10</sup> (Snover et al., 2006). COMET was again chosen as the primary evaluation metric this year. It was calculated on the test set using automatic resegmentation of the hypothesis based on the reference translation by mwerSegmenter,<sup>11</sup> employing a detailed script made accessible to participants.<sup>12</sup> To enhance the soundness and completeness of the evaluation, human assessment was also conducted on the best-performing submission from each participant.

### 3.3 Submissions

This year, 7 teams participated in the offline task, submitting a total of 30 runs through the “SPEECHM” evaluation server. Table 2 provides a breakdown of the participation in each sub-task showing, for each training data condition, the number of participants, the number of submitted runs and, for each training data condition (constrained, constrained<sup>+LLM</sup>, unconstrained), the number of submitted runs. Below, we provide a short description of the systems, whose creators submitted a system description paper.

CUNI-NL (Luu and Bojar, 2025) participated with an end-to-end en-de system trained under the “constrained with Large Language Models” condition. The model consists of an audio encoder that transforms the input audio into embeddings that are then passed to the LLM, which generates the output texts (transcript or translation). Both a length adapter and a modality adapter are added to facilitate the integration of the audio embeddings into the LLM. Two speech encoders (Seamless-v2-large and Whisper-v3-large) and three LLMs (Llama3 8B Instruct, EuroLLM 9B Instruct, and gemma3 12B Instruct) have been tested. To enhance the performance, multitask training was performed, teaching the model to transcribe, translate, and simultaneously transcribe and translate. The training data are limited to the CoVoST2 dataset and a large multilingual corpus built from the Common Voice corpora.

KIT (Koneru et al., 2025) participated with a cascade en-de system trained under the “unconstrained” condition. The cascade model comprises several components. The segmenter aims to identify the optimal point at which to segment an audio file. Various techniques were tested, demonstrating that fixed-window chunking with a chunk size of 25 consistently yields the best performance. The second component is an ensemble of ASR systems trained under different conditions, which is used to transcribe the audio. The produced transcripts are then recombined by a task-adapted LLM based on Llama3 8 B. The final transcripts are translated using a version of Tower 7B enhanced for the en-de translation direction. A final component was introduced to post-edit the translations with an APE model based on Tower 13 B. All the data used to train each component has been previously cleaned and selected to obtain high-quality samples.

NAIST (Widiaputri et al., 2025) participated with end-to-end en-de, en-zh systems, where the version based on SALMONN technology was trained under the “unconstrained” condition, while the in-house version was trained under the “constrained with Large Language Models” condition. SALMONN is an end-to-end speech-to-text model that integrates Whisper large-v2 as the speech encoder, a fine-tuned BEATs encoder for non-speech audio and the Vicuna 13B LLM as the decoder. The two audio encoders and the LLM are connected via a window-level Q-Former module. The customised end-to-end version is based on the Whisper large-v3 encoder, a DeCo projector, and the Qwen2.5 LLM. The en-de models are fine-tuned using a combination of CoVoST and Europarl, while the en-zh models are fine-tuned only on CoVoST. Different prompts have been tested to maximise translation performance.

NYA (Wang et al., 2025) participated with cascade en-ar, en-de, en-zh systems trained under the “unconstrained” condition. The ASR is based on Whisper medium, while the MT system combines an NMT model based on the Transformer technology and an LLM model based on X-ALMA. The NMT model is enhanced by leveraging data augmentation with backwards and forward translations and domain adaptation via data filtering. The LLM model is obtained by fine-tuning X-ALMA on in-domain data and leveraging Low-

<sup>9</sup>nrefs:1+case:mixed+eff:yes+nc:6+nw:0+space:no+version:2.4.2

<sup>10</sup>nrefs:1+case:lc+tok:tercom+norm:no+punct:yes+asian:no+version:2.4.2

<sup>11</sup>[www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz](http://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz)

<sup>12</sup>[github.com/isl-mt/SLT.KIT/blob/master/scripts/evaluate/Eval.sh](https://github.com/isl-mt/SLT.KIT/blob/master/scripts/evaluate/Eval.sh)



English-German				
Participants	Runs	Constrained	Constrained <sup>+LLM</sup>	Unconstrained
6	16	0	4	12
English-Chinese				
Participants	Runs	Constrained	Constrained <sup>+LLM</sup>	Unconstrained
4	10	0	2	8
English-Arabic				
Participants	Runs	Constrained	Constrained <sup>+LLM</sup>	Unconstrained
2	4	0	0	4

Table 2: Breakdown of the participation in each sub-task (English→German, English→Chinese, English→Arabic) of the IWSLT offline ST track. For each language direction, we report the number of participants, the number of submitted runs and, for each training data condition (constrained, constrained<sup>+LLM</sup>, unconstrained), the number of submitted runs.

Rank Adaptation fine-tuning. The NMT n-best and the LLM list of candidates are merged and reranked using COMET-based MBR decoding. The MT training data are filtered using a semantic metric based on sBERT. The in-domain specific data are generated by crawling domain-specific videos and leveraging the existing bilingual subtitles. The audio is segmented using SHAS.

### 3.4 Results

#### 3.4.1 English to German

**Overall result** Table 21 shows the aggregated result of the systems participated in the four test sets. In terms of ranking based on automatic evaluation metrics, KIT is ranked 1st, followed by NYA and NeMo. These top-3 systems perform better than the others by a large margin, e.g., a 0.1 COMET score, and most of them are based on the cascaded architecture rather than end-to-end. Unlike last year, where the winning system is metric dependent, the ranking between the top-3 systems remains consistent across all six metrics.

Unlike the top-3 systems, NAIST (U) and CUNI-NL presents a scenario where the ranking is metric dependent. In particular, NAIST (U) performs better in both COMET and BLEURT (neural metrics) but worse in both BLEU and chrF (string-based metrics).

**Domains** This year, a new set of domains has been introduced for evaluation. The long-standing TED domain has been removed, whereas the accent (data) and the ITV (only the domain) remain. Similar to last year’s edition, we evaluated each submitted system on different domains.

In spite of having diverse set of domains, the top-3 systems (KIT, NYA and NeMo) perform consistently well. The much better numbers on the evaluation metrics indicate that both Scientific

Presentations and Business News domains are less challenging to translate than the accent and ITV domains. Although the top-3 systems perform similarly in both the accent and ITV domains, the remaining systems achieve far worse scores on the ITV domain, making ITV possibly the most challenging domain.

Furthermore, the ranking across domains is quite consistent meaning that a system performing good in one domain as performs good in the other domain. The only exception is AIB, which performs good on three domains, but has challenges in the ITV domain.

**Data conditions** On top of the above, we can also observe the improvement of translation quality by increase the training data size. In all the test domains, the top three systems are from the “unconstrained” conditions, whereas the “constrained LLMs” submissions are ranked the bottom, except in the ITV domain. Among all the participants, NAIST is the only team which submitted both “unconstrained” and “constrained with LLMs” conditions. Their “unconstrained” system outperforms the constrained condition substantially in all metrics, showing the importance of training data size despite using LLMs for the tasks. However, it is worth noting that the pretrained models and the architectures between the two conditions are quite dissimilar. Another noteworthy comparison is between the CUNI-NL and the NAIST (U). Despite being a “constrained with LLMs” submission, the CUNI-NL performs better in Business News, ITV and Scientific Presentations domains in almost all metrics. This smaller performance gap could be attributed to the choice of the pretrained models, which the CUNI-NL has substantially tested on.

### 3.4.2 English to Arabic

For the en-ar direction, we evaluate the submitted systems on the Business News domain. This is a newly added language pair this year, and there are 3 submissions that are all based on “unconstrained” conditions.

Table 22 summarizes the results. The NYA is ranked 1st, followed by the NeMo and the AIB. The ranking is consistent across all the evaluation metrics. Furthermore, the ranking is also consistent with the ranking in English to German.

### 3.4.3 English to Chinese

For the en-zh direction, we evaluate the submitted systems on the Scientific Presentations domain. Unlike last year, there are both cascaded and end-to-end submissions this year.

Table 23 summarizes the results. The NYA is ranked 1st in the COMET metric among the six submitted systems. In addition to COMET, it is also ranked 1st in both BLEU and character-TER. While it does not score the highest on chrF and BLEURT, it ranks second overall. The AIB takes the second place with performance similar to the NYA in most evaluation metrics, and it is even ranked 1st according to BLEURT.

Regarding the data condition, NAIST submitted both “unconstrained” and “constrained with LLMs” conditions. Similar to en-de language direction, their “unconstrained” system outperforms the “constrained with LLMs” system substantially in all metrics. Despite the stronger performance, possibly caused by the larger training data size, NAIST (U) and NAIST (C<sup>+</sup>) use different pre-trained components.

## 3.5 Human Evaluation

Similar to previous editions, each participant’s primary submission has been further assessed by professional translators. The details of the human evaluation and its results are described in A.

Examining the results, it is interesting to note that, in most cases, human evaluation confirms the ranking provided by automatic metrics, with only minor discrepancies. This is true for the English to Arabic direction, where NYA outperforms other models, and for the English to Chinese direction, where only the top position shifted in favour of NYA, leading to a better average DA score than AIB (despite automatic metrics showing minimal difference between the two submissions). The human evaluation also corroborates the findings

from the automatic metrics regarding the impact of data conditions: the models trained in the unconstrained data condition generally outperform those trained in the constrained condition.

For English to German, the results confirm the trends observed in other language directions for the TV series test set, with human evaluation validating the rankings generated by the automatic metrics. More variations are shown for the accent and scientific presentation test sets.<sup>13</sup> For the accent test set, KIT outperforms all other systems, achieving the best score. The most surprising results concern the AIB submission, which, despite a significant difference from the best model in terms of COMET score (5.4 points), is indistinguishable from KIT from the human evaluation standpoint. It is difficult to hypothesise a possible reason for this discrepancy due to the lack of a system description paper, but this confirms the need to test a model under different conditions and validate its results with human evaluation. The AIB submission also shows similar behaviour for the scientific presentation test set, where it is penalised by the automatic evaluation (fourth with a gap of 3.8 COMET scores from the top-ranked system), but rewarded by human evaluation.

The fact that some of the test sets are shared across different tasks gives us the possibility to present a single ranking including systems developed under different conditions and tasks. Examining Tables 14, 15 and 17 shows that the systems built for the offline task without any latency (simultaneous task), task-sharing (instruction task), and length (subtitling task) constraints attain the best performance, with a margin of more than 1 average DA score over the other submissions.

## 4 Simultaneous track

Simultaneous speech translation focuses on translating speech in real-time, in a manner similar to simultaneous interpreting. The system is designed to begin translating before the end of an utterance. This technology is particularly useful in scenarios such as international conferences, personal travel, or public emergency events.

The task included two tracks: cascaded and direct. Submissions to the cascaded track contain systems that produce intermediate text, i.e. the transcription of the source audio, that is imme-

<sup>13</sup>The Asharq News test set has not been human-evaluated due to budget constraints.

diately consumed by a simultaneous text-to-text agent. In contrast, direct, or end-to-end, systems avoid any intermediate text and directly generate target-language (text) translations from the source audio. Both tracks covered four language directions as in the previous year: English to German, English to Chinese, English to Japanese, and Czech to English.

## 4.1 Challenge

### 4.1.1 Changes from the last year

This year’s simultaneous translation task had two major changes compared to the last year:

**Long-form speech** We introduced a more realistic condition for simultaneous speech translation on unsegmented speech (Papi et al., 2025a). Participants had to develop streaming translation systems processing long-form speech.

**Large language models** Participants were allowed to use LLMs under the same conditions as *Constrained with large language models* in the Offline task described in Section 3.2.

### 4.1.2 Latency regimes

Two latency regimes, *low* and *high*, were introduced for each of the tracks to evaluate translation quality in different latency conditions.

**English-to-German and Czech-to-English** 0 to 2 seconds (low), 2 to 4 seconds (high)

**English-to-Chinese** 0 to 2.5 seconds (low), 2.5 to 4 seconds (high)

**English-to-Japanese** 0 to 3.5 seconds (low), 3.5 to 5 seconds (high)

### 4.1.3 Submission

Participants were allowed to submit no more than one system per track, language direction, and latency regime. The latency regime of a submission was determined by its results on the development set. This year, we allowed two submission options: *Docker image* and *System log* submissions. The latter option was easier for the participants because they did not need to wrap their systems into a deployable form. Systems of the Docker image submissions were executed by the organizers on the blind-test set in a controlled environment using a NVIDIA H200 GPU. An example implementation was provided using the SimulEval toolkit (Ma et al., 2020).

## 4.2 Data

To simplify the setting and allow participants to focus on the new modeling aspects of simultaneous translation, we adhere to the constraints with large language models as defined for the offline SLT task, see Section 3.2 above. This is the only data condition for the task. The test and dev sets differ across language pairs:

### English to German, Chinese, and Japanese

The test data are the speech translation section of the IWSLT25Instruct benchmark created for the Instruction Following task (Section 9) and derived from scientific talks (ACL Anthology presentations). The dev data are the ACL 60/60 benchmark (Salesky et al., 2023). In addition, we use *Accented English Conversations* test set for English to German.

split	domain	#utter.	#words/ utter.	duration (min)
dev	ParCzech	276	24	56
	ELITR	314	13	28.6
test	ParCzech	636	20.53	108.58
	Non-Native	1298	6.33	86.85

Table 3: Statistics of the dev and test sets for the Czech-English simultaneous task.

**Czech to English** The dev set was created from two sources:

- From ParCzech 3.0 (Kopp et al., 2021), we took a subset of the test recordings in the variant called “context”, which consists of parliamentary speeches in their original partitioning, preserving the natural flow of the speech.
- From the ELITR test set (Ansari et al., 2021),<sup>14</sup> we took an entire recording of a debate about AI.

The reference translations of the devset were done by students of translation studies from the Faculty of Arts at Charles University.

The test set was also collected from two sources:

- Selected recordings (complete speeches) of the Parliament of the Czech Republic, ensuring that there is no speaker overlap with the recordings allowed for training.
- Recordings of Czech language proficiency exams at the A2 level (Novák et al., 2024).

<sup>14</sup>[github.com/ELITR/elitr-testset/tree/master/documents/2021-theatre-related/robothon-debate](https://github.com/ELITR/elitr-testset/tree/master/documents/2021-theatre-related/robothon-debate)



The reference translations of the testset were provided by a professional translation agency. The statistics of both sets are provided in Table 3.

### 4.3 Evaluation

#### 4.3.1 Automatic Evaluation

We automatically evaluate two aspects of models: quality and latency.

**Quality** We conducted both automatic and human evaluation. BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022a) are used for automatic quality evaluation. The ranking of the submission is based on the BLEU score on the blind test set.

**Latency** We only conducted automatic evaluation using StreamLAAL (Papi et al., 2024).

#### 4.3.2 Human evaluation

For English-to-German and Czech-to-English, human evaluation was conducted using the Continuous Rating method proposed by Javorský et al. (2022). Further details on the method and score calculation are provided in Appendix A.2. This evaluation covered systems operating in the high-latency regime (with the exception of the CMU submission, which participated only in the low-latency regime).

For Czech-to-English, we additionally collected two independent human interpretations—one by a professional and one by a student interpreter—and evaluated them in the exact same manual evaluation style as system outputs, i.e. presenting them as gradually growing text in their authentic timing. The professional interpreter has been working full-time in the field since 2013, has been accredited by EU institutions since 2018, and regularly interprets for clients such as Czech Television, Czech Radio, CNN, and the World Bank. The student interpreter was a second-year master’s student at the Institute of Translation Studies, with three completed semesters of simultaneous interpreting training. The interpreting was carried out remotely, transcribed by WhisperX (Bain et al., 2023) and post-edited by an annotator fluent in English. For preparation of the sessions, both interpreters got a brief summary of each speech in three sentences using Llama 3.3 language model (Grattafiori et al., 2024). According to the professional interpreter, the interpretation differed from real-world conditions for three main reasons: (1) the absence of visual input, as the recordings were provided in audio-only format; (2) the absence of

a second interpreter, who would normally assist by noting down numbers and looking up specific terminology; and (3) limited preparation time, as the speeches covered a wide range of topics — unlike in real interpreting settings, where the subject matter is typically more stable.

For English-to-Japanese, another human evaluation was conducted by a professional interpreter using MQM-based metric (JTF, 2018) as in the last years.

Human evaluation using Direct Assessment was also conducted for comparison with other tasks, as described in A.1.

### 4.4 Submissions

Five teams in total participated this year, with three of those participating submissions containing testable systems for computationally-aware latency measurements. All teams entered the English-to-German track; four teams entered the English-to-Chinese, two teams entered the English-to-Japanese tracks; and one team entered the Czech-to-English track.

**BASELINES** were built for all directions. We use two approaches, a cascaded and a direct one. Both approaches used simultaneous segmenters to accommodate the long-form regime. We used fixed-length and VAD segmenters as described in Polák and Bojar (2024). For the cascaded system, we use Whisper-Large-V3-Turbo (Radford et al., 2023) for ASR and M2M100 (Fan et al., 2021) for MT. Both the Whisper and M2M100 models were onlinized using the Local Agreement policy (Polák et al., 2022, 2023). To make the ASR more robust to segmentation, we used the transcript of the previous segment as a context. For the direct approach, we selected SeamlessM4T (Seamless Communication et al., 2023) as the backbone of our system. We also used the Local Agreement policy for onlinizing the offline SeamlessM4T model.

CUNI (Macháček and Polák, 2025) participated in the direct track for English to German, Chinese, and Japanese, as well as Czech to English directions. They proposed two system architectures based on the language direction. For the from-English direction, their system is based on Whisper-Large-V3 (Radford et al., 2022) in the role of ASR and EuroLLM (Martins et al., 2024) as MT. The Whisper model was onlinized using the AlignAtt (Papi et al., 2023) policy,

while the EuroLLM model was onlineized using the Local Agreement policy (Polák et al., 2022, 2023). For the Czech-to-English direction, they used a direct approach, leveraging Whisper-Large-V3. They also explored improving translation quality by including previous translation as context and prompting for in-domain terminology.

CMU (Ouyang et al., 2025) participated in the direct track for the English to Chinese and German directions. Their system integrates a chunk-wise causal Wav2Vec2.0 speech encoder (Baevski et al., 2020), an adapter, and the Qwen2.5-7B-Instruct (Qwen et al., 2025) as the decoder. The training is conducted in two stages on speech segments curated from LibriSpeech (Panayotov et al., 2015), CommonVoice (Ardila et al., 2020b), and VoxPopuli (Wang et al., 2021) datasets, which are translated into Chinese and German with the 4-bit quantized Qwen2.5-32B-Instruct. The latency is controlled through a configurable latency multiplier, ensuring translations are generated after accumulating a predefined number of chunks, and the decoder uses a sliding window strategy to maintain the context through KV cache concatenation.

OSU (Raffel et al., 2025) participated in the cascaded track for the English-to-German and Chinese directions. Their system employs Whisper-Large-V2 (Radford et al., 2022) with a voice-activity-detection (VAD) segmenter (Siler Team, 2021) for ASR with a 4-bit quantized Gemma3-12B-Instruct (Team et al., 2025) and context-aware conversational prompting (Wang et al., 2024a) for translation. For fine-tuning, they re-purpose a prior framework (Agostinelli et al., 2024; Raffel et al., 2024) and its conversational prompting implementation alongside semantic similarity-based filtering to curate noisy subtitling data (Lison et al., 2018) before fine-tuning with LoRAs (Hu et al., 2021). In addition, this system augments basic conversational prompting for ST by leveraging a single-sentence sliding window memory bank for prior context.

UPV (Sanchez et al., 2025) participated in the cascaded track for the English-to-German direction. Their system employs Whisper-Large-V3-Turbo (Radford et al., 2022) with a modified longest-common-prefix (LCP) decoding policy for ASR alongside NLLB-3.3B (NLLB Team et al., 2022) with relaxed-agreement LCP (RALCP)

(Wang et al., 2024b) with a *wait-k* policy (Ma et al., 2019) for simultaneous translation. Additionally, this system features a similar, but simplified and more efficient, segmentation process to AlignAtt (Papi et al., 2023), leveraging attention maps to judge necessary model context. For training, they randomly prepended up to 10 sentences of prior context to a given sample so as to better leverage the unsegmented audio of this year’s task.

NAIST (Tan et al., 2025) participated in English-to-German, Chinese, and Japanese language directions of the direct track. Their system employs SHAS (Tsiamas et al., 2022) for speech segmentation, Whisper-large-v3 (Radford et al., 2022) for encoding input speech, DeCo (Yao et al., 2024) for projecting Whisper features into acoustic embeddings for the LLM, and Qwen-2.5-7B-Instruct (Qwen et al., 2025) LLM. It was fine-tuned with LoRA by joint training of ST and ASR, and the offline-trained ST system was used for simultaneous translation using Local Agreement (Liu et al., 2020; Polák et al., 2022).

## 4.5 Results

### 4.5.1 Automatic Evaluation

We rank the system performance based on BLEU scores. Cascaded systems are marked with an asterisk (\*). The detailed results can be found in the respective tables in Appendix B.3.

**Low-Latency** The ranking of systems for the the low-latency condition is as follows:

- English to German (Table 24):  
CMU, NAIST, OSU \*, BASELINES-Direct
- English to Chinese (Table 25):  
CMU, NAIST, OSU \*, BASELINES-Direct
- English to Japanese (Table 26):  
NAIST, BASELINES-Direct
- Native Czech to English (Table 27):  
CUNI, BASELINES-Direct
- Non-native Czech to English (Table 28):  
CUNI, BASELINES-Direct
- Accented English to German (Table 29):  
OSU \*, NAIST, CMU, BASELINES-Direct

**High-Latency** The ranking of systems for the high-latency condition is as follows:

- English to German (Table 24):  
CUNI \*, UPV \*, OSU \*, BASELINES-Casc.\*,  
NAIST, BASELINES-Direct
- English to Chinese (Table 25):  
CUNI \*, NAIST, OSU \*, BASELINES-Direct

- English to Japanese (Table 26):  
CUNI, NAIST, BASELINES-Direct
- Native Czech to English (Table 27):  
BASELINES-Direct, CUNI, BASELINES-Casc.\*
- Non-native Czech to English (Table 28):  
CUNI, BASELINES-Casc.\*, BASELINES-Direct
- Accented English to German (Table 29):  
OSU \*, UPV \*, BASELINES-Casc. \*, NAIST,  
CUNI \*, BASELINES-Direct

#### 4.5.2 Human Evaluation

Details of the human evaluation are provided in Section A.2 of the Appendix and results are shown in Table 18 for Czech-to-English, in Table 19 for English-to-German, and in Table 20 for English-to-Japanese. For Czech-to-English and English-to-German, we selected only one baseline that has a higher BLEU score.

#### 4.6 Conclusions

This year’s simultaneous translation shared task marks a significant shift in the focus of simultaneous translation system evaluations. With the introduction of unsegmented source audio in the test-set, participants are incentivized to address critical opportunities and challenges in real applications that have largely been avoided in prior years at the IWSLT. Unlike last year, submissions for this year’s shared task all feature large language models (LLMs), with the exception of the CUNI Czech-to-English submission, which were tailored for simultaneous translation in a variety of ways. Interestingly, a range of LLMs were represented in this year’s submissions. CUNI’s submission leveraged EuroLLM, a model built for translation across numerous languages, whereas other teams employed more general-purpose models.

On the IWSLT25Instruct test set, the CUNI submission outperformed almost all other systems at high-latency regimes, aside from on English-to-Chinese, where the NAIST submission produced a slightly higher BLEU score. At low-latency regimes, CMU produced the highest quality translations at comparatively low latency for English-to-German and English-to-Chinese. While the OSU and UPV submissions performed worse on the IWSLT25Instruct test set, they both performed significantly better on the challenging accented English-to-German test set, with the OSU system performing best at the cost of comparably high latency.

Human evaluation of the Czech-to-English lan-

guage pair shows that the quality of CUNI is comparable to that of the student interpreter but worse than that of the professional interpreter. However, the latency of CUNI is 1.51, approximately three times lower, i.e. faster than human interpreting.<sup>15</sup> BLEU scores for human interpretations are very low, which is expected, as interpreting often involves paraphrasing, summarization, and explanation. While both latency and BLEU favor CUNI, the professional interpreting still delivers the highest overall quality and in the shortest time, by going beyond the literal translation and conveying information in a more comprehensive way.

Human evaluation for English-to-German and English-to-Japanese aligns well with the results of automatic evaluation. Neural network-based evaluations are similarly aligned with automatic evaluations, yielding no major surprises.

Regarding promising directions for investigations and improvements to the shared task, the accented and non-native test sets emerged as the most difficult for current systems, and more studies on these scenarios could drive simultaneous translation models to be more robust. Moreover, enhancing the task accessibility—such as allowing log-based submissions as this year—can encourage broader participation. However, this comes at the cost of losing compatibility in computationally-aware latency metrics, which are crucial for simultaneous translation systems. Striking a balance between accessibility and fair evaluation will be key to enabling more meaningful progress in future editions.

### 5 Subtitling track

In recent years, the task of automatically creating subtitles for audiovisual content in another language has gained a lot of 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** was introduced for the first time in 2023 (Agarwal et al., 2023) and proposed again in 2024 (Ahmad et al., 2024) as part of the IWSLT Evaluation Campaigns.

The automatic subtitling task has been continued this year.<sup>16</sup> Participants were asked to gen-

<sup>15</sup>The latency is even lower than 2 seconds. The reason is that the systems were bucketed according to the latency on the devset, which for CUNI is 2.63.

<sup>16</sup>The **subtitle compression** sub-track, introduced in 2024, was proposed this year as well but we received no submis-

erate subtitles in German and/or Arabic from English speech in audiovisual documents.

## 5.1 Challenge

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 characters per line. Most audio-visual companies define their own subtitling guidelines, which can slightly differ from each other. In the case of IWSLT participants, we asked to generate subtitles according to specific guidelines provided by TED, including:

- The maximum subtitle reading speed is 21 characters per second;
- Lines cannot exceed 42 characters, including white spaces;
- Subtitles cannot exceed 2 lines.

Participants were expected to use only the audio track from the provided videos (dev and test sets), as the video track could be of low quality and primarily intended to check the temporal synchronicity and other aspects of displaying subtitles on screen.

The subtitling track required participants to automatically subtitle audio-visual documents in German and/or Arabic, where the spoken language is always English. The documents were collected from the following sources:

- TV series from **ITV Studios**;<sup>17</sup>
- Financial news content recordings from the **Asharq Business with Bloomberg** media group.<sup>18</sup>

## 5.2 Data and Metrics

**Data.** This track proposed two training data conditions:

- **Constrained:** the official training data condition, in which the allowed training data is limited to a medium-sized framework<sup>19</sup> to keep the training time and resource requirements manageable;
- **Unconstrained:** a setup without data restrictions (any resource, pre-trained language mod-

sion for it.

<sup>17</sup>[www.itvstudios.com](http://www.itvstudios.com)

<sup>18</sup>[asharqbusiness.com](http://asharqbusiness.com)

<sup>19</sup>[iwslt.org/2025/subtitling#training-and-data-conditions](https://iwslt.org/2025/subtitling#training-and-data-conditions)

els included, can be used) to allow also the participation of teams equipped with high computational power and effective in-house solutions built on additional resources.

For each language and domain, a development set and three test sets were released, those of previous evaluations (**tst2023** and **tst2024**), used for measuring progress over years, and a new one (**tst2025**). Table 4 provides some statistics on these sets.

domain	set	AV docs	hh:mm	#ref subtitles	
				de	ar
ITV	dev	7	06:01	4489	-
	tst23	7	05:08	4807	-
	tst24	7	05:54	4564	-
	tst25	3	02:07	1845	-
Asharq-Bloomberg	dev	2	03:01	3662	2974
	tst25	2	03:03	3543	2759

Table 4: 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**, primary metric, used also for ranking (Wilken et al., 2022);<sup>20</sup>
  - Translation quality vs. reference translations:
    - **BLEU**<sup>21</sup> and **CHR**<sup>22</sup> via sacreBLEU (Post, 2018);
    - **BLEURT** (Sellam et al., 2020).
- Automatic subtitles are realigned to the reference subtitles using mwerSegmenter (Matusov et al., 2005)<sup>23</sup> before running sacreBLEU and BLEURT.
- Subtitle compliance:<sup>24</sup>
    - rate of subtitles with more than 21 characters per second (**CPS**);
    - rate of lines longer than 42 characters, white-space included (**CPL**);
    - rate of subtitles with more than 2 lines (**LPB**).

<sup>20</sup>[github.com/apptek/SubER](https://github.com/apptek/SubER)

<sup>21</sup>sacreBLEU signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

<sup>22</sup>sacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0

<sup>23</sup>[www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz](http://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz)

<sup>24</sup>[github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech\\_to\\_text/scripts/subtitle\\_compliance.py](https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/subtitle_compliance.py)



### 5.3 Submissions

The subtitling track saw the participation of only one team: APPTEK (Petrick et al., 2025). Details about their systems follow.

**AppTek:** The APPTEK cascaded system includes the AppTek<sup>25</sup> production ASR and MT systems, adapted to the domains of this evaluation (ITV and Asharq-Bloomberg).

- **ITV:** In addition to other speech data from various domains, APPTEK’s hybrid ASR system was trained on entertainment data (audio and corresponding subtitles) provided by AppTek’s major media and entertainment localization customer. Similar data, in the form of professionally created English and German subtitle files, was used to adapt the English-to-German Transformer-based neural MT system.
- **Asharq-Bloomberg en-de:** The cascade of the AppTek’s general domain ASR system and an adapted English-German MT system was used. The MT model was adapted on a subset of parallel data selected from available public sources (like CCMatrix), based on semantic similarity with the Asharq-Bloomberg en-de parallel development data (clustering based on sentence embedding similarity).
- **Asharq-Bloomberg en-ar:** In this case too, the cascade consisted of the AppTek’s general domain ASR system and an adapted English-Arabic MT system. Here, the MT model was adapted on parallel data of human-curated (post-edited) Asharq-Bloomberg financial news programs. This data was available to AppTek as part of their cooperation with Asharq business with Bloomberg.

AppTek’s Intelligent Line Segmentation (ILS, proprietary technology) neural model was used in the source language after ASR to create subtitle blocks, timed mostly according to word boundaries but extended where possible for a comfortable (lower) reading speed. ILS was also used to segment the translated sentences into these blocks, optimizing line breaks for human acceptance and readability while, at the same time, respecting the subtitling constraints.

AppTek’s NMT systems support length control. For all primary submissions, whenever the default translation violated either the lines-per-block (LPB) limit or the characters-per-second (CPS)

limit, the source transcript was re-translated with a stricter length control parameter (e.g., “short”, “shorter”, “shortest”).

For the primary ITV submission, an increased reading speed limit of 23 CPS was chosen for a better translation quality/subtitle compliance trade-off. The Contrastive 1 submission is without MT length control, while the Contrastive 2 submission uses the default CPS value of 21. For Asharq-Bloomberg, the Contrastive 1 is without domain adaptation, en-ar Contrastive 2 is without length control, en-de Contrastive 2 differs in setting MT meta-data controls to genre “news” and style “formal”.

### 5.4 Results

#### 5.4.1 Automatic Evaluation

Scores on tst2025 of all APPTEK runs calculated using automatic metrics are shown in Tables 30 and 31. Tables 32 and 33 refer to progressive tst2024 and tst2023 sets, respectively, where the primary runs of 2024 and 2023 participants are reported as well to allow comparisons and quantification of progresses.

**tst2025 ITV en-de** (Table 30, ITV rows): Scores confirm the expectations based on the setups of the various runs. The primary run actually provides the best trade-off between translation quality and subtitle compliance using a smoothed setup of the length control mechanism: indeed, its BLEURT score lies between those of Contrastive 1 (for which translation quality was the priority, obtained by disabling the length control mechanisms) and Contrastive 2 (for which subtitle compliance was prioritized using the default setup of the length control mechanism). On the other side, the CPS of the primary run is better than that of Contrastive 1 but worse than Contrastive 2. The SubER value, being the best among all, confirms that the working point of the primary run optimizes the compromise between the two contrasting features.

**tst2025 Asharq-Bloomberg en-de** (Table 30, Asharq-Bloomberg rows): In the financial news domain, the length control configuration is common to all runs and so it is not surprising to observe CPS values that are close to each other. The MT model used to produce the Contrastive 1 submission was not domain-adapted, which caused the lowest BLEURT value. It is evident

<sup>25</sup>[www.apptek.com](http://www.apptek.com)



that the generation of translations according to the “news” genre and “formal” style (Contrastive 2) does not have effects that automatic metrics can capture.

**tst2025 Asharq-Bloomberg en-ar** (Table 31): In this case, the domain adaptation does not help too much for the primary run as compared to the use of the original generic MT model (Contrastive 1). The deactivation of the length control mechanism (Contrastive 2) allows to obtain the best translation quality at the expense of the lowest CPS.

**tst2024 ITV en-de** (Table 32): The results of APPTEK’s runs on the tst2024 essentially confirm the main outcome from the 2025 testset, i.e. that the length control mechanism allows to adjust the subtitle compliance at the expense of translation quality. The main difference observed between tst2025 and tst2024 results is that, for the latter, the best SubER—corresponding to the optimal trade-off between the two contrastive features—is obtained with the Contrastive 2 setup, not the primary one.

Concerning the comparison with the primary submission of the past edition, the improvement observed for the 2025 APPTEK system is impressive from all point of views, including translation quality, subtitle compliance, and trade-off between them. The only 2024 system that beats the primary AppTek 2025 submission is HW-TSC in terms of (only) BLEURT, but at the cost of significantly worse subtitle compliance values.

**tst2023 ITV en-de** (Table 33): The same considerations made on APPTEK’s runs for tst2024, in particular on the impact of the length control mechanism, also apply to tst2023.

The results on tst2023 also assess the progress among all participants of the current and past two editions of the subtitling track. As noted last year, the two best primary 2024 systems (APPTEK and HW-TSC) achieved SubER values similar to those of the two best 2023 systems (APPTEK and TLT), having generally better translation quality but worse subtitle compliance. This seemed to indicate that in 2024 more attention was paid to the quality of translation than to subtitle compliance. On the contrary, this year both aspects were taken into consideration, allowing to establish working points that are better than in the past from all perspectives.

Overall, the results discussed here demonstrate

a clear evolution in subtitling technology over the years. Despite limited participation, the task appears to have successfully met its objectives of fostering research in this area by providing a shared evaluation framework for sound comparisons across diverse and challenging settings, as well as enabling comparative analyses of progress on blind test sets from previous years.

#### 5.4.2 Human Evaluation

Human evaluation was also conducted for the subtitling task, with the aim of gaining a general and purely indicative understanding of the quality of the systems’ output in this challenging condition, as compared to systems developed under different conditions, including the much less restrictive ones of the offline task. A crucial premise in interpreting the results reported in Section A.1 is that these results stem from an evaluation setup that is inherently penalizing for subtitling systems. The scores shown in Tables 13 and 16 were obtained by asking human assessors to compare the systems’ outputs against verbatim translations, without access to the reference transcripts in the source language - a process that inevitably disadvantages the often shortened or condensed outputs produced by subtitling systems. That said, the results are not surprising. On the en-ar task (Table 13), the gap with the three competitive, unconstrained offline systems is substantial. On the en-de task (Table 16), the APPTEK system obtains rank 4 out of the 8 evaluated systems.

## 6 Model compression track

The Model Compression Track, introduced for the first time at IWSLT 2025, addresses a growing concern in the NLP community: how to reconcile the impressive capabilities of foundation models with the practical constraints of real-world deployment. As a matter of fact, while large-scale text and speech models have revolutionized tasks such as end-to-end speech-to-text translation, their substantial size and computational demands introduce significant challenges in resource-constrained settings—including mobile devices, embedded systems, and edge computing environments. This is particularly problematic when low-latency, on-device inference is required. Model compression offers a promising path forward, enabling reductions in model size and complexity while striving to minimize performance degradation as much as possible. By foregrounding this challenge, the

track aims to establish a shared evaluation framework for monitoring future advancements in the development of more efficient, accessible, and deployable SLT systems.

## 6.1 Challenge

This year’s objective was to assess participants’ ability to reduce the size of a large multilingual speech-to-text foundation model while minimizing performance degradation in English→German and English→Chinese speech translation settings. The chosen model, Qwen2-Audio (Chu et al., 2024), was selected due to its substantial yet manageable size (8.2 billion parameters, requiring approximately 16 GB of memory storage), its support for various speech processing task across multiple language directions, and its permissive Apache 2.0 license. Altogether, its computational cost, memory-intensive nature, and versatility make it an ideal candidate for task-oriented model compression.

Regarding compression techniques, admissible approaches were required to exclusively focus on modifying or optimizing the model’s internal parameters, ensuring that the final compressed model remained strictly derived from the original Qwen2-Audio. Therefore, eligible techniques included pruning (i.e. the removal of less important neurons and/or entire layers within the model, by identifying and eliminating parameters that contribute minimally to its output), quantization (i.e. the reduction of the numerical precision of the model’s weights—e.g., from 32-bit to 16-bit, 8-bit, or less—to lower its memory footprint), distillation (i.e. the creation of a smaller “student” model derived from Qwen2-Audio, for instance through pruning, trained to replicate the behavior of the original “teacher” model), as well as any other method that produces a compressed counterpart of the original model. Compression techniques could be applied either individually or in combination.

## 6.2 Data and Metrics

**Test Data** Participants were provided with test data representative of a specific domain, **scientific presentations**, which is shared across other tracks—specifically, the offline, simultaneous, and instruction-following tracks. This dataset (IWSLT25Instruct, fully described in Section 9) comprises 21 recordings, each approximately 5.5 minutes in length, featuring transcripts of scientific oral presentations along with their corre-

sponding translations from English into several languages (including German and Chinese).

**Training and Development Data** Participants were offered the possibility to submit systems developed under two distinct training data conditions, which differed in the datasets allowed to support the model compression process. Specifically, while the **unconstrained** condition imposed no restrictions on data usage, the **constrained** condition limited the permitted training data to the ACL60/60 dataset.<sup>26</sup> This dataset is identical in both size and source audio content for the two language directions involved in the task and, although small, is domain-consistent with the evaluation sets.

**Evaluation** As an initial step toward a comprehensive evaluation framework for benchmarking compression techniques that strike a balance between compactness and performance, this first round of the task focused on a subset of the relevant dimensions of the problem,<sup>27</sup> specifically addressing two interconnected challenges, each with its own evaluation criteria:

- **Model Reduction:** Reduce the size of the foundation model, defined by its number of parameters and memory usage, to improve suitability for deployment in resource-limited settings.
- **Translation Performance:** Preserve translation quality despite model size reduction, ensuring that the compressed models remain both practically valuable and reliable.

Focusing on these two dimensions, the evaluation protocol was designed to follow a two-step approach.

**STEP 1:** Categorization of the submitted models into five size bins based on their storage requirements (S),<sup>28</sup> representing increasingly aggressive levels of compression. The bins were defined as follows:

- Bin1:  $2\text{ GB} \leq S < 4\text{ GB}$
- Bin2:  $1\text{ GB} \leq S < 2\text{ GB}$
- Bin3:  $500\text{ MB} \leq S < 1\text{ GB}$

<sup>26</sup><https://aclanthology.org/attachments/2023.iwslt-1.2.dataset.zip>

<sup>27</sup>While computational efficiency (i.e., speed) is recognized as a critical factor for deploying models in resource-constrained environments, it was excluded from the evaluation framework in this initial round. However, we plan to adopt a phased evaluation strategy in future editions, with subsequent rounds incorporating computational efficiency and thereby broadening the overall evaluation scope.

<sup>28</sup>Self-reported by participants at the submission stage.

Model	Num. Params (↓)	Storage (↓)	en-de (↑)	en-zh (↑)
Qwen2-Audio-7B-Instruct	8.4B	16.8GB	0.672	0.743
TCD_constrained_primary	5.0B	9.7GB	0.764	0.806
TCD_unconstrained_contrastive	4.1B	8.8GB	0.693	-

Table 5: Results on the IWSLT25-Instruct ST test set in terms of translation quality (COMET-22 scores) and model size (expressed in terms of number of parameters and storage size).

- Bin4:  $200 \text{ MB} \leq S < 500 \text{ MB}$
- Bin5:  $S < 200 \text{ MB}$

**STEP 2:** Translation quality assessment using COMET, following the same procedure adopted in the offline track (i.e., computing COMET scores on the test sets after automatically resegmenting the system hypotheses and aligning them with the reference translations using mwerSegmenter<sup>29</sup>).

The rationale behind this evaluation protocol was to enable an independent assessment of models within the same size bin, thereby ensuring fairness and meaningfulness in the comparisons.

### 6.3 Submissions and Results

The task had only one participant, **TCD** (Moslem, 2025), that submitted a constrained primary system and an unconstrained contrastive one. The constrained system reduced the number of parameters by 40% by means of 4-bit quantization and QLoRa finetuning, after a full-finetuning of the base model (Qwen2-Audio-7B-Instruct) on the in-domain data. During the QLoRa finetuning, sequence-level knowledge distillation from the full-finetuned model is employed. For the unconstrained system, the method is similar, but after the first finetuning of the whole model a layer pruning strategy on the decoder (from 32 to 24 layers) is applied to further streamline the model, followed by another full-finetuning of the resulting model.

As seen from Table 5, the submitted runs exhibit mixed results with respect to our two evaluation dimensions. On the one hand, looking at **model reduction**, the number of parameters (5.0B for the constrained submission, 4.1B for the unconstrained one) and storage usage (9.7 GB and 8.8 GB, respectively) of the compressed models are notable but insufficient to meet the most relaxed size requirements defined by Bin1 (i.e., a maximum of 4 GB of storage). This highlights the difficulty of the task and the need to further

explore more aggressive techniques, as there remains significant room for improvement.

On the other hand, considering **translation performance**, it is encouraging to observe that, although the reductions were insufficient to fall into any of the target compression bins, the output quality across both target languages is even higher than the original model, thanks to dedicated finetuning on in-domain data, despite the applied compression techniques. The COMET scores show relative increases up to 13.43% on English→German and 9.46% on English→Chinese compared to the original, uncompressed Qwen2-Audio model. This is a non-trivial outcome, especially given the typical trade-offs involved when attempting to reduce the computational requirements of a large model.

In light of these findings, we believe that the challenges introduced in this first round of the model compression track remain open. The substantial margin for improvement observed should encourage broader participation in future rounds, driven by the growing need for efficient, accessible, and deployable SLT systems.

## 7 Low-resource SLT

The 5<sup>th</sup> edition of the Low-resource Spoken Language Translation track focused on the translation of speech from a variety of data-scarce languages. The target language is typically a higher-resource one, generally of similar geographical or historical linkages. The goal of this 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, many of the world’s languages and dialects lack the parallel data at scale needed for standard supervised learning.

Recognizing that the biggest bottleneck towards truly language-inclusive speech translation systems is data availability, this year’s edition included a data track, inviting participants to contribute newly collected speech translation datasets

<sup>29</sup> [www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz](http://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz)

for under-resourced language pairs.

## 7.1 Challenge

**Systems Track** This year’s task significantly expanded the typological and geographical diversity of the languages, language families, and scripts represented. The ten subtasks were:

- North Levantine Arabic → English (apc-eng)
- Tunisian Arabic → English (aeb-eng)
- Bemba → English (bem-eng)
- Fongbe → French (fon-fra)
- Irish → English (gle-eng)
- Bhojpuri → Hindi (bho-hin)
- Estonian → English (est-eng)
- Maltese → English (mlt-eng)
- Marathi → Hindi (mar-hin)
- Quechua → Spanish (que-spa)

Teams were allowed to submit to as few as one language pair, up to all ten. Both constrained and unconstrained submissions were allowed, to be separately ranked. For the constrained scenario, teams were only allowed to submit systems using the data provided by the shared task. For the unconstrained systems, teams were allowed to use any data as well as any pre-trained models.

**Data Track** This track aimed to empower language communities to contribute datasets. Such datasets are essential for expanding the reach of spoken language technology to more languages and varieties.

Participants of this track were encouraged to get creative with data creation strategies, while also ensuring data quality. As such, data track instructions included the following:

- Translations should be performed, wherever possible, by qualified, native speakers of the target language. We strongly encouraged verification of the data by at least one additional native speaker.
- Submitted datasets should be accompanied by dataset cards.<sup>30</sup> These should detail precise language information and the translation workflow that was employed. In particular, we asked 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.
- We highly encouraged new contributions to be released under CC BY-SA 4.0 or other similarly

permissive licenses. By contributing data to this shared task, participants agreed to have this data released under these terms. At a minimum, data should be made available for research use.

- While post-editing of automatic output was allowed, we required that any data submitted for the shared task are 100% verified by humans, if not directly created by humans. Raw, unverified machine translated outputs were not allowed. If using MT, we tasked participants with ensuring that the terms of service of the model they used allow re-using its outputs to train other machine translation models (for example, popular commercial systems such as DeepL, Google Translate and ChatGPT explicitly disallow this).

## 7.2 Data and Metrics

Table 6 provides a summary of the training data that were part of the shared task. We describe in more detail the data for each language pair below.

### **North Levantine Arabic–English (apc-eng)**

Levantine Arabic (ISO code: `apc`) is a well-established unit within the Arabic dialectal continuum, spoken mainly in Syria, Jordan, Lebanon, and Palestine. Although historically often split into *North* and *South* sub-dialects, recent ISO categorizations unite them under a common variant. Nonetheless, we maintain this finer split to emphasize the distinct phonological features and linguistic variations that characterize regional accents.

As in the first run of the *apc-eng* language pair, participants were provided with the UFAL Parallel Corpus of North Levantine 1.0 (Sellat et al., 2023), which includes about 120k lines of multi-parallel North Levantine-Modern Standard Arabic-English textual data, that can be downloaded from the LINDAT/CLARIAH-CZ Repository.<sup>31</sup> For additional speech data in Levantine Arabic, participants were pointed to two LDC resources: the BBN/AUB DARPA Babylon Levantine corpus (Makhoul et al., 2005) and the Levantine Arabic QT Training Data Set 5 corpus (Maamouri et al., 2006). Participants were also encouraged to make use of the Tunisian Arabic and Modern Standard Arabic resources made available in previous IWSLT editions.

Given the limited amount of publicly available corpora, we adopted the design of the initial *apc-eng* language pair run and focused exclusively on the unconstrained scenario.

<sup>30</sup>[github.com/openlanguagedata/oldi.org/blob/main/resources/dataset-card-template.md](https://github.com/openlanguagedata/oldi.org/blob/main/resources/dataset-card-template.md)

<sup>31</sup>[hdl.handle.net/11234/1-5033](https://hdl.handle.net/11234/1-5033)



The development<sup>32</sup> and test<sup>33</sup> data consist of recordings of native speakers of the dialect and are a mix of spontaneous monologues and dialogues on topics of everyday life (health, family life, sports) as well as characteristics of the country of origin (Syrian traditions, education system, culture, etc.). The transcription and translation team consisted of students of Arabic at Charles University, with an additional quality check provided by the native speakers of the dialect.

**Tunisian Arabic–English (aeb-eng)** Tunisian Arabic (ISO code: aeb) is the main spoken language in Tunisia. It is heavily influenced by the Arabic language. Due to its geographic position, the spoken language of Tunisia was also influenced by other languages including Tamazight, French and Turkish. As was the case of IWSLT22 and 23, the provided Tunisian Arabic–English corpus consists of around 323 hours of Tunisian Conversational Telephone Speech (CTS) along with manual transcripts made available by LDC. A subset of the above transcript (200k lines that represent 167 hours of speech) was manually translated into English and provided as training data for the speech translation task. In this 2025 evaluation campaign, participants also had access to an additional Tunisian dialect corpus of manually transcribed 08 hours of conversational speech (Mdhaifar et al., 2024).

All train and test sets are time-segmented at the utterance level. The development and test sets are the same official sets used during IWSLT 2022 and 2023.

**Bemba–English (bem-eng)** Bemba (also known as IciBemba) is a Bantu language (ISO code: bem), spoken predominantly in Zambia and other parts of Africa by over 10 million people. 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 Northern, 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 Anastopoulos, 2022) datasets.

**Fongbe–French (fon-fra)** Fongbé (also spelled Fongbè or Fon) is a Gbe language (ISO 639-3: fon). Fongbe, a tonal African language, is the most spoken dialect of Benin, by more than 50% of Benin’s population, including 8 million speakers. Fongbe is also spoken in Nigeria and Togo. The provided dataset contains over 48 hours of Fongbe audio recordings aligned with French translations. Additionally, a validation set of over 6 hours is included. The data used for this shared task is the extended version of the FFSTC corpus recently released (Kponou et al., 2025a). All recordings are derived from reading sessions by native Fongbe speakers, making this dataset a valuable resource for speech translation and low-resource language processing research.

**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 recognized 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., 2020a),<sup>34</sup> and Living-Audio-Dataset.<sup>35</sup> The Irish-to-English corpus comprises approximately 12 hours of Irish speech data (see Table 6), translated into English texts.<sup>36</sup> 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.<sup>37</sup>

<sup>34</sup>[commonvoice.mozilla.org/en/datasets](https://commonvoice.mozilla.org/en/datasets)

<sup>35</sup>[github.com/Idlak/Living-Audio-Dataset](https://github.com/Idlak/Living-Audio-Dataset)

<sup>36</sup>[github.com/shashwatup9k/iwslt2025\\_ga-eng](https://github.com/shashwatup9k/iwslt2025_ga-eng)

<sup>37</sup>[hf.co/collections/ymoslem/irish-english-speech-](https://hf.co/collections/ymoslem/irish-english-speech-)

<sup>32</sup>IWSLT 2024 devset and testset (with references): [hdl.handle.net/11234/1-5518](https://hdl.handle.net/11234/1-5518), [hdl.handle.net/11234/1-5519](https://hdl.handle.net/11234/1-5519)

<sup>33</sup>[hdl.handle.net/11234/1-5924](https://hdl.handle.net/11234/1-5924)



**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 23.31 hours of Bhojpuri speech data (see Table 6) from the news domain, extracted from News On Air<sup>38</sup> and translated into Hindi texts.<sup>39</sup> Additionally, the participants were directed that they may use monolingual Bhojpuri audio data (with transcription) from ULCA-asr-dataset-corpus<sup>40</sup> as well as Bhojpuri Language Technological Resources (BHLTR) (Ojha et al., 2020; Ojha, 2019)<sup>41</sup> and Bhojpuri-wav2vec2 based model.<sup>42</sup>

**Estonian–English (est-eng)** Estonian (ISO code: est) belongs to Finnic branch of the Uralic language family. It is the official language of Estonia and is spoken natively by about one million people.

The provided training set consists of 581,647 utterances (1,258 hours), while the development set includes 1,601 utterances (3.6 hours). The training data is sourced from the TalTech Estonian Speech Dataset 1.0 (Alumäe et al., 2023), a manually transcribed corpus primarily comprising broadcast material, created for training speech recognition models. All recordings are long-form speech, transcribed and time-aligned at the utterance level. In this dataset, long recordings have been segmented into individual utterances. The transcripts have been automatically translated into English using Google Translate in 2024 (Sildam et al., 2024).

The development and test sets include speech from government and municipal press confer-

ences, TV news, radio shows and talk shows, covering a variety of topics (sports, AI, international relations). The English translations have been manually created by professional translation agencies, instructed to translate without using any MT systems for post-editing. Both the original Estonian transcriptions and their English translations are provided for all utterances.

**Maltese–English (mlt-eng)** Maltese (ISO code: mlt) is a Semitic language with a heavy influence from Italian and English. It is spoken primarily in Malta, as well as in migrant communities abroad, notably in Australia, parts of the United States, and Canada.

The data release for this shared task comprises over 14 hours (split into development and training sets) of audio data, along with their transcription in Maltese and translation into English. Participants were allowed to use additional Maltese data, including the text corpus used to train BERTu (Micallef et al., 2022), a Maltese monolingual BERT model, the MASRI Data speech recognition data (Hernandez Mena et al., 2020), and any data available at the Maltese Language Resource Server.<sup>43</sup>

**Marathi–Hindi (mar-hin)** Marathi (ISO code: mar) is an Indo-Aryan language and is dominantly spoken in the state of Maharashtra in India. It is one of the 22 scheduled languages of India and the official language of Maharashtra and Goa. As per the 2011 Census of India, it has around 83 million speakers which covers 6.86% of the country’s total population.<sup>44</sup> Marathi is the third most spoken language in India.

The provided Marathi–Hindi corpus consists of 25.12 hours of Marathi speech data (see Table 6) from the news domain, extracted from News On Air<sup>45</sup> and translated into Hindi texts.<sup>46</sup> The dataset was manually segmented and translated by Panlingua.<sup>47</sup> Additionally, the participants were directed that they may use monolingual Marathi audio data (with transcription) from Common Voice (Ardila et al., 2020a),<sup>48</sup> as well as the corpus provided by He et al. (2020)<sup>49</sup> and the Indian Language Cor-

translation-datasets-665dd9e8fbba279db3474ca0

<sup>38</sup> [newsonair.gov.in](https://newsonair.gov.in)

<sup>39</sup> [github.com/panlingua/iwslt2025\\_bho-hi](https://github.com/panlingua/iwslt2025_bho-hi)

<sup>40</sup> [github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus](https://github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus)

<sup>41</sup> [github.com/shashwatup9k/bho-resources](https://github.com/shashwatup9k/bho-resources)

<sup>42</sup> [www.openslr.org/64/](https://www.openslr.org/64/)

<sup>43</sup> [mlrs.research.um.edu.mt/](https://mlrs.research.um.edu.mt/)

<sup>44</sup> [censusindia.gov.in/nada/index.php/catalog/42561](https://censusindia.gov.in/nada/index.php/catalog/42561)

<sup>45</sup> [newsonair.gov.in](https://newsonair.gov.in)

<sup>46</sup> [github.com/panlingua/iwslt2025\\_mr-hi](https://github.com/panlingua/iwslt2025_mr-hi)

<sup>47</sup> [panlingua.co.in/](https://panlingua.co.in/)

<sup>48</sup> [commonvoice.mozilla.org/en/datasets](https://commonvoice.mozilla.org/en/datasets)

<sup>49</sup> [www.openslr.org/64/](https://www.openslr.org/64/)

Language Pairs	Train Set	Dev Set	Test Set	Additional Data
North Levantine–English	apc-eng -	2.5	1.39	IWSLT 2024 test set (with references)
Tunisian Arabic–English	aeb-eng 323.0	-	-	A 160 hours out of this 323 hours are manually translated into English. 8h of transcribed speech from TARIC data set are also provided. Evaluation sets are same as IWSLT23.
Bemba–English	bem-eng 167.17	5.89	5.83	28.12 hours of monolingual audio with transcriptions (ASR) and 60 hours of untranscribed audio data.
Fongbe–French	fon-fra 48	6.1	5.9	A 57 hours of spoken Fongbe with corresponding French translations
Irish–English	ga-eng 9.46	1.03	0.66	A 196 hours of Synthetic Data, IWSLT 2023 and 2024 test set (with references) and MT data (monolingual and parallel corpora)
Bhojpuri–Hindi	bho-hi 19.88	2.07	0.54	IWSLT 2024 test set (with references ) and Monolingual audio with transcription (ASR) and monolingual text
Estonian–English	est-eng 1258.0	3.6	4.22	Remark: training data is synthetic (ASR data, machine-translated to English)
Maltese–English	mlt-eng 11.83	2.52	2.0	Monolingual audio with transcriptions (ASR), monolingual text
Marathi–Hindi	mr-hi 15.88	3.66	0.46	Monolingual audio with transcriptions (ASR), IWSLT 2023 and 2024 test set (with references) and monolingual text
Quechua–Spanish	que-spa 1.60	1.03	1.03	48.0 hours of monolingual audio with transcriptions (ASR) and post-edited translations (new) along with extra MT data

Table 6: Training, development and test data details (hours) for the language pairs of the low-resource shared task.

pora (Abraham et al., 2020).<sup>50</sup>

**Quechua–Spanish (que-spa)** Quechua (macro-language ISO code: que) 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 are found to be similar to other languages like Finnish. The average number of morphemes per word (synthesis) is about two times larger than in English. English typically has around 1.5 morphemes per word and Quechua has about 3 morphemes per word.

There are two main regional 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 a “southern” language. We label the data set with que - the ISO norm for Quechua II mixtures.

Due to the lack of data and low performance in previous work ((Salesky et al., 2024; E. Ortega

et al., 2024)), the organizers decided to allow only *unconstrained* submissions this year. The unconstrained setting consists of 1 hour and 40 minutes of training data and divided into 573 training files, 125 validation files, and 125 test files which are excerpts from the Siminchik corpus translated by native Quechua speakers. (Cardenas et al., 2018) Additionally, participants were directed to another larger data set from the Siminchik corpus which consisted of 48 hours of fully transcribed Quechua audio (monolingual). In this year’s task (2025), the organizers also included post-edited translation from Google of the 48 Siminchik hours which did not have translations last year (2024). Another MT dataset is offered in a parallel format, similar to last year. (Ortega et al., 2020) It consists of 100 daily magazine article sentences and 51140 sentences which are of religious context in nature.

### 7.2.1 Metrics

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

<sup>50</sup>[www.cse.iitb.ac.in/~pjyothi/indicorpora/](http://www.cse.iitb.ac.in/~pjyothi/indicorpora/)

### 7.3 Submissions

The Shared Task received a record 109 submissions (for speech translation) from 12 teams for all 10 language pairs. The submissions that provided an accompanying system paper are described in detail below and outlined in Table 7.

**AIB-MARCO** This team employed a cascade speech translation system consisting of Whisper/SeamlessM4T and Qwen2.5-7B-instruct. They performed sliding window ASR on the input audio then segment-level translation based on the transcription from the ASR model.

For primary systems of apc-eng and est-eng they used Whisper-large as the ASR model, whereas for gle-eng they used SeamlessM4T as the ASR model. For the contrastive systems, they employed different ASR models. The LLM used in translation is an optimized Qwen2.5-7B-instruct model.<sup>51</sup>

**ALADAN** (Kheder et al., 2025) provided a submission for the North Levantine Arabic to English direction, building on the same team’s efforts from last year (Kheder et al., 2024). It is a cascade of ASR and MT systems. For the MT part data sparsity is alleviated via a crowd-sourced parallel corpus that covers five major Arabic dialects (Tunisian, Levantine, Moroccan, Algerian, Egyptian), curated via rigorous qualification and filtering. They also include an additional experiment with a large, high-quality Levantine Arabic corpus from LDC, which does not benefit from adding the crowdsourced data. ASR is done with a TDNN-F model and a Zipformer, whereas compared to the previous year’s submission, a 4-times bigger model is taken for Zipformer (253M parameters). The methodology also includes dialect-specific normalization of Arabic text.

**BUINUS** (Tjjaranata et al., 2025) focused on the mlt-eng direction. Their system employs a cascade architecture, combining ASR and translation to handle the low-resource setting better. For ASR, they use Whisper (Radford et al., 2022), which was further fine-tuned with the data provided in the shared task. For the translation step, they use NLLB model (NLLB Team et al., 2022), employing both direct fine-tuning and data augmentation techniques designed to modify the target

sequences and thereby reinforce encoder reliance and decoder robustness. Fine-tuning of NLLB was carried out in two stages: an initial stage used a combination of real and augmented data, followed by a second stage fine-tuning exclusively on the main task to refine the model further. To efficiently fine-tune larger models under computational constraints, they used QLoRA (Dettmers et al., 2023), achieving better performance with the 3.3B parameter model compared to smaller versions. Notably, their analysis revealed that data augmentation yielded comparatively greater performance gains for smaller models, underscoring the value of data-driven strategies in resource-constrained scenarios. They note, however, that the performance difference between the larger and smaller NLLB models was modest, and the errors at the ASR stage hurt the translation component.

**GMU** (Meng and Anastasopoulos, 2025) submitted systems for all language pairs except apc-eng. Their approach focuses on fine-tuning SeamlessM4T-v2 for ASR, MT, and ST tasks. The fine-tuned ASR and MT models are used to construct cascaded ST systems. They also explored various training paradigms for ST fine-tuning, including direct end-to-end (E2E) fine-tuning, parameter initialization using fine-tuned ASR and/or MT model components, and multi-task training. The multi-task training setup includes ST, MT and knowledge distillation (KD) objectives, where KD leverages the MT components to enhance the ST components. They found that direct E2E fine-tuning yielded strong overall results, and initializing the ST encoder with an in-domain fine-tuned ASR encoder further improved performance on languages SeamlessM4T-v2 had not been previously trained on. Multi-task training, on the other hand, provided marginal improvements.

**JHU** Johns Hopkins University’s team, (Robinson et al., 2025), participated in all language pairs continuing their tradition from last year (Romney Robinson et al., 2024). As with the previous year, the motivation was to assess the robustness of the methods they were employing across a variety of domains and typologically diverse languages. However, the main focus this year was on ensembling methods, and in particular, Minimum Bayes Risk (MBR) decoding (Bickel and Doksum, 1977; Kumar and Byrne, 2004). In order to do so, they aimed to gather a variety of different submissions

<sup>51</sup>This description was provided by the participants. No associated paper was submitted.

Team Name	Language Pairs									
	apc-eng	aeb-eng	bem-eng	fon-fra	bho-hin	gle-eng	est-eng	mlt-eng	mar-hin	que-spa
Systems Track										
AIB-MARCO	✓					✓	✓			
ALADAN (Kheder et al., 2025)	✓	✓								
BUINUS (Tjjaranata et al., 2025)								✓		
GMU (Meng and Anastasopoulos, 2025)		✓	✓	✓	✓	✓	✓	✓	✓	✓
IIITH-BUT (Akkiraju et al., 2025)					✓					
JHU (Robinson et al., 2025)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
KIT (Li et al., 2025)	✓	✓	✓							
KREASOF-TCD (Farouq et al., 2025)			✓							
LIA (Chellaf et al., 2025)	✓	✓		✓						
QUESPA (Ortega et al., 2025)										✓
SYSTRAN (Avila and Crego, 2025)		✓								
Teams per Pair:	5	6	4	3	3	3	3	3	2	3
Data Track										
KUVOST (Mohammadamini et al., 2025)	English - Central Kurdish									
URDU (Mehmood and Rauf, 2025)	Urdu - English									
FFSTC-2 (Kponou et al., 2025b)	Fongbe - French									

Table 7: Breakdown of the teams and the language pairs subtasks that they participated in for the Low-Resource Shared Task.

for each language pair. They relied on both end-to-end translation systems, as well as cascaded systems. In addition, they looked at combining similar languages for mixed data training. Overall, the results were mixed with ensembling helping in some language pairs and hurting in others. However, a key takeaway is that for practioners, MBR is still helpful because you do not need to know which system is the best in advance.

LIA (Chellaf et al., 2025) participated in three language pairs - both of the arabic dialects, as well as Fongbe to French. All of their submissions were in the unconstrained setting relying on pre-trained models. They explored both pipelined systems and end-to-end systems. They investigated various ways of augmenting systems with varying data, such as combining Modern Standard Arabic (MSA) data with dialectal arabic, or looking at including Fongbe transcriptions both with and without diacritics. For the Tunisian-to-English translation task, their primary system was an end-to-end system based on a language-agnostic semantically aligned speech encoder. They trained it following the SAMU-XSLR framework (Khurana et al., 2022) from the w2v-bert 2.0 (Seamless Communication et al., 2023) model as a student and BGE-M3 text model (Chen et al., 2024) as a teacher. For the North Levantine-to-English task, their primary system was based on a combination of cascaded systems. The two ASR modules were based on Whisper-large-v3: these models have been fine-

tuned on the Levantine data released by the organizers but also on Modern Standard Arabic data. The two MT models applied to the ASR outputs were based on NLLB-200 1.3B fine-tuned on the official data augmented with the Levanti corpus, available on Hugging Face <sup>52</sup>. Each MT model fed by different ASR output generated 10 translation hypotheses. The final selection was made by using BLASER (Chen et al., 2023). Last, for the Fongbe to French translation task, their primary system was also a cascaded system using an ASR module built on the AfriHuBERT SSL speech encoder (Alabi et al., 2024) and an MT module based on the NLLB model.

KIT (Li et al., 2025) participated in the Bemba-to-English, North Levantine Arabic-to-English, and Tunisian Arabic-to-English tasks under the unconstrained condition. They explored both cascaded and end-to-end ST systems. All approaches were based on pretrained models: SeamlessM4T (Seamless Communication et al., 2023) for end-to-end ST, NLLB (NLLB Team et al., 2022) for MT, and MMS (Pratap et al., 2024) and XEUS<sup>53</sup> for ASR. The main focus was on using synthetic data for data augmentation and applying model regularization techniques. Two types of synthetic data generation were studied: (1) translating source language ASR data using MT systems to create ST training data, and (2) generating source lan-

<sup>52</sup>[huggingface.co/datasets/guymorlan/levanti](https://huggingface.co/datasets/guymorlan/levanti)

<sup>53</sup>[huggingface.co/espnet/xeus](https://huggingface.co/espnet/xeus)



guage speech via text-to-speech from MT training data. Results showed that ST models trained only on synthetic data can outperform cascaded systems, provided that a strong MT system is used. The impact of TTS-based augmentation varied: it was effective only when the TTS quality was high. Regularization experiments used intradistillation (Romney Robinson et al., 2024), which proved to be a reliable and broadly applicable method across all tasks in low-resource settings.

IIITH-BUT (Akkiraju et al., 2025) fine-tuned SeamlessM4T models for Bhojpuri-Hindi speech translation. To address data scarcity, they applied speed perturbation and SpecAugment data augmentation techniques. Moreover, they examined cross-lingual transfer learning through joint training with Marathi and Bhojpuri speech data.

The team experimented with two variants of SeamlessM4T, medium (1.2B parameters) and large v2 (2.3B parameters). For hyperparameter optimisation, they explored a range of values for batch size, learning rate, label smoothing, and warmup steps. For data augmentation, they used SpecAugment to apply spectrogram masking with time masks and frequency masks, and implemented speed perturbation with speed factors of 0.9x, 1.0x, and 1.1x. This data augmentation process resulted in expanding the training data by three times. Finally, they combined the Marathi-Hindi parallel data with the limited Bhojpuri-Hindi dataset to fine-tune the SeamlessM4T models.

KREASOF-TCD (Farouq et al., 2025) This team participated in the Bemba-to-English shared task under unconstrained conditions. The team submitted three speech translation systems based on the cascading method. The *Primary* submission system is based on medium-sized Whisper for the ASR and NLLB-200 3.3B for the MT. The *Contrastive 1* and *Contrastive 2* systems use the small-sized Whisper model for the ASR, while the MT systems are based on the NLLB-200 3.3B and NLLB-200 600M models, respectively. The team explored fine-tuning pre-trained models and data augmentation for their strategy to develop the systems. The ASR systems were obtained by fine-tuning the Whisper models on the data from the BembaSpeech (Sikasote and Anastopoulos, 2022) and BIG-C (Sikasote et al., 2023a) datasets. The MT systems are based on

the NLLB-200 (NLLB Team et al., 2022) model, which is fine-tuned on bilingual segments of the BIGC and "dev" split of the FLORES-200 (Goyal et al., 2022) datasets. To improve the quality of speech translations, the team explored augmenting the Bemba-to-English training data with the portion of the Tatoeba (Tiedemann, 2020) dataset that was back-translated from English into Bemba using the NLLB-200 600M model. The back-translations were filtered using cross-entropy scores.

SYSTRAN (Avila and Crego, 2025) participated in one language pair, Tunisian Arabic to English, under the *constrained* condition using the resources provided by LDC for this task that included MSA data from broadcast news and Tunisian Arabic conversational telephone speech. The focus of their contribution was on tightly coupling an ASR encoder (Whisper, the Medium and Large-v3 versions tested) with an NMT decoder (NLLB, the 3.3B parameter version). Embeddings from the Whisper encoder are fed into the NLLB decoder via a Reshape module consisting of a convolutional layer and linear projection layer instead of using the standard word embeddings. The motivation is for parameter-efficient training in low-resource settings, ensuring quality translations while being scalable. They fine-tuned their model using the available LDC parallel corpora, with additional filtering and cleaning strategies to optimize domain robustness and translation consistency.

QUESPA (Ortega et al., 2025) submitted three *unconstrained* systems this year as the Quechua-Spanish shared task organizers only allowed unconstrained setting submissions. Team QUESPA were able to improve the previous year's results despite the baseline task data remaining mostly the same with exception of newly machine-translated text from the original Siminchik corpus. The three unconstrained systems ranged from 14.8 BLEU to 26.7 BLEU where QUESPA's best performing systems from last year (2024) ranged from 11.1 to 19.7 BLEU. The 7 BLEU points of improvement of their best system is attributed to a new ASR dataset released from the "quy" ISO code called Collao, a dialect of Quechua spoken mostly in southern Peru. (Paccotacya-Yanque et al., 2022)

QUESPA's *unconstrained* systems were once again a novel introduction for the QUE-SPA



Language Pair	Winning Team	System	Constrained	BLEU
apc-eng	KIT	primary	no	23.3
aeb-eng	KIT	primary	no	21.4
bem-eng	GMU	primary	no	31.7
fon-fra	LIA	primary	no	39.6
bho-hin	JHU	primary	no	10.7
gle-eng	GMU	primary	no	13.4
est-eng	AIB-MARCO	primary	no	30.9
mlt-eng	GMU	primary	no	57.5
mar-hin	GMU	contrastive1	no	44.3
que-spa	QUESPA	contrastive2	no	26.7

Table 8: Winning submissions for each language pair of the Low-Resource Shared Task.

task and outperformed last year’s best systems. The Primary System was not previously used by QUESPA in IWSLT. It is comprised of a cascaded ASR + MT system where ConMamba (Jiang et al., 2024), based on a Conformer architecture (Gulati et al., 2020) is used for ASR using publicly available recipes<sup>54</sup>, experimenting with small (S) and large (L) configurations (144/512 dimensions, 12+4/12+6 layers). The resulting transcribed text is then passed into a newly created (fine-tuned) NLLB (NLLB Team et al., 2022) machine translation system that was tested in development with several combinations that finally resulted in 18 BLEU on the test set. The *Contrastive 1* system is similar to QUESPA’s submission from 2024, however, Whisper **Version 3** is used this year along with ESPnet (Watanabe et al., 2018). Results from the Whisper V3 model were then passed into the NLLB-based MT model used in the Primary system. The *Contrastive 2* system is QUESPA’s biggest achievement yet and can be considered the most novel system to data for speech translation on the Quechua–Spanish language pair. It is a pre-trained SpeechT5 (Ao et al., 2022) model fine-tuned for Speech Translation using the unconstrained training data along with the 48 hours of newly created post-edited MT data. Furthermore, they applied a data augmentation technique called *nlpaug* (noise, distortion, duplication)(Ma, 2019) which resulted in a total of 96h: 48h original + 48h of new synthetic data. Lastly, their best addition to the Contrastive 2 system was the inclusion of a Collao speech translation corpus that contains 15 hours of Quechua Collao translated speech (quz). The resultant training data set for the Contrastive 2 system thus was: 96 + 15 (111) total hours of Quechua speech translations. (Paccotacya-Yanque et al., 2022)

<sup>54</sup>[github.com/xi-j/Mamba-ASR](https://github.com/xi-j/Mamba-ASR)

## 7.4 Results

**General Notes** Table 8 summarizes the winning submissions for each language pair. Detailed results for all teams’ systems and settings are available in Appendix B.5.

Of the 10 language pairs, 6 different teams had the top performing system on at least one language pair. This shows how competitive the shared task was, and that a multitude of approaches are helpful for low-resource speech translation.

Compared to previous iterations of the shared task, some of the language pairs had marked improvements with large gains in the official automatic metrics. For example, BLEU scores for Quechua-Spanish, the least resourced language pair, improved from 19.7 to 26.7 BLEU points (this was largely the result of the use of additional data by the winning team). However, for other continuing language pairs, performance is rather stagnated, remaining in exactly the same levels (if not worse) for Bemba-English, Bhojpuri-Hindi, Irish-English, and the two Arabic dialects language pairs. This might suggest that we have perhaps reached a performance ceiling of sorts in the current datasets under the current data-scarce conditions, especially for the language pairs that lie in the low-end of data availability. It should be noted, though, that this ”ceiling” performance nevertheless still lags substantially behind the translation quality we observe for high-resource pairs, still reinforcing the need for further data collection and research in the area.

For the language pairs included for the first time in the shared task, we find that Estonian-English, our highest resourced language pair with more than 1,200 hours of translated audio, ends up with speech translation systems of decent quality with BLUE scores in the 29–31 range by multiple participants. On the other hand lies Fongbe-Frneh,

which even though does end up with decent systems yielding BLEU scores over 30 by two participants. Similar to last year’s findings, we see our current technologies can produce good ST systems for language pairs with more than 50 hours of high-quality translated speech.

We note that almost all submissions followed the unconstrained setting – a clear indication that pre-trained multilingual systems seem to be the best option for building ST for low-resource languages, at least under the current data, architectural, and compute constraints.

**Notes on *apc-eng*** Compared to the initial run of the *apc-eng* language pair in the previous year, the performance gap between the top-ranked system (KIT) and the remaining participants has narrowed. The second-place team (LIA) achieved results within 1 BLEU point of the winner, while the third-place team (ALADAN, last year’s champion) trailed by only 3.5 BLEU points. Although the absolute BLEU score achieved by the top-performing system is notably lower than that of the previous year (23.34 vs. 28.71), we attribute this discrepancy not to a decline in overall system quality (quite the contrary!) but rather to differences in the test set composition (2024: 974 lines, 12,263 words; 2025: 1,026 lines, 8,833 words). Although the ranking based on chrF corresponds with the BLEU evaluation, COMET indicates that the LIA system surpasses KIT, emphasizing the minimal performance differences among the leading submissions. The integration of end-to-end and cascaded systems, particularly through MBR decoding to combine translation hypotheses, proved to be a successful strategy for enhancing overall system performance. Due to their strong multilingual capabilities, the Whisper and NLLB models continue to be among the most widely adopted solutions for ASR and MT, respectively. Top-performing systems demonstrated substantial quality improvements through the use of additional speech and text resources, often curated internally. Notably, LIA showcased the benefits of carefully filtering a general Arabic corpus using a dialect identification system. The generation of synthetic textual data via back-translation, forward-translation, and paraphrasing remains an effective method. The winning team, KIT, also experimented with synthetic speech data generation using a TTS model; however, this approach was found to be ineffective, primarily due to the

lack of high-quality speech data, leading to an under-trained TTS system. Several teams also reported findings regarding the impact of domain alignment in training and evaluation datasets, emphasizing the critical importance of developing resources for low-resource languages that are tailored to the practical needs of end users.

## 7.5 Data Track Results and Discussion

The data track received 3 submissions, each producing usable datasets for 3 low-resource language pairs: English-Central Kurdish, Urdu-English, and Fongbe-French. This successful first iteration reinforces the desire by researchers and communities to contribute open-source datasets. The organizers will plan to use these datasets in future iterations of the low-resource shared task as appropriate. We discuss the submissions below.

**KUVOST** (Mohammadamini et al., 2025) produced a large-scale English speech to Central Kurdish dataset by relying on the publicly available Common Voice dataset. This effort produced more than 1,000 hours of parallel speech translation data, by leveraging community volunteer work: more than 230 volunteers manually translated and revised more than 240k English sentences, which were then paired with their utterances in Common Voice. The effort included an extensive data validation process. The participants also ensure the quality of the data by producing pre-determined train-dev-test splits, and building baseline systems on top of fine-tuned Whisper v3 and Seamless M4T, leading to BLEU scores over 32 on the test set.

Note that this effort, in contrast to the norm for the systems track, produced data where the low-resource language (Central Kurdish) is on the target side and the high-resource one (English) is on the source speech side.

**URDU** (Mehmood and Rauf, 2025) produced an Urdu-English speech translation dataset. They relied on Common Voice 13.0 and its Urdu speech portion. The Urdu transcripts were first automatically translated into English, but then checked and corrected by 19 bilingual volunteers, as well as validated by a professional translator. This multi-stage quality assurance approach disentangles the correction of potential syntactic or grammatical errors from a secondary stage that ensures high-quality, fluent translations for idiomatic or po-

etic texts, highlighting the potential need for more careful handling of some data subdomains.

FFSTC-2 (Kponou et al., 2025b) presented an extension of the previous FFSTC dataset, adding another 36 hours to bring the available total to 61 hours of Fongbe-French data. Unlike the other two submissions, this team started with target-side text (French), which was first automatically translated to Fongbe. Then the text translations were reviewed by bilingual experts, and only at the end was read speech of these Fongbe translations collected. The effort included a validation process, e.g. to remove utterances with excessive background noise, where the validators re-recorded the utterances, yielding an additional 42k recordings.

The participants also confirmed the utility of these additional data, developing ST systems (both cascade and end-to-end) as well as ASR systems that improve over systems trained on the previous iteration of the corpus.

## 8 Indic Languages Track

The growing demand for inclusive digital access has highlighted the need for seamless cross-lingual communication, especially in linguistically rich regions like India. While English dominates global technology and information spheres, millions of speakers of Indic languages such as Bengali, Hindi, and Tamil still lack adequate speech and language technologies. Despite their large combined speaker base of over 700 million and significant cultural and economic importance, these languages remain underrepresented in NLP and speech research due to limited high-quality parallel text and audio data.

Compounding this challenge are the inherent complexities of Indic languages including rich morphology, high inflection, and frequent code-mixing in real-world discourse, which make Spoken Language Translation (SLT) development especially difficult (Sethiya and Maurya, 2024). Addressing this gap, the Indic Shared Task track at IWSLT 2025 focuses on SLT for Bengali, Hindi, and Tamil in both English→Indic and Indic→English directions. The latter is emphasized due to its higher complexity and the inclusion of STEM and broadcast media domains, demanding systems capable of handling technical vocabulary and varied speech styles.

By releasing the first benchmark dataset tailored to these low-resource languages across critical do-

main, this task aims to drive research that tackles real-world multilingual challenges. It seeks to advance digital inclusion, foster equitable access to global knowledge, and support the preservation and technological integration of Indic languages.

### 8.1 Challenge

The IWSLT 2025 Indic Shared Task track focuses on speech-to-text translation (ST) across six language directions: English-to-Bengali (en→bn), English-to-Hindi (en→hi), English-to-Tamil (en→ta), Bengali-to-English (bn→en), Hindi-to-English (hi→en), and Tamil-to-English (ta→en). This year’s challenge expands beyond previous iterations by including both Indic-to-English and English-to-Indic directions, though the data sources for each direction are distinct.

The track allows participants to submit in both the constrained and unconstrained conditions. The constrained condition permits only the use of the provided dataset, while the unconstrained condition allows the incorporation of additional external resources and pre-trained models. Systems can be either end-to-end (E2E) or cascaded, and participants may submit both monolingual and multilingual systems across any or all of the six language directions.

### 8.2 Data and Metrics

The Indic track at IWSLT 2025 provides a comprehensive speech-to-text translation (ST) corpus spanning three Indic languages: Bengali, Hindi, and Tamil. The dataset is constructed from two distinct sources, reflecting the two translation directions.

For the English-to-Indic (en→xx) direction, the data is derived from the Indic-ST corpus (Sethiya et al., 2024), which consists English speech paired with English transcripts and Indic translations. These data is from domains like Mann ki Baat, and NPTEL, unlike IWSLT 2024, which had data from TED talks (Sethiya et al., 2024). The dataset is segmented using provided YAML files, ensuring consistent alignment across audio, English transcripts, and Indic translations. Table 9 reports the number of lines and audio hours, partitioned into training, validation, and test splits. Note that due to linguistic differences, the token counts between English and the target Indic languages naturally vary.

For the Indic-to-English (xx→en) direction, the data is sourced from a curated subset of

the BhasaAnuvaad dataset (Sankar et al., 2025), which draws from rich educational and broadcast domains. Specifically, it includes material from the National Programme on Technology Enhanced Learning (NPTEL), the Spoken-Tutorial project, and Mann-ki-Baat addresses, covering specialized STEM content as well as public broadcast speech. This direction provides a new challenge for participants, requiring systems to handle domain-specific terminology, varied accents, and spontaneous speech phenomena.

**English-Bengali (en↔bn):** Bengali, the seventh most spoken language globally, has around 228 million speakers and belongs to the Indo-Aryan family. It is the official language of Bangladesh and is widely spoken in the Bengal region of India, written in the Bengali-Assamese script. The en→bn dataset comprises 815 hours of English speech aligned to Bengali translations, while the bn→en set contains 157.95 hours of Bengali speech aligned to English text.

**English-Hindi (en↔hi):** Hindi is the third most spoken language in the world, with approximately 615 million speakers. It belongs to the Indo-Aryan family and is primarily spoken in India, where it serves as one of the official languages, written in Devanagari script. The en→hi dataset contains 815 hours of English speech and Hindi translations, while the hi→en dataset provides 653.88 hours of Hindi speech with aligned English translations.

**English-Tamil (en↔ta):** Tamil, a classical Dravidian language with approximately 91 million speakers, is spoken predominantly in the Tamil Nadu state of India and parts of Sri Lanka. It is written in the Tamil script derived from Brahmi. The en→ta dataset offers 815 hours of English speech with aligned Tamil translations, whereas the ta→en data includes 378.16 hours of Tamil speech with English translations.

**Evaluation Metrics:** For system evaluation, we primarily employ the chrF++ metric (Popović, 2017), chosen for its high correlation with human judgments—especially in the context of Indian languages (Sai B et al., 2023)—making it particularly well-suited to our task. All chrF++ scores are computed using the standardized sacreBLEU toolkit (Post, 2018) to ensure consistency and reproducibility. In addition, we report BLEU scores for completeness, although they are not used in ranking the systems.

Lang.	Train		Valid		Test	
	Hours	Samples	Hours	Samples	Hours	Samples
en→hi	680.54	205.2k	40.48	11.67k	93.13	36.25k
en→bn	680.54	205.2k	40.48	11.67k	93.13	36.25k
en→ta	680.54	205.2k	40.48	11.67k	93.13	36.25k
bn→en	157.95	64.8k	1.00	395	1.25	858
hi→en	653.88	248.8k	1.00	397	1.34	579
ta→en	478.16	211.3k	1.00	457	2.18	956

Table 9: Summary of provided data for each language direction, including hours and number of samples.

### 8.3 Submissions

The 2nd edition of the Indic shared task track of IWSLT received 32 submissions for all six language pairs from five teams: the CDAC-SVNIT team from SNLP Lab, the CDAC Noida and SVNIT, Surat; the JU-CS-NLP team from Jadavpur University; another team, JU from Jadavpur University; team IITM from Speech Lab, IIT Madras; and team HITSZ from Harbin Institute of Technology, Shenzhen. The participants submitted their results under various constraints, including end-to-end constrained and unconstrained, cascaded constrained, and unconstrained approaches. Below, we provide an overview of each team’s approach and their results.

**CDAC-SVNIT (Roy et al., 2025):** This team submitted 12 systems, two for each of the six language pairs. Their submissions featured both cascaded and end-to-end approaches. The cascaded systems operated under an unconstrained setting, while the end-to-end systems adhered to a constrained setup. For the cascaded approach, they fine-tuned a pre-trained CLSRIL-23 model for ASR and a pre-trained IndicTrans2 model for MT. The end-to-end systems utilized a transformer-based encoder-decoder architecture from the Fairseq toolkit, pretrained on the provided data.

**JU-CS-NLP (Dhar et al., 2025):** This team submitted six systems, one for each language pair, under the unconstrained cascaded setting. For En → xx translation, the system employed OpenAI’s pre-trained Whisper Base model for ASR and a fine-tuned version of Meta’s NLLB-200-distilled-600M model for MT. For xx → En, it used the pre-trained IndicConformer model for ASR and the fine-tuned IndicTrans2 model for MT, both developed by AI4Bharat. The MT models are fine-tuned on the provided dataset.

**JU (Das et al., 2025):** The submission includes



Direction	Team ID	chrF++ / BLEU
en→bn	CDAC-SVNIT	62.21 / 36.96
	JU-CSE-NLP	<b>74.58 / 51.70</b>
	IITM	60.81 / 26.67
en→hi	CDAC-SVNIT	64.17 / 44.09
	JU-CSE-NLP	<b>72.98 / 57.61</b>
	IITM	62.30 / 41.09
en→ta	CDAC-SVNIT	66.15 / 29.34
	JU-CSE-NLP	<b>73.81 / 36.17</b>
	IITM	62.33 / 21.35
bn→en	CDAC-SVNIT	44.89 / 14.77
	JU-CSE-NLP	53.99 / 23.69
	JU	35.56 / 8.69
	IITM	<b>55.27 / 22.90</b>
hi→en	CDAC-SVNIT	67.06 / 41.04
	JU-CSE-NLP	67.91 / 44.13
	IITM	<b>68.14 / 41.59</b>
ta→en	CDAC-SVNIT	41.16 / 15.70
	JU-CSE-NLP	<b>49.34 / 17.66</b>
	JU*	39.02 / 13.39
	IITM	47.44 / 18.41

Table 10: Performance of unconstrained cascaded systems on different language pairs in terms of chrF++ and BLEU scores. The \* symbol denotes a system that used a multilingual base model without any finetuning.

an unconstrained cascade setting for 2 language pairs from Bengali and Tamil to English. A pre-trained Whisper Small model is used for ASR, which is pretrained for Bengali on the Bangla Mozilla Common Voice dataset and for Tamil on multiple publicly available datasets. For MT, the system utilized the fine-tuned MarianMT model for Bengali to English translation and the fine-tuned facebooknllb-200-distilled-600M model for Tamil to English translation.

**IITM (Sarkar et al., 2025):** The team submitted six systems under the unconstrained cascaded setting. For ASR, they used the Phi-4 model, fine-tuned separately for each language: Bengali using SKNahin/open-large-bengali-asr-data, Hindi using SpringLab/Hindi-1482hrs and AI4Bharat/SeamlessAlign, and Tamil using Prajwal-143/ASR-Tamil-cleaned. For MT, they employed the NLLB model, fine-tuned on the SPRINGLab/shiksha and SPRINGLab/BPCC-cleaned datasets for xx → English translation.

**HITSZ (Wei et al., 2025):** The team made 6 submissions for the unconstrained end-to-end setting for each of the 6 language pairs. The end-to-end system utilizes the encoder-decoder based Dhvani model, where the speech signals are encoded using the whisper speech encoder and the

Direction	chrF++ / BLEU
en→bn	52.69 / 27.00
en→hi	52.50 / 33.84
en→ta	54.67 / 22.81
bn→en	53.07 / 25.02
hi→en	62.94 / 39.29
ta→en	43.91 / 19.27

Table 11: Performance of unconstrained end-to-end systems by HITSZ on different language pairs in terms of chrF++ and BLEU scores.

non-speech audio signals are encoded using the BEAT’s encoder, which are bridged to the language model with the help of Q-former. The transformed tokens are decoded using the Krutrim large language instruct model.

## 8.4 Results

Tables 10, 11 & 12 present the performance of the submitted systems across six translation directions, evaluated primarily using the chrF++ (Popović, 2017) metric. Each direction was evaluated under both unconstrained and constrained settings, and systems were categorized as either cascaded or end-to-end (E2E) in design. The unconstrained setting permitted the use of any external data, while the constrained setting required systems to be trained using only the provided shared data. Below, we summarize the key findings per translation direction.

**en→bn** In the English-to-Bengali direction, the highest chrF++ score was achieved by the JU-CSE-NLP team using a cascaded system in the unconstrained setting, with a score of 74.58. CDAC-SVNIT and IITM also submitted strong cascaded systems, achieving 62.21 and 60.81 chrF++, respectively. Among end-to-end (E2E) systems, HITSZ obtained a chrF++ of 52.69, while in the constrained setting, CDAC-SVNIT’s E2E model led with 58.22 chrF++, indicating the effectiveness of their model despite the data restrictions.

**en→hi** For English-to-Hindi, the best chrF++ score again came from JU-CSE-NLP’s cascaded system under the unconstrained condition, reaching 72.98. CDAC-SVNIT and IITM followed closely with scores of 64.17 and 62.30, respectively. The E2E system from HITSZ achieved 52.50 chrF++, and CDAC-SVNIT’s constrained E2E model attained a respectable 54.48, outperforming several unconstrained E2E systems.

Direction	chrF++ / BLEU
en→bn	58.22 / 31.57
en→hi	54.48 / 34.61
en→ta	56.08 / 21.35
bn→en	14.30 / 00.46
hi→en	42.97 / 15.42
ta→en	26.25 / 05.05

Table 12: Performance of constrained systems submitted by CDAC-SVNIT using an end-to-end (E2E) approach. Only the provided shared data was used for training.

**en→ta** In the English-to-Tamil direction, JU-CSE-NLP led with a chrF++ of 73.81 using a cascaded approach under the unconstrained setting. This was followed by IITM (62.33) and HITSZ’s E2E model (54.67). Under the constrained condition, CDAC-SVNIT’s E2E model achieved 56.08 chrF++, showing competitive performance despite being limited to shared training data.

**bn→en** For Bengali-to-English, the highest chrF++ score was reported by HITSZ’s E2E system with 53.07, outperforming all cascaded systems including CDAC-SVNIT (44.89), IITM (55.27), and JU (35.56). Under the constrained condition, the best result was 14.3 chrF++ from CDAC-SVNIT’s E2E model, underscoring the difficulty of this direction when relying solely on shared data.

**hi→en** In the Hindi-to-English direction, the top-performing system was submitted by IITM with a chrF++ of 68.14 using a cascaded architecture under the unconstrained setting. This was closely followed by JU-CSE-NLP (67.91) and CDAC-SVNIT (67.06). HITSZ’s E2E model achieved 62.94 chrF++, while CDAC-SVNIT’s constrained E2E system reached 42.97, indicating a substantial drop in performance under constrained data.

**ta→en** For Tamil-to-English, JU-CSE-NLP’s cascaded system achieved the highest chrF++ score under the unconstrained setting with 49.34. Other strong systems included IITM (41.16) and JU\* (47.44), the latter of which utilized a multilingual model without fine-tuning. Among E2E approaches, HITSZ led with 43.91. CDAC-SVNIT’s constrained E2E system attained 26.25 chrF++, again reflecting the challenges imposed by data limitations in this direction.

## 8.5 Conclusion

This edition of the Low-Resource Indic Multilingual Speech Translation track marked the first time that translation from Indic languages to English was included alongside the English-to-Indic directions. This expansion provided a more comprehensive evaluation of multilingual translation capabilities and highlighted the unique challenges of translating into English from morphologically rich and syntactically diverse Indic languages.

Across the six language directions, systems demonstrated strong performance in both unconstrained and constrained settings, with cascaded architectures generally outperforming end-to-end approaches in the unconstrained track. However, several constrained end-to-end systems showed promising results, indicating progress toward robust low-resource translation without reliance on external data.

The wide range of approaches submitted—spanning cascaded pipelines, multilingual pre-training, and direct speech-to-text modeling—reflects growing diversity in system design for low-resource speech translation. These results offer valuable insights into the current state of the field and set a strong baseline for future editions of the task, especially in further improving Indic-to-English performance and in exploring more unified multilingual modeling techniques.

## 9 Instruction-Following Track

In recent years, large language models (LLMs) have redefined the landscape of natural language processing by demonstrating the ability to perform a wide range of tasks without requiring task-specific architectures or fine-tuning. These models offer a single, unified interface for diverse applications such as translation, summarization, and question answering, simply by conditioning on textual instructions (Hendy et al., 2023). Initially restricted to textual input, LLMs are now evolving into multimodal systems, incorporating modalities such as vision and speech to expand their applicability beyond the text domain (Li et al., 2024). In parallel, speech foundation models (SFM) have emerged as powerful architectures capable of processing spoken language at scale (Latif et al., 2023). When combined with the instruction-following capabilities of LLMs (Ouyang et al., 2022), they open new opportunities for building general-purpose speech models that are not lim-

ited to handling a pre-defined set of tasks (Rubenstein et al., 2023). This integration, often referred to as SpeechLLM or SFM+LLM (Gaido et al., 2024), promises to deliver very versatile systems, making it possible to interact with spoken language in flexible and controllable ways.

To explore this promising direction, this year we introduce, for the first time at IWSLT, a new shared task focused on evaluating instruction-following models for the speech modality. The goal is to assess models that can perform multiple speech-to-text tasks—such as automatic speech recognition, speech translation, spoken question answering, and summarization—by following natural language prompts, using either short audio segments or long-form spoken content as input.

## 9.1 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 a natural language prompt. The model receives both an audio 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 audio: **SHORT**, where the input is represented by automatically segmented audio (usually of a few seconds), and **LONG**, where the input is a long-form audio. Depending on the sub-track, the following tasks have to be supported by the model:

- **SHORT Sub-Track**
  - **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);
- **LONG Sub-Track**
  - **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);

- **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.

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, and S2TSUM, and English to German (de), Italian (it), and Chinese (zh) for S2TT, multilingual SQA, and multilingual S2TSUM. 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 audio in inglese”[it], “Translate this audio into English”[en]).

## 9.2 Data and Metrics

**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<sup>55</sup> and Large Language Model<sup>56</sup>, training their systems on designated datasets:

- EuroParl-ST (Iranzo-Sánchez et al., 2020) and CoVoST2 (Wang et al., 2020) for ASR/S2TT<sup>57</sup> tasks,
- Spoken-SQuAD (Li et al., 2018) for SQA,
- NUTSHELL (Züfle et al., 2025) for S2TSUM.

Development data is provided through the ACL 60/60 dataset (Salesky et al., 2023), which contains transcripts, translations, and summaries that can be retrieved using video IDs. Importantly, 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

<sup>55</sup>[hf.co/facebook/seamless-m4t-v2-large](https://hf.co/facebook/seamless-m4t-v2-large)

<sup>56</sup>[hf.co/meta-llama/Llama-3.1-8B-Instruct](https://hf.co/meta-llama/Llama-3.1-8B-Instruct)

<sup>57</sup>EuroParl-ST: en→{it, de}, CoVoST2: en→{zh, de}

models. No training data is provided for cross-lingual SQA or S2TSUM 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.

**Evaluation Data.** We evaluate the submitted models with IWSLT25Instruct, a novel resource, representing the first cross-lingual multimodal benchmark for instruction-following tasks across speech, text, and vision modalities in four languages: English, German, Italian, and Chinese. IWSLT25Instruct is extracted from the ASR, S2TT, SQA, and S2TSUM sections of the MMIF benchmark (Papi et al., 2025b), built upon scientific domain data retrieved from the ACL Anthology.<sup>58</sup> The dataset contains 21 videos, corresponding to 2 hours. Source audio and video content in English (talks of about 5-6 minutes each) are enriched with multilingual annotations and translations to support: *i*) ASR (en→en); *ii*) S2TT (en→de, it, zh), *iii*) S2TSUM (en→es, de, it, zh); *iv*) SQA (en→es, de, it, zh). In SQA, questions (about 10 for each video) are provided both in the speech language (English) and in other target languages (German, Italian, Chinese), and answers must be given in the same language as the one of the question (e.g., Italian questions require answers in Italian). The SQA task includes unanswerable questions, to which the only correct response is “*Not answerable*” or its corresponding translations in the other languages.<sup>59</sup> For S2TSUM, the dataset contains 100 abstracts (including those of the 21 videos), for a total of 17k words. The audio data are provided as complete audio files (5-6 minutes, 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.

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.

**Metrics.** The evaluation was carried out by computing separate scores for each of the tasks involved. Namely, for ASR, we computed WER using the jiWER library<sup>60</sup> after normalizing the test using the Whisper normalizer<sup>61</sup> (Radford et al., 2022). For S2TT, we used COMET<sup>62</sup> (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. Lastly, 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).<sup>63</sup> The code used for the evaluation is available at: [github.com/hlt-mt/if-iwslt2025](https://github.com/hlt-mt/if-iwslt2025).

### 9.3 Submissions

In total, we received 16 submissions from 5 different teams. Two teams submitted under the constrained setting. Only one submission was contrastive. Two teams (NLE and KIT) participated in all language directions, while others (CUNI-NL and IST) submitted for a subset. One team (MEETWEEN) submitted for English only. The participants’ systems in the SHORT (CUNI-NL, IST, MEETWEEN, NLE) and LONG (KIT) sub-tracks are detailed below.

**CUNI-NL** (Luu and Bojar, 2025) participated in the unconstrained LONG sub-track, submitting to ASR (en→en) and S2TT (en→de). Their submission explores the combination of speech encoders and instruction-tuned LLMs. Specifically, they compare Whisper and Seamless as encoders, alongside LLaMA, EuroLLM-9B-Instruct (Martins et al., 2024), and Gemma-3-12B-IT (Team

<sup>58</sup>[aclanthology.org](https://aclanthology.org)

<sup>59</sup>Namely, in Italian “*Non è possibile rispondere*”, German “*Nicht zu beantworten.*”, and Chinese 无法回答。

<sup>60</sup>[github.com/jitsi/jiwer](https://github.com/jitsi/jiwer)

<sup>61</sup>Specifically, we used version 0.0.10.

<sup>62</sup>With model `Unbabel/wmt22-comet-da`.

<sup>63</sup>See [github.com/Tiiiger/bert\\_score/blob/master/journal/rescale\\_baseline.md](https://github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md)



et al., 2025) as LLMs. For Seamless, the original length adapter is used, while for Whisper, a convolution-based length adapter is applied. A trainable feed-forward projection connects the frozen encoder with the frozen LLM, and LoRA adapters (Hu et al., 2021) are applied on top of the LLM. Training is conducted exclusively on the CoVoST dataset. Their results show that combining Seamless as the encoder with EuroLLM as the LLM yields the strongest performance.

**IST** (Attanasio et al., 2025) participated in the SHORT unconstrained sub-track, submitting to the en→en, de, and zh language pairs. Their system adapts small language models: audio is encoded with wav2vec 2.0 (Baevski et al., 2020), and a two-layer MLP projects features into the input space of a frozen Qwen2.5–1.5B (Qwen et al., 2025). Seven ASR datasets are used, along with CoVoST2 for S2TT, and Spoken-SQuAD for SQA. To increase coverage, ASR transcripts and Spoken-SQuAD are translated into German and Chinese using multiple LLMs and unanswerable questions are synthesized to improve SQA robustness. Task and language tags are prepended to prompts to enable multilingual, multitask instruction following. Training then proceeds in two stages: first, the speech encoder and MLP are jointly trained on ASR data for modality alignment; then, the encoder is frozen and only the MLP is fine-tuned on ASR, AST, and SQA.

**MEETWEEN** participated in the SHORT unconstrained sub-track, submitting to the ASR and SQA tasks. The system<sup>64</sup> combines the Seamless speech encoder with a Q-Former (Li et al., 2023; Tang et al., 2024) modality adapter and a LLaMA decoder. Training is performed in three stages. In the first stage, an ASR warmup is conducted with the encoder and LLM frozen and only the modality adapter is trained. The second stage, all-task warmup, retains the frozen encoder and LLM while training the adapter across ASR, S2TT, SQA, S2TSUM, MT, SLU, and lip reading tasks. Finally, in end-to-end training, the encoder remains frozen while both the adapter and LLM are fine-tuned on the same set of tasks.

**NLE** (Lee et al., 2025) participated in the SHORT constrained sub-track, submitting to all language pairs: en→en, de, it, zh. They augmented training data by translating SpokenSQuAD and

generating more fluent, abstractive answers. Their model employs a Seamless encoder with additional downsampling, a Transformer-based projection module, and LLaMA with LoRA (Hu et al., 2021) applied. Training occurs in three stages using two-level sampling process (Zanon Boito et al., 2024): first, the projector is trained with frozen encoder and LLM on ASR+ST or ASR+ST+SQA data; second, LoRA adapters are trained on the LLM using text-only MT and QA data; finally, both are jointly fine-tuned on all tasks for 1000 steps, with strong performance evident after 100 steps. Models trained with SQA in stage two initially underperform on SQA. However, after final tuning, all models perform similarly, with those trained only on ASR and S2TT slightly better on S2TT.

**KIT** Koneru et al. (2025) participated in the LONG constrained sub-track, submitting to all language pairs: en→en, de, it, zh. They augmented data by synthesizing NUTSHELL speech with TTS for ASR adaptation, and using LLaMA to generate multilingual QA pairs and translated summaries from NUTSHELL for SQA and S2TSUM. Their architecture connects Seamless and LLaMA via a trainable Q-Former (Li et al., 2023; Tang et al., 2024). Training involved contrastive pretraining (Züfle and Niehues, 2025) on ASR data followed by task-specific fine-tuning. Chain-of-thought reasoning was applied to improve SQA robustness by detecting unanswerable questions. For long audio, VAD-based segmentation (Sohn et al., 1999) was used in ASR and S2TT. For SQA and S2TSUM, audio segments were encoded separately, with embeddings concatenated before projection and LLM input to maintain end-to-end trainability. A context-aware post-editing model trained on NUTSHELL TTS data improved domain-specific terminology and restored context lost to segmentation.

## 9.4 Results

### 9.4.1 Automatic Evaluation

The complete results for both SHORT and LONG sub-tracks are presented in Table 47. For comparison, we include the results of the Phi4-Multimodal model (Abouelenin et al., 2025), a state-of-the-art baseline model trained on a broader range of tasks (including the IF task) and datasets (both in-house and public).

<sup>64</sup>[huggingface.co/meetweeen/Llama-speechlmm-1.0-l](https://huggingface.co/meetweeen/Llama-speechlmm-1.0-l)

**Monolingual English.** In the monolingual scenario—comprising ASR and SQA in the *SHORT* sub-track, and ASR, SQA, and S2TSUM in the *LONG* sub-track—all participating teams submitted systems, including a contrastive submission (CUNI-NL). In the *SHORT* sub-track, the best ASR performance is achieved by the baseline (7 WER). Among participants, NLE obtains the best result (13 WER), followed by CUNI-NL and IST, both with 15 WER. For SQA, NLE outperforms all other systems with a BERTScore of 0.50—exceeding the baseline by 0.04 points. Notably, the NLE’s system, even if trained in the constrained settings, still emerged as the top-performing participant, though it lagged behind the baseline in ASR by nearly double the WER. In the *LONG* sub-track, KIT, which is the only team that submitted a system, is able to outperform the baseline in two out of three tasks (ASR and S2TSUM), and its SQA performance (0.41 BERTScore) is nearly on par with the baseline (0.42). Nonetheless, there remains a performance gap compared to short-form processing: for example, the constrained systems NLE (*SHORT*) and KIT (*LONG*) differ by 0.02 WER in ASR and 0.08 BERTScore in SQA.

**Crosslingual German.** In the English-to-German (en-de) direction, the best S2TT result in the *SHORT* sub-track is achieved by the baseline (0.77 COMET). Among participants, CUNI-NL’s primary submission (0.72 COMET), NLE (0.71), and CUNI-NL’s contrastive (0.69) perform similarly. For SQA, NLE achieves the best score, surpassing the baseline by 0.02 BERTScore. In the *LONG* sub-track, KIT outperforms the baseline in all three tasks (ST, SQA, and S2TSUM), with substantial margins in some cases (e.g., 0.74 vs. 0.55 COMET in S2TT). While short-form processing remains easier for current systems, the gap is smaller in this case, with the constrained NLE system achieving only 0.03 COMET improvement on S2TT and 0.03 BERTScore in SQA compared to the constrained KIT.

**Crosslingual Italian.** In the English-to-Italian (en-it) direction, the baseline again achieves the best S2TT result in the *SHORT* sub-track, outperforming the only participant (NLE) by 0.06 COMET. However, NLE surpasses the baseline in SQA with a 0.02 BERTScore improvement. In the *LONG* sub-track, KIT outperforms the base-

line across all three tasks, including a large gain of 0.21 COMET in S2TT. As with other language directions, performance on long-form input remains consistently lower than short-form.

**Crosslingual Chinese.** In the English-to-Chinese (en-zh) direction, the baseline also leads in S2TT for the *SHORT* sub-track, outperforming NLE—the best-performing participant—by 0.05 COMET. For SQA, however, NLE achieves a 0.02 BERTScore improvement over the baseline. In the *LONG* sub-track, KIT once again outperforms the baseline and, interestingly, achieves better performance in long-form SQA (0.41) than those obtained by NLE in the short-form SQA (0.35), suggesting that the system was able to effectively exploit the long context.

#### 9.4.2 Human Evaluation

Similar to the other tracks of this year’s IWSLT Evaluation Campaign, each participant’s primary submission<sup>65</sup> has been manually evaluated. The human evaluation involves the speech translation outputs in German and Chinese, and the manual process that has been conducted is explained in Appendix A. The results are also compared with those of the other tracks in Table 15 and Table 17.

The human evaluation results largely confirm the trends observed in automatic evaluation. For en-de, the top-ranked KIT system (with a COMET score of 0.74) achieved the best human-evaluated performance, followed by CUNI primary and NLE (with a COMET of 0.72 and 0.71, respectively). However, human evaluators found the second and third-ranked systems indistinguishable, suggesting that COMET score differences of 0.01 fall below the threshold of human perceptual sensitivity. Similarly, for en-zh, the KIT and NLE systems were perceived as equivalent by humans, confirming their close automatic scores (of 0.77 and 0.76, respectively). Compared to the other tracks, the IF track results align with expectations, performing worse than the systems of the offline track but better than those of the simultaneous track, especially under low-latency constraints. This performance reflects two key factors: offline and simultaneous tracks’ systems benefit from larger training datasets and task-specific optimization for speech

<sup>65</sup>We have excluded from the human evaluations the submissions with COMET scores below 0.4, as they were significantly worse than other participants, making the comparison meaningless.

translation, while IF models are more general-purpose architectures, supporting multiple tasks. These findings highlight that while automatic metrics provide valuable performance insights, human perception may be less sensitive to small metric differences, particularly when systems achieve relatively high performance levels.

## 9.5 Discussion and Conclusions

As this was the first edition of the Instruction-Following (IF) shared task at IWSLT, our primary goal was to understand the interest of our community in evaluating general-purpose speech models across a variety of tasks and languages, and explore the feasibility of leveraging these models for long-form speech processing. The task was met with strong interest, with 16 submissions from 5 teams, and provided valuable insights into the current capabilities and limitations of IF systems for speech-based tasks.

Among the four tasks, ASR emerged as the most accessible, with most participants achieving a WER below 18. Monolingual SQA was also relatively approachable, with BERTScores up to 0.50. In contrast, crosslingual SQA proved more challenging, with best-case BERTScores between 0.38 and 0.41. The S2TT task showed consistent translation quality across language pairs, with best COMET scores ranging from 0.74 to 0.77. S2TSUM, however, stood out as the most difficult task, with no system exceeding a score of 0.37—even in the best case (en-zh).

Comparing performance across tracks, short-form processing (SHORT) consistently outperformed long-form (LONG) processing in all languages. Surprisingly, the difference appears to be more pronounced in the monolingual tasks instead of the crosslingual tasks, which are inherently more difficult, suggesting that ASR and monolingual SQA are better mastered by current short-form models. It is also noteworthy that the best results in both tracks were achieved by systems trained under constrained settings, demonstrating that these settings represent a promising *starter pack* for IF model development, allowing for building competitive systems even with limited resources.

In terms of top-performing systems, NLE’s submissions led the SHORT track across all language directions. In the LONG track, the KIT system—despite being the only submission—outperformed

the state-of-the-art Phi4-Multimodal baseline in nearly every task.

Given the success of this first edition and the encouraging level of participation, we plan to continue the IF shared task in future editions of IWSLT, expanding its scope and challenges to further advance research in speech processing.

## Acknowledgements

We gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018002. The work by FBK has received funding from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU, and from the European Union’s Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BetWEEN People). The work by Charles University received funding from the Project OP JAK Mezišektorová spolupráce Nr. CZ.02.01.01/00/23\_020/0008518 named “Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím” (Ondřej Bojar), from the grant 272323 of the Grant Agency of Charles University (Dávid Javorský), and the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO (Peter Polák). The work was also supported by the internal university grant number 260 821 (SVV). The work by Mateusz Krubiński and Pavel Pecina was funded by the European Commission via its H2020 Program (contract no. 870930: WELCOME).

Atul Kr. Ojha and John P. McCrae would like to thank Research Ireland under Grant Number SFI/12/RC/2289\_P2 Insight\_2 and thank RTÉ/TG4 for sharing the Irish speech data. We would also like to thank Panlingua for providing the Marathi-Hindi and Bhojpuri-Hindi speech translation data.

The work by Tsz Kin Lam was funded by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee (grant number 10039436: UTTER)

## References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. [Crowdsourcing speech data for low-resource languages from low-income workers](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826, Marseille, France. European Language Resources Association.
- 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, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 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.
- Victor Agostinelli, Max Wild, Matthew Raffel, Kazi Fuad, and Lizhong Chen. 2024. [Simul-LLM: A framework for exploring high-quality simultaneous translation with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10530–10541, Bangkok, Thailand. 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, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemanek, and Rodolfo Zevallos. 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.
- 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, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydin, and Marcos Zampieri. 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.
- Bhavana Akkiraju, Aishwarya Pothula, Santosh Kesiraju, and Anil Vuppala. 2025. IIITH-BUT system for IWSLT 2025 low-resource Bhojpur to Hindi speech translation. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Jesujoba O Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2024. Afrihubert: A self-supervised speech representation model for african languages. *arXiv preprint arXiv:2409.20201*.
- Tanel Alumäe, Joonas Kalda, Külli Bode, and Martin Kaitsa. 2023. [Automatic closed captioning for Estonian live broadcasts](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 492–499, Tórshavn, Faroe Islands. University of Tartu Library.
- Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTeV: Comprehensive Evaluation of Spoken Language Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Demo Papers*, Kyiv, Ukraine. 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. 2020a. Common voice: A massively-multilingual speech corpus. In *Lrec*.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020b. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4218–4222.
- Giuseppe Attanasio, Sonal Sannigrahi, Ben Peters, and André F.T. Martins. 2025. IST at IWSLT 2025: Multilingual Efficient Learning for Speech-Text Models. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Marko Avila and Josep Crego. 2025. SYSTRAN @ IWSLT 2025 Low-resource track. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.
- Bbc. 2019. [BBC Subtitle Guidelines](#). BBC © 2018 Version 1.1.8.
- Peter J Bickel and Kjell A Doksum. 1977. *Mathematical statistics: basic ideas and selected topics*. Holden-Day Inc.
- Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *Isi-nlp 2*, page 21.
- 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.
- Chaimae Chellaf, Haroun Elleuch, Othman Istaiteh, D. Fortuné Kponou, Fethi Bougares, Yannick Estève, and salima Mdhaifar. 2025. LIA and ELYA-DATA systems for the IWSLT 2025 low-resource speech translation shared task. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. [BLASER: A text-free speech-to-speech translation evaluation metric](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. 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](#).
- Sayan Das, Soham Chaudhuri, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. IWSLT 2025 Indic Track System Description Paper: Speech-to-Text Translation from Low-Resource Indian Languages (Bengali and Tamil) to English. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, Nips ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. [A statistical analysis of summarization evaluation metrics using resampling methods](#). *Transactions of the Association for Computational Linguistics*, 9:1132–1146.
- Debjit Dhar, Soham Lahiri, Tapabrata Mondal, and Sivaji Bandyopadhyay. 2025. JU-CSE-NLP’s Cascaded Speech to Text Translation Systems for IWSLT 2025 in Indic Track. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- David Draper, James S Hodges, Colin L Mallows, and Daryl Pregibon. 1993. Exchangeability and data analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 156(1):9–28.
- John E. Ortega, Rodolfo Joel Zevallos, Ibrahim Said Ahmad, and William Chen. 2024. [QUESPA submission for the IWSLT 2024 dialectal and low-resource speech translation task](#). In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 125–133,

- Bangkok, Thailand (in-person and online). Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Muhammad Hazim Al Farouq, Aman Kassahun Wassie, and Yasmin Moslem. 2025. Bemba Speech Translation: Exploring a Low-Resource African Language. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- RA Fisher. 1935. *The design of experiments*. Oliver & Boyd.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. [Are LLMs breaking MT metrics? results of the WMT24 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. [Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.
- 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.
- Phillip Good. 2002. Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, 1:243–247.
- Phillip Good. 2013. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.
- 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.
- Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmongkol Sarin, and Knot Pipatsrisawat. 2020. [Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503, Marseille, France. European Language Resources Association.
- 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*.
- Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. [Masri-headset: A maltese corpus for speech recognition](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- 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.
- Dávid Javorský, Dominik Macháček, and Ondřej Bojar. 2022. [Continuous rating as reliable human evaluation of simultaneous speech translation](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 154–164, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Xilin Jiang, Yinghao Aaron Li, Adrian Nicolas Florea, Cong Han, and Nima Mesgarani. 2024. Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis. *arXiv preprint arXiv:2407.09732*.
- Japan Translation Federation JTF. 2018. [JTF Translation Quality Evaluation Guidelines, 1st Edition \(in Japanese\)](#).
- Waad Ben Kheder, Josef Jon, André Beyer, Abdel Mes-saoudi, Rabea Affan, Claude Barras, Maxim Tychonov, and Jean-Luc Gauvain. 2024. ALADAN at IWSLT24 Low-resource Arabic Dialectal Speech Translation Task. In *Proceedings of the 21th International Conference on Spoken Language Translation (IWSLT)*.
- Waad Ben Kheder, Josef Jon, André Beyer, Abdel Mes-saoudi, Rabea Affan, Claude Barras, Maxim Tychonov, and Jean-Luc Gauvain. 2025. ALADAN at IWSLT25 Low-resource Arabic Dialectal Speech Translation Task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1493–1504.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thammie Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. [Findings of the 2023 conference on machine translation \(WMT23\): LLMs are here but not quite there yet](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.
- Sai Koneru, Maike Züfle, Thai-Binh Nguyen, Seymanur Akti, Jan Niehues, and Alexander Waibel. 2025. KIT’s Offline Speech Translation and Instruction Following Submission for IWSLT 2025. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Matyáš Kopp, Vladislav Stankov, Ondřej Bojar, Barbora Hladká, and Pavel Straňák. 2021. [ParCzech 3.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- D. Fortune Kponou, Salima Mdhaffar, Fréjus A. A. Laleye, Eugène C. Ezin, and Yannick Estève. 2025a. Extending the fongbe to french speech translation corpus: Resources, models and benchmark. In *Proceedings of Interspeech 2025*.
- D. Fortuné Kponou, Salima Mdhaffar, Fréjus A. A. Laleye, Eugène Cokou Ezin, and Yannick Estève. 2025b. FFSTC 2: Extending the Fongbe to French Speech Translation Corpus. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.
- 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*.
- Beomseok Lee, Marcely Zanon Boito, Laurent Besacier, and Ioan Calapodescu. 2025. NAVER LABS Europe Submission to the Instruction Following Track. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. [Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension](#). In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Hyderabad, India, September 2-6, 2018*, pages 3459–3463. Isca.
- 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.



- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning*.
- Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2025. KIT’s Low-resource Speech Translation Systems for IWSLT2025: System Enhancement with Synthetic Data and Model Regularization. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. [OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. In *Proceedings of Interspeech 2020*, pages 3620–3624.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchart. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció*, 12:455–463.
- Nam Luu and Ondřej Bojar. 2025. CUNI-NL@IWSLT 2025: End-to-end Offline Speech Translation and Instruction Following with LLMs. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Edward Ma. 2019. Nlp augmentation. <https://github.com/makcedward/nlpaug>.
- 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.
- Xutai Ma, Mohammad Javad Dousti, Chaghan Wang, Jiatao Gu, and Juan Pino. 2020. [SIMULEVAL: An evaluation toolkit for simultaneous translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.
- Mohamed Maamouri, Tim Buckwalter, David Graff, and Hubert Jin. 2006. Levantine arabic qt training data set 5. *Speech Linguistic Data Consortium, Philadelphia*.
- Dominik Macháček and Peter Polák. 2025. Simultaneous Translation with Offline Speech and LLM Models in CUNI Submission to IWSLT 2025. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- John Makhoul, Bushra Zawaydeh, Frederick Choi, and David Stallard. 2005. Bbn/aur darpa babylon levantine arabic speech and transcripts. *Linguistic Data Consortium (LDC), LDC Catalog No.: LDC2005S08*.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Fayssé, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. [Euroollm: Multilingual language models for europe](#).
- 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.
- Salima Mdhaffar, Fethi Bougares, Renato De Mori, Salah Zaiem, Mirco Ravanelli, and Yannick Estève. 2024. Taric-slu: A tunisian benchmark dataset for spoken language understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15606–15616.
- Humaira Mehmood and Sadaf Abdul Rauf. 2025. Human-Evaluated Urdu-English Speech Corpus: Advancing Speech-to-Text for Low-Resource Languages. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Chutong Meng and Antonios Anastasopoulos. 2025. GMU Systems for the IWSLT 2025 Low-Resource Speech Translation Shared Task. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. [Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese](#). In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.
- Mohammad Mohammadamini, Daban Jaff, Sara Jamal, Ibrahim Ahmed, Hawkar Omar, Darya Sabr, Marie Tahon, and Antoine Laurent. 2025. Kuvost: A Large-Scale Human-Annotated English to Central Kurdish Speech Translation Dataset Driven from



- English Common Voice. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- 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.
- Yasmin Moslem. 2025. Efficient Speech Translation through Model Compression and Knowledge Distillation. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- 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, John Hoffman, Searley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint*.
- Michal Novák, Peter Polák, Kateřina Rysová, Magdaléna Rysová, and Ondřej Bojar. 2024. Towards automated spoken language assessment: A study of asr transcription of examinations for non-native speakers of czech.
- 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).
- John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.
- John E. Ortega, Rodolfo Joel Zevallos, William Chen, and Idris Abdulmumin. 2025. QUESPA Submission for the IWSLT 2025 Dialectal and Low-resource Speech Translation Task. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 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. CMU’s IWSLT 2024 Simultaneous Speech Translation System. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- 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.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.
- 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, Peter Polák, Dominik Macháček, and Ondřej Bojar. 2025a. [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, et al. 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. 2025b. MCIF: Multimodal Crosslingual Instruction-Following Benchmark from Scientific Talks.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

- Frithjof Petrick, Patrick Wilken, Evgeny Matusov, Nahuel Roselló, 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)*.
- Edwin James George Pitman. 1937. Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4(2):225–232.
- 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.
- Peter Polák, Brian Yan, Shinji Watanabe, Alex Waibel, and Ondřej Bojar. 2023. [Incremental blockwise beam search for simultaneous speech translation with controllable quality-latency tradeoff](#). In *Proc. INTERSPEECH 2023*, pages 3979–3983.
- Peter Polák and Ondřej Bojar. 2024. [Long-form end-to-end speech translation via latent alignment segmentation](#). In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1076–1082.
- Maja Popović. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaozheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.
- 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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#).
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust Speech Recognition via Large-Scale Weak Supervision](#).
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. Pmlr.
- Matthew Raffel, Victor Agostinelli, and Lizhong Chen. 2024. [Simultaneous masking, not prompting optimization: A paradigm shift in fine-tuning LLMs for simultaneous translation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18302–18314, Miami, Florida, USA. Association for Computational Linguistics.
- Matthew Raffel, Victor Agostinelli, and Lizhong Chen. 2025. BeaverTalk: Oregon State University’s IWSLT 2025 Simultaneous Speech Translation System. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, 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. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Nathaniel Romney Robinson, Niyati Bafna, Xiluo He, Tom Lupicki, Lavanya Shankar, Cihan Xiao, Qi Sun, Kenton Murray, and David Yarowsky. 2025. JHU IWSLT 2025 Low-resource System Description. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.

- Nathaniel Romney Robinson, Kaiser Sun, Cihan Xiao, Niyati Bafna, Weiting Tan, Haoran Xu, Henry Li Xinyuan, Ankur Kejriwal, Sanjeev Khudanpur, Kenton Murray, and Paul McNamee. 2024. [JHU IWSLT 2024 dialectal and low-resource system description](#). In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 140–153, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.
- Mukund K Roy, Karunesh K Arora, Praveen Kumar Chandaliya, Rohit Kumar, and Pruthwik Mishra. 2025. CDAC-SVNIT submission for IWSLT 2025 Indic track shared task. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalan 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, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. [Audiopalm: A large language model that can speak and listen](#).
- Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. [IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.
- Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. [Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors. 2024. *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*. Association for Computational Linguistics, Bangkok, Thailand (in-person and online).
- 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.
- Jorge Iranzo Sanchez, Jorge Civera Saiz, and Adrià Giménez Pastor. 2025. MLLP-VRain UPV System for the IWSLT 2025 Simultaneous Speech Translation Task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- Ashwin Sankar, Sparsh Jain, Nikhil Narasimhan, Devilal Choudhary, Dhairya Suman, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2025. [Towards building large scale datasets and state-of-the-art automatic speech translation systems for 13 Indian languages](#). In *The 63rd Annual Meeting of the Association for Computational Linguistics*.
- Sankalpa Sarkar, Samridhi Kashyap, Advait Joglekar, and Srinivasan Umesh. 2025. Effectively combining Phi-4 and NLLB for Spoken Language Translation: SPRING Lab IITM’s submission to Low Resource Multilingual Indic Track. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinash Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. [Seamlessm4t: Massively multilingual & multimodal machine translation](#).
- 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.
- Hashem Sellat, Shadi Saleh, Mateusz Krubiński, Adam Pospíšil, Petr Zemanek, and Pavel Pecina. 2023. [UFAL parallel corpus of north levantine 1.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.



- Nivedita Sethiya and Chandresh Kumar Maurya. 2024. [End-to-end speech-to-text translation: A survey](#). *Computer Speech & Language*, page 101751.
- Nivedita Sethiya, Saanvi Nair, and Chandresh Maurya. 2024. [Indic-TEDST: Datasets and baselines for low-resource speech to text translation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9019–9024, Torino, Italia. ELRA and ICCL.
- 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.
- Tiia Sildam, Andra Velve, and Tanel Alumäe. 2024. [Finetuning end-to-end models for Estonian conversational spoken language translation](#). In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 166–174, Bangkok, Thailand. Association for Computational Linguistics.
- Silero Team. 2021. [Silero vad: pre-trained enterprise-grade voice activity detector \(vad\), number detector and language classifier](#).
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. [A statistical model-based voice activity detection](#). *IEEE Signal Processing Letters*, 6(1):1–3.
- 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.
- Haotian Tan, Ruhayah Faradishi Widiaputri, Jan Meyer Saragih, Yuka Ko, Katsuhito Sudoh, Satoshi Nakamura, and Sakriani Sakti. 2025. [NAIST Simultaneous Speech Translation System for IWSLT 2025](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. [SALMONN: Towards generic hearing abilities for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culli-



- ton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shrivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#).
- 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).
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Filbert Aurelian Tjjaranata, Vallerie Alexandra Putra, Eryawan Presma Yulianrifat, and Ikhlul Akmal Hanif. 2025. BUINUS System Description for IWSLT 2025 Maltese to English Low-Resource Speech Translation Track. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Ioannis Tsiamas, José Fonollosa, and Marta Costajussà. 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. Costajussà. 2022. [SHAS: Approaching optimal Segmentation for End-to-End Speech Translation](#). In *Proc. Interspeech 2022*, pages 106–110.
- Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. [VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, 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.
- Minghan Wang, Thuy-Trang Vu, Yuxia Wang, Ehsan Shareghi, and Gholamreza Haffari. 2024a. [Conversational simulmt: Efficient simultaneous translation with large language models](#).
- Minghan Wang, Thuy-Trang Vu, Jinming Zhao, Fateh Shiri, Ehsan Shareghi, and Gholamreza Haffari. 2024b. [Simultaneous machine translation with large language models](#). In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 89–103, Canberra, Australia. Association for Computational Linguistics.
- Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. [CharacTer: Translation edit rate on character level](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Association for Computational Linguistics.
- Wenxuan Wang, Yingxin Zhang, Yifan Jin, Binbin Du, and Yuke Li. 2025. NYA’s Offline Speech Translation System for IWSLT 2025. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.
- Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. [ESPnet: End-to-end speech processing toolkit](#). In *Proceedings of Interspeech*, pages 2207–2211.
- Xuchen Wei, Yangxin Wu, Yaoyin Zhang, Henglyu Liu, Kehai Chen, Xuefeng Bai, and Min Zhang. 2025. HITSZ’s End-To-End Speech Translation Systems Combining Sequence-to-Sequence Auto Speech Recognition Model and Indic Large Language Model for IWSLT 2025 in Indic Track. In *Proceedings of the 22th International Conference on Spoken Language Translation (IWSLT)*.

- Ruhiyah Faradishi Widiaputri, Haotian Tan, Jan Meyer Saragih, Yuka Ko, Katsuhito Sudoh, Satoshi Nakamura, and Sakriani Sakti. 2025. NAIST Offline Speech Translation System for IWSLT 2025. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT)*.
- 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.
- Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. [Deco: Decoupling token compression from semantic abstraction in multimodal large language models](#).
- ZamStats. 2012. [2010 census of population and housing - national analytical report](#).
- Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. [mhubert-147: A compact multilingual hubert model](#). In *Interspeech 2024*, pages 3939–3943.
- 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*.
- Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025a. [How to select datapoints for efficient human evaluation of nlg models?](#)
- Vilém Zouhar, Maïke Züfle, Beni Egressy, Julius Cheng, and Jan Niehues. 2025b. [Early-exit and instant confidence translation quality estimation](#).
- Maïke 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.
- Maïke Züfle and Jan Niehues. 2025. [Contrastive learning for task-independent speechllm-pretraining](#).

## Appendix A. Human Evaluation

### A Human Evaluation

Human evaluation includes direct assessment for offline, simultaneous, subtitling, and instruction following tasks (A.1), in addition to continuous rating and MQM for the simultaneous task (A.2, A.3).

#### A.1 Direct Assessment

For the offline translation track (Section 3), simultaneous translation track (Section 4), subtitling track (Section 5), and instruction following track (Section 9), 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), expressed as scores ranging from 0 to 100. The *business news* test set does not include reference transcripts, so the human assessment is performed monolingually, comparing the system outputs against reference translations. We exclude the English to German direction from this test set for budget reasons. All other sets are graded in full, with no subsampling performed. No annotator normalization was performed this year.

Since many tasks have standardized their test sets, we evaluate all outputs for a given testset, across any task that used said testset. 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 have different objectives – for example, length in the case of subtitling and latency in the case of online systems. Additionally, in the case of the *business news* testset, we use the verbatim version of the reference; the subtitle systems would likely have been judged more favorably if we had instead used the more terse subtitle reference.

##### 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.<sup>66</sup> 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 sentence and system translation but also with the system translation of the previous and following segments. Annotators are then instructed as follows: “*Sentence boundary errors are expected and should not be taken into account when judging translation quality. This is when the target appears to be adding or missing words (including being completely empty) while the source was segmented in a different place. To this end, we have included the previous and next sentence targets for context. If the content of the source and target are only different because of sentence boundary issues, do not let this affect your scoring judgement.*”

*Example of a good translation (shown English-only for illustration purposes) suffering only from sentence boundary issues that should not be penalized:*

*Source: you’ll see that there’s actually a sign near the road.*

*Target: is a sign near the*

*Previous target: [...] and you will see that there actually Next target: road. [...]*

No video or audio context is provided. 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.

For monolingual grading (*business news* test set, English to Arabic), we add the following instruction: “*You’ll be shown a candidate translation from English into Arabic, while the “source” is the Arabic reference translation. Please rate the correctness of the candidate, given the reference..*”

##### A.1.2 Computing Pairwise Statistical Significance and System Rankings

Last year, we used the Wilcoxon rank-sum test (also called the Mann-Whitney U) to determine statistical significance of the human evaluation scores. The Wilcoxon rank-sum test is non-parametric, which is advantageous because DA scores do not follow a normal or other known distribution (see Figure 1).

<sup>66</sup>[www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz](http://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz)

However, the Wilcoxon rank-sum test also assumes independent samples, whereas our data samples are not in fact independent. This is because a given source sentence is translated by two or more MT systems and then those outputs are scored by a human annotator. We generally expect correlation between scores for the same source sentence (e.g. a source sentence which is very difficult to translate will likely result in lower than average scores for all MT systems).

An alternative to the Wilcoxon rank-sum test is the Wilcoxon signed-rank test, which assumes dependent (i.e. paired) data, but it adds an assumption that the distribution of scores is symmetric around a mean, which Figure 1 illustrates is not true in our case.

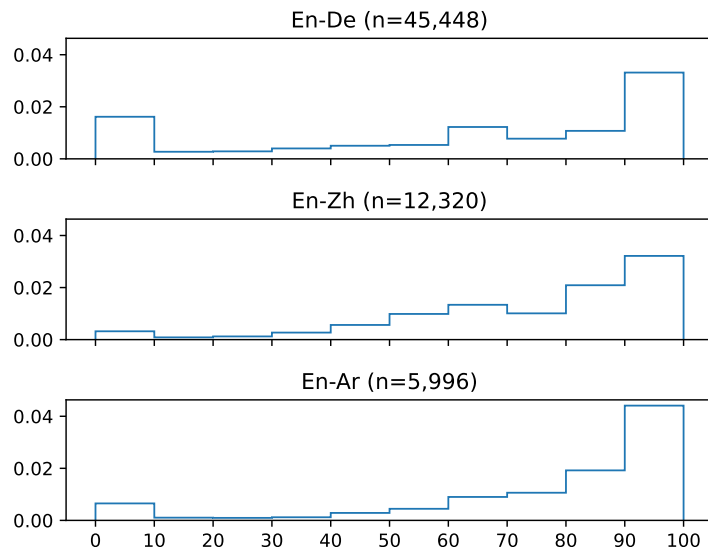


Figure 1: Direct assessment score histograms, normalized, per language pair.

This year, we chose to use a permutation test (Fisher, 1935) to estimate the statistical significance of the difference in the means of the segment-level DA scores for each pair of MT systems. Permutation tests are appealing because they don’t require any assumptions about the underlying distribution of the data. Instead, they have the assumption of exchangeability (Pitman, 1937; Draper et al., 1993; Good, 2002)—that is, under the null hypothesis (in our case, that the two MT systems are of equal quality) the joint distribution of the observations is invariant under permutations of the data labels. We first randomly split the segment-level scores (ignoring the labels, i.e. which MT system produced each segment) into two parts and compute the difference in DA score mean. Repeating this process many times provides a set of mean differences we can reasonably expect under the null hypothesis that the two systems are of the same quality. We compute a one-tailed  $p$ -value by calculating the fraction of the time that the random splits produce differences greater than or equal to the mean difference we observe for the two systems. To help ensure exchangeability, we perform permutations such that each split has exactly one translation of each test set sentence, commonly referred to as a paired permutation test (Good, 2013). In the context of machine translation, paired permutation tests are widely used in automatic metric evaluation (Deutsch et al., 2021; Freitag et al., 2023, 2024; Thompson et al., 2024). We use the paired permutation implementation from Thompson et al. (2024).<sup>67</sup>

Given the  $p$ -values from all pairwise system comparisons, system rankings are trivially computed by ordering the systems by mean DA score and then finding the rank of the highest and lowest ranked system(s) that are not statistically significantly different from each system. We use a 95% confidence (i.e.  $p$ -value  $< 0.05$ ).

English → Arabic, *business news* testset rankings are given in Table 13, with  $p$ -values in Figure 2. English → German, *accented English conversations* testset rankings are given in Table 14, with  $p$ -values in Figure 3. English → German, *scientific presentations* testset rankings are given in Table 15, with  $p$ -

<sup>67</sup>[github.com/thompsonb/mt-metrics-eval/blob/main/mt\\_metrics\\_eval/pairwise\\_paired\\_permutation\\_test.py](https://github.com/thompsonb/mt-metrics-eval/blob/main/mt_metrics_eval/pairwise_paired_permutation_test.py)



values in Figure 4. English → German, *TV series* testset rankings are given in Table 16, with  $p$ -values in Figure 5. English → Chinese, *scientific presentations* testset rankings are given in Table 17, with  $p$ -values in Figure 6.

As one would expect, we find that across all language pairs / test sets, offline systems tend to be the highest ranked, and high-latency online systems tend to rank higher than low-latency online systems.

Table 13: English → Arabic, *business news* testset. Human direct assessment scores and corresponding rankings. Rank range based on 95% confidence interval, from pairwise  $p$ -values in Figure 2.

Task	System	Data/Condition	Human Score	Human Rank
Offline	NYA	unconstrained	84.451	1
Offline	NEMO	unconstrained	82.017	2
Offline	AIB-MARCO	unconstrained	80.228	3
Subtitling	APPTek		60.524	4

Table 14: English → German, *accent English conversations* testset. Human direct assessment scores and corresponding rankings. Rank range based on 95% confidence interval, from pairwise  $p$ -values in Figure 3.

Task	System	Data/Condition	Human Score	Human Rank
Offline	KIT	unconstrained	74.865	1-2
Offline	AIB-MARCO	unconstrained	74.705	1-2
Offline	NYA	unconstrained	72.576	3-4
Offline	NEMO	unconstrained	72.298	3-4
Simultaneous	UPV	high	70.679	5-6
Simultaneous	OSU	high	70.372	5-6
Simultaneous	OSU	low	67.550	7
Offline	NAIST	unconstrained	63.622	8
Offline	NAIST	constrained	58.610	9
Offline	CUNI	constrained	51.407	10
Simultaneous	CMU	low	44.099	11

Table 15: English → German, *scientific presentations* testset. Human direct assessment scores and corresponding rankings. Rank range based on 95% confidence interval, from pairwise  $p$ -values in Figure 4.

Task	System	Data/Condition	Human Score	Human Rank
Offline	KIT	unconstrained	90.626	1
Offline	NEMO	unconstrained	86.583	2-4
Offline	NYA	unconstrained	86.536	2-4
Offline	AIB-MARCO	unconstrained	85.372	2-5
Simultaneous	CUNI	high	84.309	4-5
Simultaneous	UPV	high	78.662	6
Simultaneous	OSU	high	76.923	7-10
Instruction.long	KIT	primary	76.382	7-10
Offline	NAIST	unconstrained	75.432	7-11
Offline	CUNI	constrained	75.367	7-11
Simultaneous	OSU	low	74.397	9-11
Simultaneous	NAIST	high	71.166	12-15
Instruction.short	CUNI-NL	primary	70.702	12-15
Simultaneous	CMU	low	70.372	12-15
Instruction.short	NLE	primary	69.607	12-16
Offline	NAIST	constrained	67.801	15-18
Simultaneous	NAIST	low	67.197	16-18
Instruction.short	CUNI-NL	contrastive	66.280	16-18

Table 16: English → German, *TV series* testset. Human direct assessment scores and corresponding rankings. Rank range based on 95% confidence interval, from pairwise  $p$ -values in Figure 5.

Task	System	Data/Condition	Human Score	Human Rank
Offline	KIT	unconstrained	61.379	1
Offline	NYA	unconstrained	56.801	2-3
Offline	NEMO	unconstrained	56.395	2-3
Subtitling	APPTEK		53.992	4
Offline	CUNI	constrained	32.278	5
Offline	NAIST	unconstrained	27.174	6
Offline	NAIST	constrained	21.674	7
Offline	AIB-MARCO	unconstrained	12.122	8

Table 17: English  $\rightarrow$  Chinese, *scientific presentations* testset. Human direct assessment scores and corresponding rankings. Rank range based on 95% confidence interval, from pairwise  $p$ -values in Figure 6.

Task	System	Data/Condition	Human Score	Human Rank
Offline	AIB-MARCO	unconstrained	85.918	1
Offline	NYA	unconstrained	84.044	2-4
Offline	BIGWATERMELON	unconstrained	83.338	2-4
Offline	NEMO	unconstrained	83.009	2-4
Simultaneous	CUNI	high	77.805	5
Offline	NAIST	unconstrained	71.593	6-9
Instruction.short	NLE	primary	70.465	6-10
Instruction.long	KIT	primary	69.995	6-10
Simultaneous	CMU	low	69.812	6-10
Simultaneous	OSU	high	69.415	7-11
Simultaneous	NAIST	high	67.761	10-12
Simultaneous	OSU	low	67.519	11-12
Simultaneous	NAIST	low	65.487	13
Offline	NAIST	constrained	58.831	14

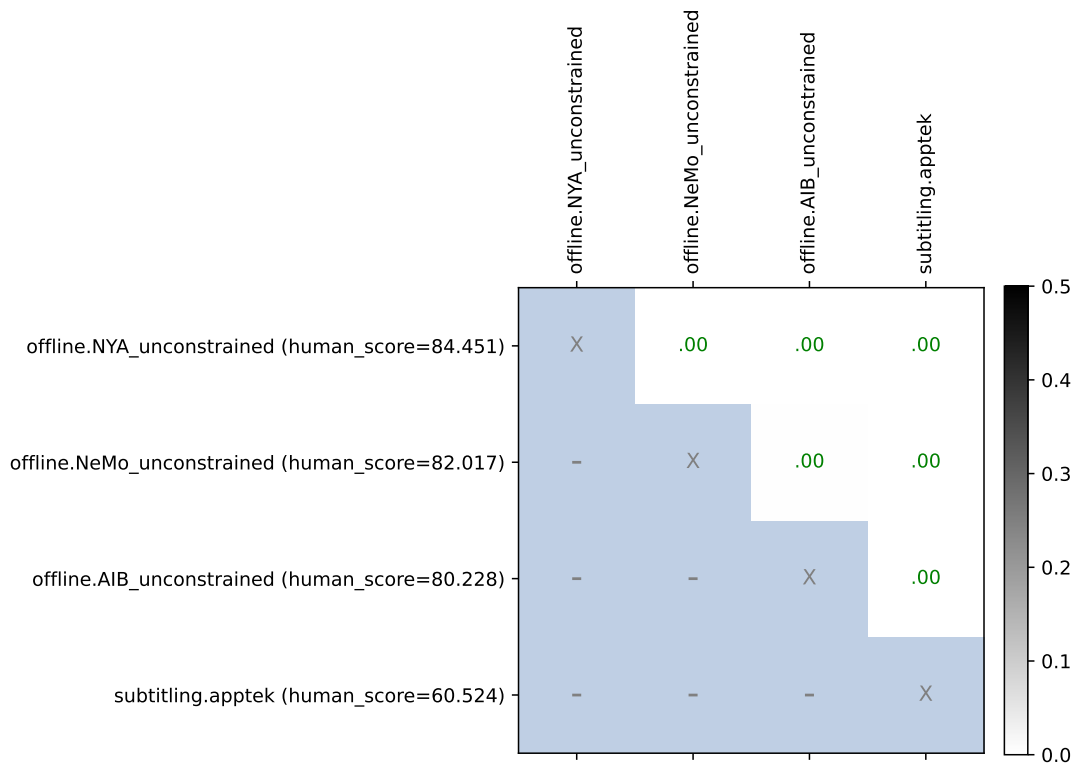


Figure 2: English  $\rightarrow$  Arabic, *business news* testset, pairwise  $p$ -values from paired permutation tests.  $p$ -values  $< 0.05$  are shown in green, while  $p$ -values  $\geq 0.05$  are shown in red. For system rankings computed from these  $p$ -values, see [Table 13](#).



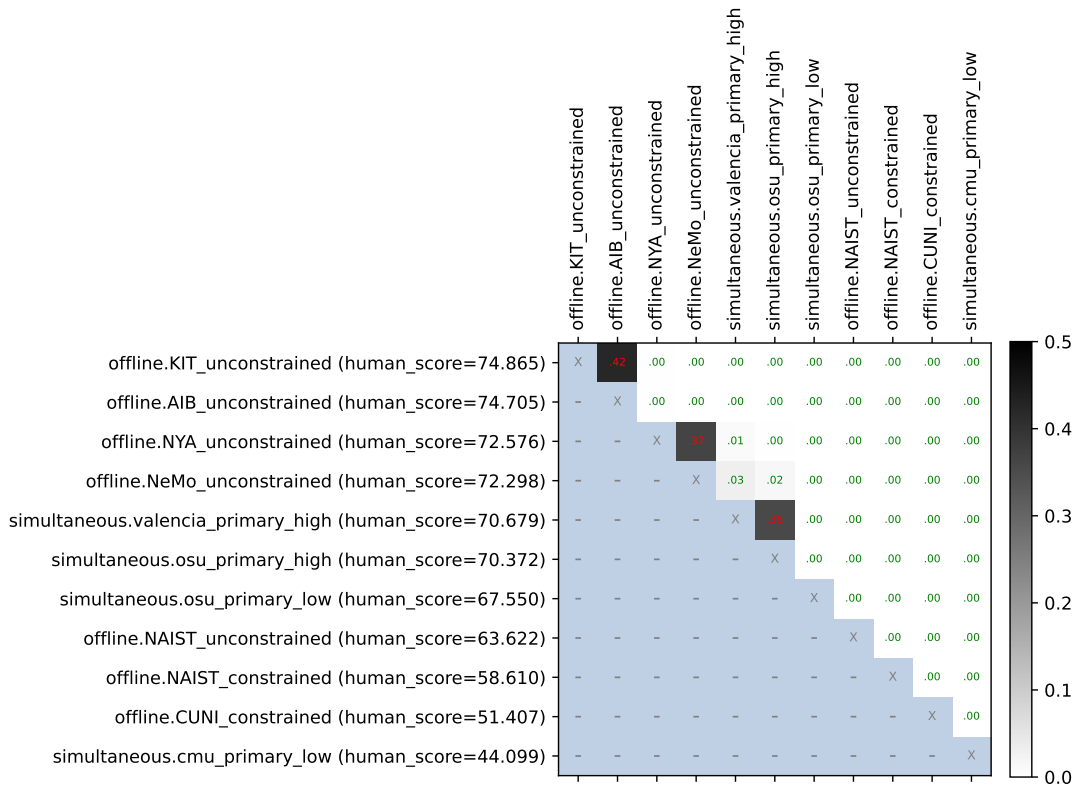


Figure 3: English → German, *accented English conversations* testset, pairwise  $p$ -values from paired permutation tests ( $n=10,000$ ).  $p$ -values  $< 0.05$  are shown in green, while  $p$ -values  $\geq 0.05$  are shown in red. For system rankings computed from these  $p$ -values, see Table 14.

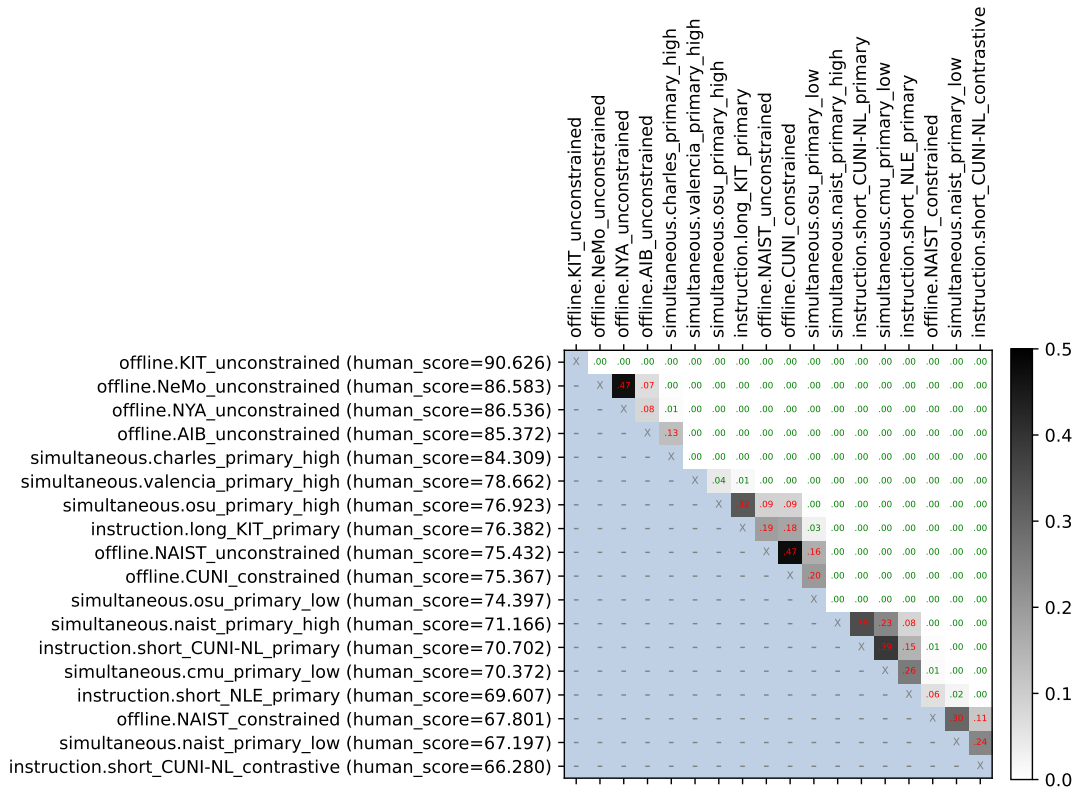


Figure 4: English → German, *scientific presentations* testset, pairwise  $p$ -values from paired permutation tests ( $n=10,000$ ).  $p$ -values  $< 0.05$  are shown in green, while  $p$ -values  $\geq 0.05$  are shown in red. For system rankings computed from these  $p$ -values, see Table 15.

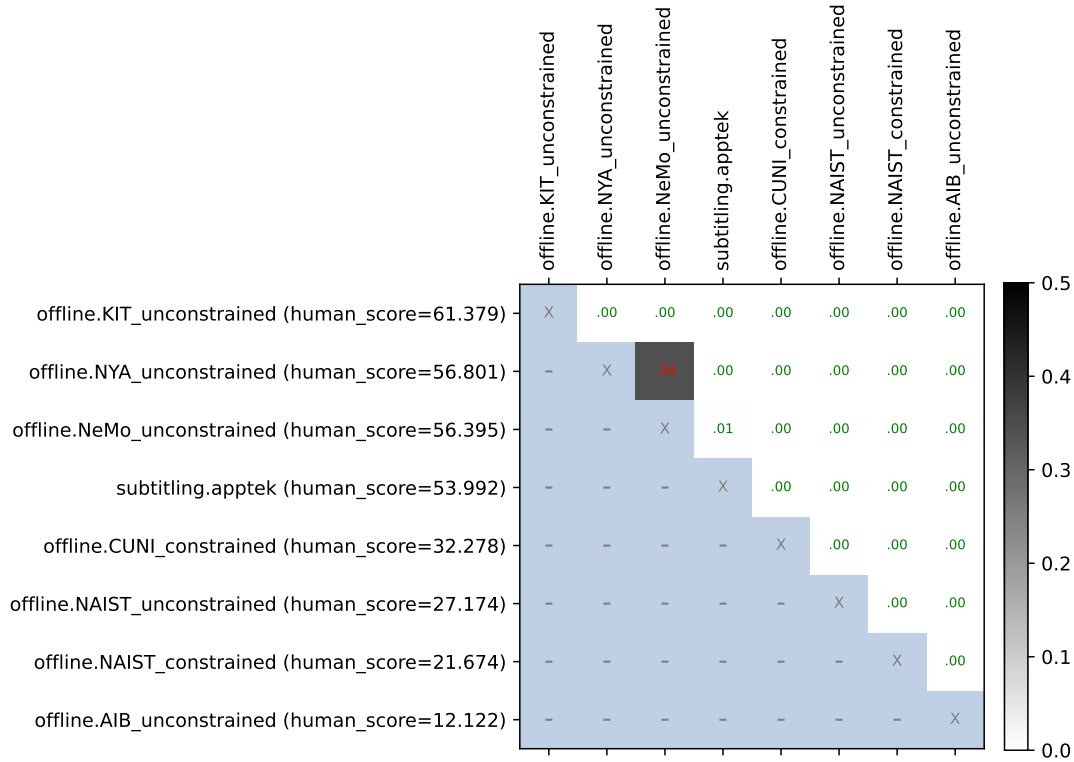


Figure 5: English → German, *TV series* testset, pairwise  $p$ -values from paired permutation tests ( $n=10,000$ ).  $p$ -values  $< 0.05$  are shown in green, while  $p$ -values  $\geq 0.05$  are shown in red. For system rankings computed from these  $p$ -values, see Table 16.

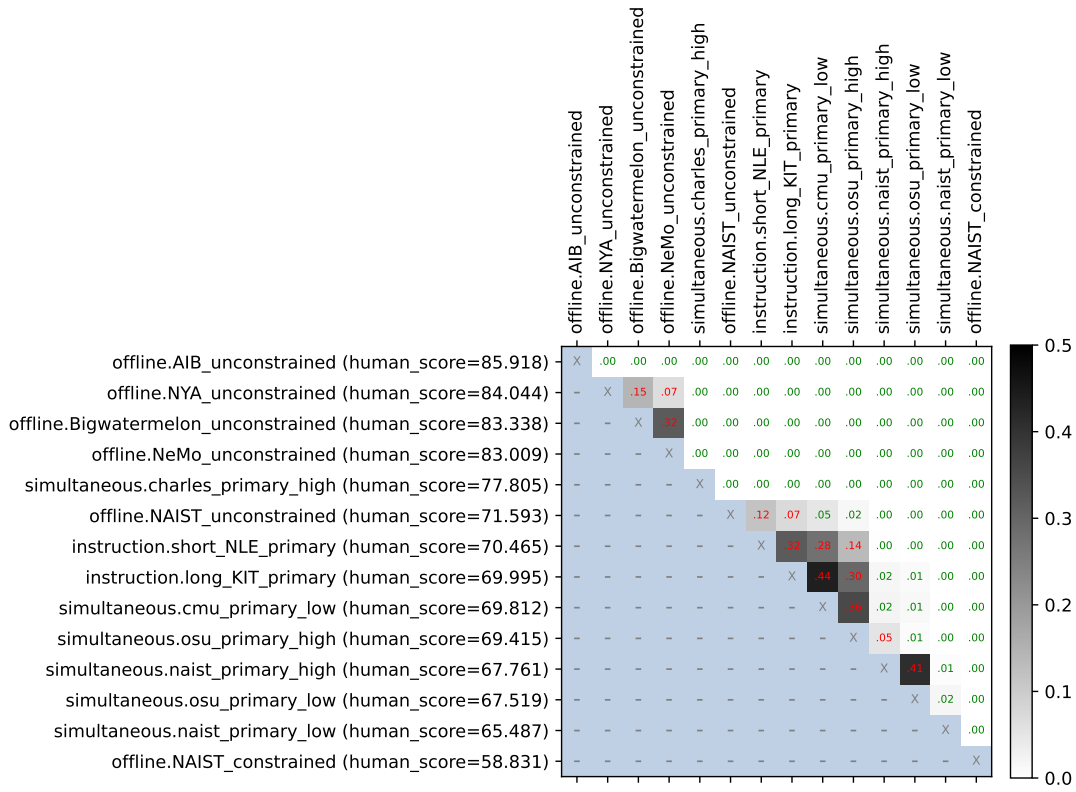


Figure 6: English  $\rightarrow$  Chinese, *scientific presentations* testset, pairwise  $p$ -values from paired permutation tests ( $n=10,000$ ).  $p$ -values  $< 0.05$  are shown in green, while  $p$ -values  $\geq 0.05$  are shown in red. For system rankings computed from these  $p$ -values, see Table 17.

### A.1.3 Deciding Which Segments to Human-Evaluate

Each year, the shared task size is limited by the amount that can be human-evaluated. Oftentimes, a random subset of segments is chosen for human evaluation to fit a specific budget. However, this uninformed selection might be suboptimal and previous works showed promise for efficient subset selection for machine translation and summarization. While the IWSLT 2025 evaluation has not used informed subset selection, this section investigates its potential for future IWSLT human evaluation campaigns.

**Setup.** Given a large set of evaluable items  $\mathcal{X}$ , the task is to select  $\mathcal{Y} \subseteq \mathcal{X}$  such that  $|\mathcal{Y}|$  fits a specific budget. Then, all systems participating in the shared task are evaluated on  $\mathcal{Y}$ . We consider the following methods for subset selection (Zouhar et al., 2025a):

- **Metric average:** Selecting examples with lowest average quality estimation scores across systems (highest difficulty). Based on [wmt22-cometkiwi-da](#) (Rei et al., 2022b).
- **Metric variance:** Selecting examples with largest variance among the quality estimation scores across systems. Same metric.
- **Metric consistency:** Selecting examples where the item-level metric ranking is predictive of the final aggregated system ranking. Same metric.
- **Diversity:** Selecting examples with which lead to most different system outputs (measured with pairwise ChrF).
- **K-means:** Selecting examples that are most dissimilar to each other (using k-means clustering).

We simulate the selection at a particular budget (subset size). We measure the success of subset selection in three ways. In all cases, the higher the better.

- **Cluster count:** Number of statistically significant clusters, as computed by [Kocmi et al. \(2023\)](#).
- **Kendall’s  $\tau_b$  rank correlation:** Similarity of final system ranking based on the subset and based on the full set.
- **Soft Pairwise Accuracy** (Thompson et al., 2024): Similarity of final system ranking based on the subset and based on the full set but with statistical significance taken into account.

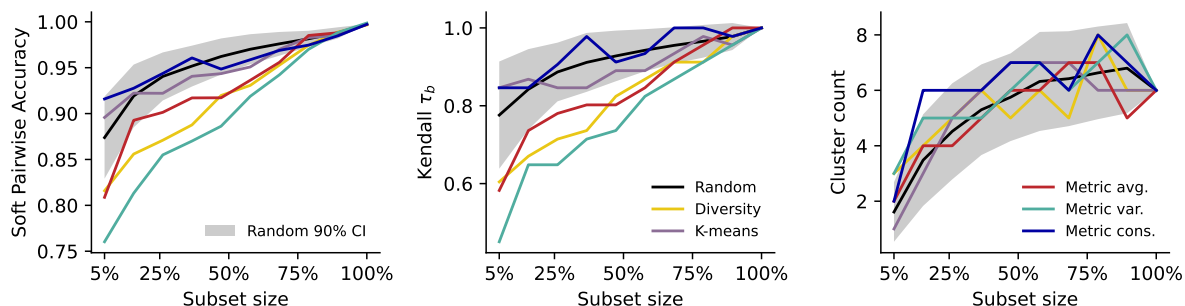


Figure 7: Results of informed subset selection for English→Arabic human-evaluated testset.

**Results.** The results in Figure 7 show that, by far, random selection remains the most robust selection, with metric consistency being on par to random when measured by soft pairwise accuracy or Kendall’s  $\tau_b$  and slightly better when measured by cluster count.

This can be partly explained by the evaluation segments not being aligned. For other tasks, such as text-to-text machine translation, given a single input, the systems produce outputs that can be compared to each other. In the current speech translation setup, the segmentation makes it so that segments with the same force-aligned source have very different outputs across systems. For example, the following are the 8 system translations aligned to the same source segment “*There were tough fights, even blood flowed.*”, which sometimes include non-relevant content, likely from previous or subsequent segments:

*viel Geld verschwendet. Scham!  
zu Besuch Sie kommen heute Nacht zu uns  
Harte Kämpfe. Es wurde Blut vergossen.  
Schwer erkämpfte Kämpfe. Es wurde Blut vergossen.*

*hart umkämpfte Schlachten, Blut wurde vergossen. Ich schäme  
Hart gekämpfte Schlachten. Das Blut wurde vergossen,  
Hart gekämpfte Schlachten. Das Blut wurde vergossen,  
war glücklich. Vier Schlachten. Es wurde geschrieben. Schande!*



Comparing the segment-level quality estimation using automatic metrics (necessary for metric average, metric variance, and metric consistency) then becomes difficult. The primary noise for the metrics comes from noisy prefixes and suffixes. Some metrics, such as [COMET-partial](#) ([Zouhar et al., 2025b](#), Appendix G) show promise to this kind of noise, though do not improve meaningfully the subset selection. The biggest hurdle to informed subset selection is thus a better alignment of system output, or the selection at higher-level units where the alignment is implicitly correct, such as the level of documents or whole audio files.

## A.2 Continuous Rating for Czech-to-English and English-to-German

Manual evaluation of English-to-German Simultaneous Task uses Continuous Rating as described by [Javorský et al. \(2022\)](#).

For both translation directions (Czech-to-English and English-to-German), we solicited students of translation studies from the Faculty of Arts, Charles University, as evaluators. All were native speakers of Czech, studying for English and (those evaluating German) also German translation.

During the evaluation, annotators were presented with the source audio and subtitles. The subtitles were displayed in two lines below the audio following the guidelines for video subtitling ([Bbc, 2019](#)). The annotators were asked to score the quality of the live-presented text output while listening to the input sound. Specifically, the instructions explicitly asked to focus on *content preservation*, or roughly the *adequacy*:

- We ask you to provide your assessment using so-called “continuous rating”, which *continuously indicates the quality of the text output given the input utterance you hear* in the range from 1 (the worst) to 4 (the best) by clicking the corresponding buttons or pressing the corresponding keys.
- The rate of clicking/pressing depends on you. However, we suggest clicking *each 5-10 seconds* or when your assessment has changed. We encourage you to provide feedback *as often as possible* even if your assessment has *not changed*.
- The quality scale should reflect primarily the meaning preservation (i.e. evaluating primarily the “content” or very approximately the “adequacy”) and the grammaticality and other qualitative aspects like punctuation (i.e. the “form” or extremely roughly the “fluency”) should be the secondary criterion.

**Processing of Collected Rankings** Once the results are collected, they are processed as follows. We first inspect the timestamps on the ratings, and remove any that appeared more than 20 seconds than the end of the audio. Because of the natural delay and because the collection process is subject to network and computational constraints, there can be ratings that are timestamped greater than the audio length. If the difference is however too high, we judge it to be an annotation error. We also remove any annotated audio where there is fewer than one rating per 20 seconds because the annotators were instructed to annotate every 5-10 seconds.

**Obtaining Final Scores** To calculate the final score for each system, we average the ratings across each annotated audio, then average across all the annotated audios pertaining to each system-latency combination. This type of averaging renders all input speeches equally important and it is not affected by the speech length or the eagerness of the annotator.

The final scores for Czech-to-English and English-to-German are provided further below in [Tables 18](#) and [19](#), respectively.

## A.3 MQM-based Human Evaluation for English-to-Japanese

For the English-to-Japanese Simultaneous Translation Task, we conducted a human evaluation using a variant of Multidimensional Quality Metrics (MQM; [Lommel et al., 2014](#)). MQM has been used in recent MT evaluation studies ([Freitag et al., 2021a](#)) and WMT Metrics shared task ([Freitag et al., 2021b](#)). For the evaluation of Japanese translations, we used *JTF Translation Quality Evaluation Guidelines* ([JTF, 2018](#)), distributed by Japan Translation Federation (JTF). The guidelines are based on MQM but include some modifications in consideration of the property of the Japanese language.

We hired a Japanese-native professional interpreter as the evaluator. The evaluator checked translation hypotheses along with their source speech transcripts and chose the corresponding error category and

severity for each translation hypothesis on a spreadsheet. Here, we asked the evaluator to focus only on *Accuracy* and *Fluency* errors, because other types of errors in Terminology, Style, and Locale convention would not be so serious in the evaluation of simultaneous translation. Finally, we calculated the cumulative error score for each system based on the error weighting presented by Freitag et al. (2021a), where *Critical* and *Major* errors have the same level of error scores. The results are shown in Table 20.

Test Set Part	Team	CR (↑)	BLEU (↑)	StreamLAAL (↓)
Non-Native	Baseline-VAD	2.34	16.46	1.85
	CUNI	3.02	24.53	1.79
ParCzech	Baseline-VAD	3.04	23.55	3.68
	Interpreting-Student	3.35	11.31	4.34
	CUNI	3.36	21.94	1.51
	Interpreting-Professional	3.51	10.09	4.16

Table 18: Human evaluation using Continuous Rating (CR) for systems from the high-latency regime of simultaneous speech-to-text translation contrasted with two variants of human interpreting, Czech-to-English. The Continuous Rating values range from 1 (worst) to 4 (best).

Team	CR (↑)	BLEU (↑)	COMET (↑)	StreamLAAL (↓)
Baseline-Fixed	3.02	19.15	0.593	3.54
NAIST	3.25	24.58	0.717	3.71
CMU	3.39	22.63	0.697	1.47
OSU	3.56	25.80	0.729	3.21
UPV	3.63	29.81	0.739	2.90
CUNI	3.72	35.25	0.790	3.32

Table 19: Human evaluation using Continuous Rating (CR) for systems from the high-latency regime (except CMU which was only in low-latency regime) of simultaneous speech-to-text translation, English-to-German. The Continuous Rating values range from 1 (worst) to 4 (best).

System	BLEU (on two ACL talks)	Error score	# Errors		
			Critical	Major	Minor
CUNI	39.4	32	0	3	17
NAIST (high)	32.8	123	8	12	23
NAIST (low)	33.1	129	12	9	24

Table 20: Human evaluation results on two ACL talks (91 lines) in the English-to-Japanese Simultaneous speech-to-text translation task. Error weights are 5 for Critical and Major errors and 1 for Minor errors.

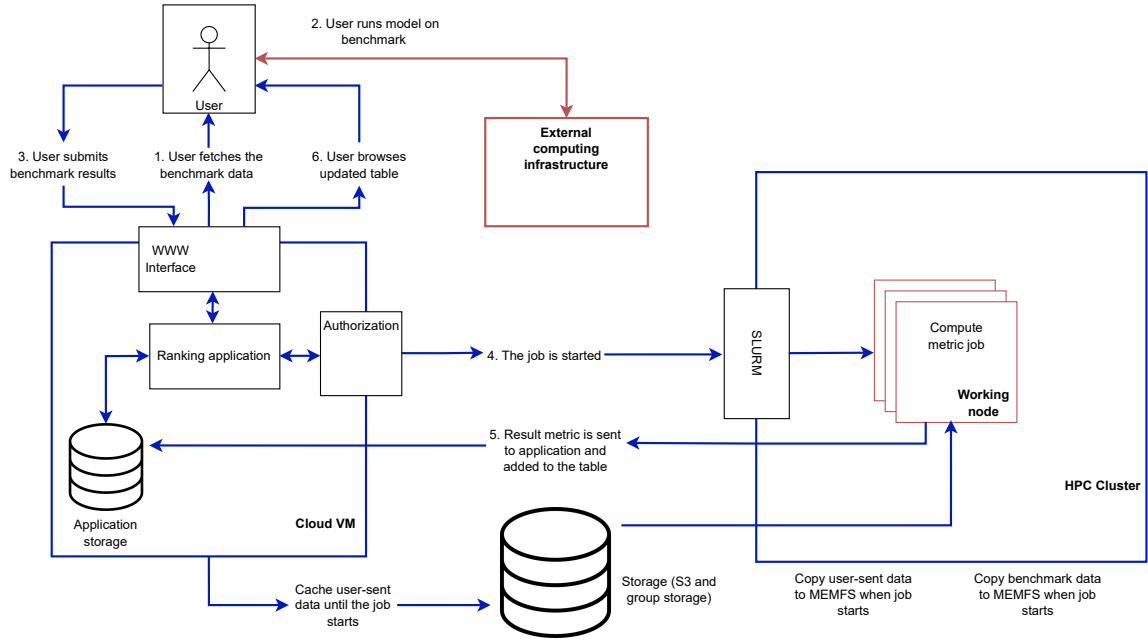


Figure 8: SPEECHM architecture. The platform is composed of the WebUI for managing user submissions and showing evaluation results, produced by the evaluation scripts executed in the scope of Slurm jobs on the HPC Ares (for CPU-based calculations) and Athena (for GPU-based calculations) HPC clusters.

## 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<sup>68</sup>, 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<sup>69</sup>. 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. submit 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 through 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 submit multiples hypothesis

<sup>68</sup>[www.meetween.eu](http://www.meetween.eu)

<sup>69</sup>[iwslt2025.speechm.cloud.cyfronet.pl](http://iwslt2025.speechm.cloud.cyfronet.pl)



Figure 9: 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 8. 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*<sup>70</sup> 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 9.

### B.1.4 The evaluation scripts

The evaluation metrics are computed through a set of scripts that run on the PLGRID clusters<sup>71</sup>. Scripts to compute the metrics that benefit from usage of GPU cards (such as COMET, BLEURT and BERT scores) run on the *Athena*<sup>72</sup> cluster while the other scripts (computing ASR, BLEU and Character scores) are executed on the *Ares* cluster<sup>73</sup>.

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 SLT

- Systems are ordered according to the COMET score (denoted by COMET, the first column).
- The “Joint” table is computed by averaging the scores of the 4 test sets, aka macro-averaging.

<sup>70</sup>[slurm.schedmd.com/slurmrestd.html](http://slurm.schedmd.com/slurmrestd.html)

<sup>71</sup>[portal.plgrid.pl](http://portal.plgrid.pl)

<sup>72</sup>[www.cyfronet.pl/en/19073,artykul,athena.html](http://www.cyfronet.pl/en/19073,artykul,athena.html)

<sup>73</sup>[www.cyfronet.pl/en/computers/18827,artykul,ares\\_supercomputer.html](http://www.cyfronet.pl/en/computers/18827,artykul,ares_supercomputer.html)



- The “D” column indicates the data condition in which each submitted run was trained, namely: Constrained (C), Constrained<sup>+LLM</sup> (C<sup>+</sup>), Unconstrained (U).
- This year, we have submissions of both cascade and end-to-end architectures.

System	D	Joint					
		COMET (↑)	BLEU (↑)	BLEURT (↑)	chrF (↑)	CharacTER (↓)	TER (↓)
KIT	U	0.783	30.2	0.660	57.4	0.451	63.1
NYA	U	0.780	28.9	0.646	56.5	0.463	64.6
NeMo	U	0.765	28.7	0.638	56.1	0.465	67.1
AIB	U	0.676	22.0	0.520	47.8	0.548	77.4
NAIST	U	0.644	17.9	0.469	43.0	0.628	77.1
CUNI-NL	C <sup>+</sup>	0.632	19.4	0.465	44.3	0.634	73.2
NAIST	C <sup>+</sup>	0.594	13.4	0.400	37.9	0.693	83.3

System	D	Accent					
		COMET	BLEU	BLEURT	chrF	CharacTER	TER
NYA	U	0.742	20.7	0.595	52.0	0.543	78.3
KIT	U	0.733	21.8	0.603	52.3	0.54	76.2
NeMo	U	0.712	18.4	0.579	51.0	0.549	92.5
NAIST	U	0.695	16.7	0.551	46.8	0.598	81.2
AIB	U	0.688	19.4	0.533	50.2	0.569	79.7
NAIST	C <sup>+</sup>	0.672	13.7	0.518	43.8	0.628	86.8
CUNI-NL	C <sup>+</sup>	0.628	15.3	0.473	40.6	0.676	80.5

System	D	Asharq News					
		COMET	BLEU	BLEURT	chrF	CharacTER	TER
KIT	U	0.833	36.2	0.722	65.0	0.382	54.5
NYA	U	0.832	37.5	0.708	64.7	0.396	52.8
NeMo	U	0.826	38.1	0.708	64.7	0.384	51.1
AIB	U	0.811	35.8	0.686	62.1	0.416	53.3
NAIST	U	0.601	18.3	0.408	40.9	0.677	74.6
CUNI-NL	C <sup>+</sup>	0.583	21.8	0.432	47.5	0.629	71.8
NAIST	C <sup>+</sup>	0.503	10.3	0.282	30.3	0.804	81.8

System	D	ITV					
		COMET	BLEU	BLEURT	chrF	CharacTER	TER
KIT	U	0.722	21.8	0.579	44.3	0.546	74.7
NYA	U	0.704	19.1	0.551	42.9	0.564	78.4
NeMo	U	0.695	19.8	0.546	42.2	0.571	75.8
CUNI-NL	C <sup>+</sup>	0.544	10.6	0.318	29.7	0.778	83.3
NAIST	U	0.513	8.20	0.287	25.7	0.787	93.8
NAIST	C <sup>+</sup>	0.491	6.90	0.249	25.4	0.786	97.5
AIB	U	0.401	2.60	0.172	17.9	0.788	121

System	D	Scientific Presentations					
		COMET	BLEU	BLEURT	chrF	CharacTER	TER
KIT	U	0.842	40.9	0.735	68.0	0.337	47.0
NYA	U	0.840	38.4	0.730	66.4	0.348	48.9
NeMo	U	0.828	38.4	0.718	66.3	0.355	49.0
AIB	U	0.804	30.0	0.688	61.0	0.418	56.1
CUNI-NL	C <sup>+</sup>	0.772	29.9	0.638	59.4	0.453	57.1
NAIST	U	0.766	28.6	0.630	58.5	0.448	59.0
NAIST	C <sup>+</sup>	0.710	22.8	0.550	52.2	0.554	67.1

Table 21: Official results of the automatic evaluation for the Offline Speech Translation Task on official test sets, **English to German**.

System	D	Asharq News					
		COMET (↑)	BLEU (↑)	BLEURT (↑)	chrF (↑)	CharacTER (↓)	TER (↓)
NYA	U	0.839	22.1	0.665	55.4	0.440	64.4
NeMo	U	0.820	19.7	0.644	52.7	0.461	66.2
AIB	U	0.812	17.2	0.627	50.3	0.496	67.0

Table 22: Official results of the automatic evaluation for the Offline Speech Translation Task on official test set, **English to Arabic**.

System	D	Scientific Presentations					
		COMET (↑)	BLEU (↑)	BLEURT (↑)	chrF (↑)	CharacTER (↓)	TER (↓)
NYA	U	0.860	56.7	0.713	49.1	0.418	32.9
AIB	U	0.856	55.7	0.719	49.1	0.427	33.0
BigWaterMelon	U	0.845	56.2	0.703	49.5	0.436	34.1
NeMo	U	0.844	46.3	0.699	40.6	0.481	39.0
NAIST	U	0.771	40.2	0.590	33.4	0.600	48.4
NAIST	C <sup>+</sup>	0.711	31.0	0.487	26.6	0.724	56.7

Table 23: 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.3 Simultaneous SLT

Latency Regime	Team	Quality Metrics		StreamLAAL	
		BLEU (↑)	COMET (↑)	dev (↓)	test (↓)
Low-Latency	Baseline-Fixed	15.74	0.551	1.87	1.70 (2.61)
	Baseline-VAD	17.81	0.595	1.82	1.99 (3.10)
	NAIST	20.85	0.680	1.92	1.82 (N/A)
	OSU *	22.04	0.708	1.84	1.73 (2.47)
	CMU	22.63	0.697	1.69	1.47 (1.81)
High-Latency	Baseline-Casc. *	24.89	0.699	3.23	3.20 (4.59)
	Baseline-Fixed	19.15	0.593	2.35	3.54 (4.57)
	Baseline-VAD	22.07	0.644	3.43	2.95 (3.82)
	NAIST	24.58	0.717	3.99	3.71 (N/A)
	OSU *	25.80	0.729	3.34	3.21 (4.41)
	UPV *	29.81	0.739	2.94	2.90 (3.37)
	CUNI *	35.25	0.790	3.77	3.32 (N/A)

Table 24: English-to-German simultaneous speech-to-text translation divided by latency regimes. Latency is measured in seconds. Values in parentheses are computationally aware latency and are provided for system submissions only on the test set. Cascaded systems are marked with an asterisk (\*).

Latency Regime	Team	Quality Metrics		StreamLAAL	
		BLEU (↑)	COMET (↑)	dev (↓)	test (↓)
Low-Latency	Baseline-Fixed	20.42	0.568	2.35	3.76 (4.64)
	Baseline-VAD	22.63	0.588	1.88	1.96 (2.74)
	OSU *	34.06	0.705	2.22	2.20 (3.34)
	NAIST	37.82	0.747	2.46	2.28 (N/A)
	CMU	43.26	0.773	2.19	2.15 (2.66)
High-Latency	Baseline-Fixed	21.84	0.595	3.12	3.11 (3.98)
	Baseline-VAD	26.19	0.638	3.28	3.15 (3.91)
	OSU *	37.07	0.733	3.52	3.49 (4.82)
	CUNI *	39.07	0.808	3.54	2.94 (N/A)
	NAIST	39.41	0.761	3.70	3.20 (N/A)

Table 25: English-to-Chinese simultaneous speech-to-text translation divided by latency regimes. Latency is measured in seconds. Values in parentheses are computationally aware latency and are provided for system submissions only on the test set. Cascaded systems are marked with an asterisk (\*).

Latency Regime	Team	Quality Metrics		StreamLAAL	
		BLEU (↑)	COMET (↑)	dev (↓)	test (↓)
Low-Latency	Baseline-VAD	11.32	0.591	2.35	2.21 (3.25)
	NAIST	23.84	0.786	3.34	2.83 (N/A)
High-Latency	Baseline-Fixed	10.05	0.610	3.74	4.62 (5.89)
	Baseline-VAD	13.76	0.667	3.66	3.54 (4.62)
	NAIST	23.99	0.787	3.98	3.25 (N/A)
	CUNI *	33.44	0.841	4.48	4.23 (N/A)

Table 26: English-to-Japanese simultaneous speech-to-text translation divided by latency regimes. Values in parentheses are computationally aware latency and are provided for system submissions only on the test set. Cascaded systems are marked with an asterisk (\*).

Latency Regime	Team	Quality Metrics		StreamLAAL	
		BLEU (↑)	COMET (↑)	dev (↓)	test (↓)
Low-Latency	Baseline-Fixed	19.96	0.647	1.87	2.31 (3.26)
	Baseline-VAD	19.94	0.642	1.78	2.46 (3.70)
	CUNI	20.78	0.715	1.76	1.41 (N/A)
High-Latency	Baselines-Casc.*	19.92	0.675	3.64	4.29 (8.11)
	Baseline-Fixed	21.44	0.662	3.41	3.34 (4.22)
	Baseline-VAD	23.55	0.677	3.34	3.68 (4.67)
	CUNI	21.94	0.729	2.63	1.51 (N/A)

Table 27: Czech-to-English simultaneous speech-to-text translation for the native speakers test set divided by latency regimes. Latency is measured in seconds. Values in parentheses are computationally aware latency and are provided for system submissions only on the test set.

Latency Regime	Team	Quality Metrics		StreamLAAL	
		BLEU (↑)	COMET (↑)	dev (↓)	test (↓)
Low-Latency	Baseline-Fixed	8.84	0.568	1.87	3.33 (4.53)
	Baseline-VAD	12.84	0.589	1.78	1.00 (1.88)
	CUNI	21.59	0.704	1.76	3.30 (N/A)
High-Latency	Baselines-Casc.*	24.00	0.698	3.64	5.30 (9.43)
	Baseline-Fixed	18.02	0.612	3.41	5.19 (6.22)
	Baseline-VAD	16.46	0.626	3.34	1.85 (2.62)
	CUNI	24.53	0.749	2.63	1.79 (N/A)

Table 28: Czech-to-English simultaneous speech-to-text translation for the non-native speakers test set divided by latency regimes. Latency is measured in seconds. Values in parentheses are computationally aware latency and are provided for system submissions only on the test set.



Latency Regime	Team	Quality Metrics		StreamLAAL	
		BLEU (↑)	COMET (↑)	dev (↓)	test (↓)
Low-Latency	Baseline-Fixed	10.89	0.490	1.87	2.48 (3.57)
	Baseline-VAD	10.22	0.487	1.82	3.43 (4.41)
	CMU	11.18	0.525	1.87	1.74 (2.26)
	NAIST	12.15	0.570	1.92	1.89 (N/A)
	OSU *	16.11	0.618	1.84	2.06 (2.90)
High-Latency	Baselines-Casc.*	13.99	0.583	3.23	3.09 (4.37)
	Baseline-Fixed	13.03	0.520	2.35	4.06 (4.92)
	Baseline-VAD	11.07	0.500	3.43	3.33 (4.33)
	CUNI *	12.51	0.626	3.77	2.99 (N/A)
	NAIST	12.92	0.585	3.99	3.70 (N/A)
	UPV *	16.26	0.599	2.94	3.58 (N/A)
	OSU *	18.73	0.643	3.34	3.81 (4.83)

Table 29: English-to-German simultaneous speech-to-text translation for the challenging accented test set divided by latency regimes. Latency is measured in seconds. Values in parentheses are computationally aware latency and are provided for system submissions only on the test set. Cascaded systems are marked with an asterisk (\*).

## B.4 Automatic Subtitling

Team	Cndt	System	Domain	Sub. qual.	Translation quality			Subtitle compliance		
				SubER	Bleu	ChrF	Bleurt	CPS	CPL	LPB
APPTeK	U	prmry	ITV	62.86	19.11	40.62	.4899	93.78	100.00	100.00
			Asharq-Bloomberg	50.87	35.21	59.03	.6057	92.44	100.00	99.19
APPTeK	U	cntrstv1	ITV	63.57	20.65	42.94	.5043	82.36	100.00	97.55
			Asharq-Bloomberg	51.93	34.28	57.83	.5869	95.69	100.00	99.85
APPTeK	U	cntrstv2	ITV	63.31	18.06	39.16	.4767	97.40	100.00	100.00
			Asharq-Bloomberg	50.94	35.02	58.99	.6052	92.13	100.00	99.07

Table 30: Subtitling Task: automatic evaluation scores on tst2025 en→de. *U* stands for *unconstrained* training condition; *prmry* and *cntrstv* for *primary* and *contrastive* systems.

Team	Cndt	System	Domain	Sub. qual.	Translation quality			Subtitle compliance		
				SubER	Bleu	ChrF	Bleurt	CPS	CPL	LPB
APPTeK	U	prmry	Asharq-Bloomberg	62.13	21.56	52.74	.5995	99.81	100.00	99.97
APPTeK	U	cntrstv1	Asharq-Bloomberg	62.62	21.22	52.25	.5982	99.81	100.00	99.97
APPTeK	U	cntrstv2	Asharq-Bloomberg	62.57	21.87	53.27	.6044	96.28	100.00	99.38

Table 31: Subtitling Task: automatic evaluation scores on tst2025 en→ar. *U* stands for *unconstrained* training condition; *prmry* and *cntrstv* for *primary* and *contrastive* systems.

Team	Cndt	System	Domain	Sub. qual.	Translation quality			Subtitle compliance		
				SubER	Bleu	ChrF	Bleurt	CPS	CPL	LPB
APPTeK	U	prmry	ITV	66.55	18.86	41.71	.5053	93.50	100.00	100.00
APPTeK	U	cntrstv1	ITV	69.32	19.07	43.08	.5164	83.53	100.00	97.87
APPTeK	U	cntrstv2	ITV	65.97	18.33	40.96	.5008	97.07	100.00	100.00
Submissions 2024										
APPTeK	U	prmry	ITV	72.38	16.98	40.42	.4683	69.23	100.00	99.92
FBK-AI4C <sub>DIR</sub>	C	prmry	ITV	78.90	9.67	28.43	.2911	70.45	90.04	99.97
FBK-AI4C <sub>CSC</sub>	U	prmry	ITV	79.92	14.86	35.16	.4048	54.20	91.12	100.00
HW-TSC	U	prmry	ITV	76.04	16.09	41.34	.5098	61.72	61.80	100.00

Table 32: Subtitling Task: automatic evaluation scores on tst2024 en→de. *C* and *U* stand for *constrained* and *unconstrained* training condition, respectively; *prmry* and *cntrstv* for *primary* and *contrastive* systems.

Team	Cndt	System	Domain	Sub. qual.	Translation quality			Subtitle compliance		
				SubER	Bleu	ChrF	Bleurt	CPS	CPL	LPB
APPTeK	U	prmry	ITV	65.26	18.79	41.62	.5064	93.32	100.00	100.00
APPTeK	U	cntrstv1	ITV	66.97	20.27	43.73	.5219	82.02	100.00	97.69
APPTeK	U	cntrstv2	ITV	65.01	18.26	40.77	.5003	97.12	100.00	100.00
Submissions 2024										
APPTeK	U	prmry	ITV	69.21	17.97	41.27	.4790	67.64	100.00	99.96
HW-TSC	U	prmry	ITV	72.16	18.35	42.95	.5244	60.15	62.37	100.00
FBK-AI4C <sub>CSC</sub>	U	prmry	ITV	74.91	16.19	35.91	.3996	54.70	92.97	100.00
FBK-AI4C <sub>DIR</sub>	C	prmry	ITV	77.15	10.40	29.13	.2939	68.73	91.00	99.97
Submissions 2023										
APPTeK	U	prmry	ITV	69.83	14.43	35.27	0.4028	86.01	100.00	100.00
TLT	U	prmry	ITV	73.11	14.92	37.13	0.4501	80.21	99.47	100.00
APPTeK	C	prmry	ITV	80.87	9.08	27.74	0.2612	91.14	100.00	100.00
FBK <sub>DIR</sub>	C	prmry	ITV	82.67	8.05	26.10	0.2255	67.75	85.12	100.00

Table 33: Subtitling Task: automatic evaluation scores on tst2023 en→de. *C* and *U* stand for *constrained* and *unconstrained* training condition, respectively; *prmry* and *cntrstv* for *primary* and *contrastive* systems.

## B.5 Low-Resource SLT

**North Levantine Arabic→English (Unconstrained Condition)**

Team	System	BLEU↓	COMET	chrF2
KIT	primary	23.34	0.704	45.09
LIA	primary	22.56	0.719	44.72
KIT	contrastive2	21.93	0.697	44.67
LIA	contrastive2	21.45	0.694	43.13
LIA	contrastive1	21.02	0.698	42.92
ALADAN	primary	20.02	0.661	39.91
KIT	contrastive1	19.11	0.683	40.95
AIB_Marco	contrastive4	16.47	0.683	37.96
AIB_Marco	contrastive3	16.22	0.667	37.48
AIB_Marco	contrastive1	15.82	0.646	36.23
JHU	contrastive1	15.39	0.657	35.91
JHU	primary	14.64	0.649	36.23
AIB_Marco	primary	12.01	0.655	34.19
AIB_Marco	contrastive2	10.53	0.573	27.69

Table 34: Automatic evaluation results for the North Levantine Arabic to English task, unconstrained Condition. A lowercase, no-punctuation variant of chrF2 is reported. The `Unbabel/wmt22-comet-da` model was used for COMET computation, with the source side (Arabic transcript) unmodified and the target side lowercased and with removed punctuation. The *AIB\_Marco* team did not submit a system description paper.

**Bemba→English (Unconstrained Condition)**

Team	System	BLEU
JHU	primary	32.6
KIT	primary	28.8
KIT	contrastive2	28.1
JHU	contrastive1	27.0
KIT	contrastive1	27.0
JHU	contrastive2	26.7

Team	System	WER
KIT ASR	primary	33.2
JHU ASR	primary	35.7

Table 35: Automatic evaluation results for the Bemba to English task, unconstrained Condition.

**Bhojpuri→Hindi (Unconstrained Condition)**

Team	System	BLEU	chrF2
GMU	contrastive1	3.4	23.0
GMU	contrastive2	2.0	16
GMU	primary	3.9	24.0
JHU	contrastive1	10.7	34.0
JHU	contrastive2	7.8	32.0
JHU	primary	10.5	34.0
IIITH_BUT	contrastive1	10.2	32.0
IIITH_BUT	primary	9.9	33.0

Table 36: Automatic evaluation results for the Bhojpuri to Hindi task, unconstrained Condition.

**Estonian→English (Unconstrained Condition)**

Team	System	BLEU	chrF2	COMET
AIB_Marco	contrastive1	29.3	55.8	0.7944
AIB_Marco	contrastive2	23.3	48.3	0.7601
AIB_Marco	primary	30.9	57.4	0.7958
GMU	contrastive1	30.2	53.4	0.7746
GMU	contrastive2	29.6	52.9	0.7760
GMU	primary	29.8	53.1	0.7767

Table 37: Automatic evaluation results for the Estonian to English task, unconstrained condition.

**Irish→English (Unconstrained Condition)**

Team	System	BLEU	chrF2
AIB_Marco	contrastive1	7.8	32.0
AIB_Marco	contrastive2	12.5	34.0
AIB_Marco	primary	12.5	34.0
GMU	contrastive1	8.4	32.0
GMU	contrastive2	6.7	30.0
GMU	primary	13.4	34.0
JHU	contrastive1	12.0	33.0
JHU	contrastive2	12.3	33.0
JHU	primary	11.6	33.0

Table 38: Automatic evaluation results for the Irish to English task, unconstrained Condition.

**Maltese→English (Unconstrained Condition)**

Team	System	BLEU	chrF2
KIT	primary	58.9	76.5
SETU-DCU	primary	56.7	81.9
KIT	contrastive2	56.2	75.0
KIT	contrastive1	55.2	74.4
SETU-DCU	contrastive1	52.6	72.1
UoM	primary	52.4	72.3
UoM	contrastive1	52.4	72.3
UoM	contrastive2	52.3	72.1
SETU-DCU	contrastive2	44.7	65.5
JHU	primary	41.4	68.6
JHU	contrastive1	36.5	64.2
UoM-DFKI	primary (e2e)	35.1	59.0
JHU	contrastive2	24.8	55.8
UoM-DFKI	contrastive1 (e2e)	18.5	42.0

Table 39: Automatic evaluation results for the Maltese to English task, Unconstrained Condition. e2e denotes end-to-end system.

**Maltese→English (Constrained Condition)**

Team	System	BLEU	chrF2
UoM	primary	0.5	15.6

Table 40: Automatic evaluation results for the Maltese to English task, Constrained Condition.



**Marathi→Hindi (Constrained Condition)**

Team	System	BLEU	chrF2
SRI-B	contrastive1	22.6	50.0
SRI-B	contrastive2	24.0	52.0
SRI-B	primary	23.7	52.0

Table 41: Automatic evaluation results for the Marathi to Hindi task, Constrained Condition.

**Marathi→Hindi (Unconstrained Condition)**

Team	System	BLEU	chrF2
GMU	contrastive1	44.3	68.0
GMU	contrastive2	41.5	66.0
GMU	primary	43.4	67.0
JHU	contrastive1	40.7	65.0
JHU	contrastive2	40.0	65.0
JHU	primary	41.4	65.0

Table 42: Automatic evaluation results for the Marathi to Hindi task, Unconstrained Condition.

**Quechua→Spanish (Unconstrained Condition)**

Team	System	BLEU	chrF2
GMU	contrastive1	12.9	46.4
GMU	contrastive2	13.0	46.4
GMU	primary	12.7	46.2
JHU	contrastive1	11.0	46.7
JHU	primary	9.0	43.5
QUESPA	contrastive1	15.0	52.4
QUESPA	contrastive2	26.7	48.6
QUESPA	primary	14.8	51.8

Table 43: Automatic evaluation results for the Quechua to Spanish task, Unconstrained Condition.

**Fongbe→French (Unconstrained Condition)**

Team	System	BLEU	chrF2
LIA	primary	<b>39.6</b>	<b>56.7</b>
LIA	contrastive1	37.23	54.96
LIA	contrastive2	32.76	50.09
LIA	contrastive3	28.32	46.08
GMU	primary	31.96	48.01
JHU	primary	5.96	23.21
JHU	contrastive1	6.26	23.27
JHU	contrastive2	5.6	23.27

Table 44: Automatic evaluation results for the Fongbe to French task, Unconstrained Condition.

## B.6 Dialectal SLT

**Tunisian Arabic→English (Unconstrained Condition)**

		test22		test23	
Team	System	BLEU	chrF	BLEU	chrF
KIT	primary	<b>22.7</b>	<b>44.4</b>	<b>21.4</b>	42.3
KIT	contrastive1	21.2	43	19.3	40.9
KIT	contrastive2	21.4	43.7	19.2	41.1
LIA	primary	22.3	44.3	21.0	<b>42.5</b>
LIA	contrastive1	22.0	43.9	20.3	41.6
LIA	contrastive2	21.6	43.4	19.2	40.3
LIA	contrastive3	21.4	43.2	19.6	41.2
GMU	primary	20.3	43.2	17.8	40.6
GMU	contrastive1	19.2	42.8	17.3	40.0
GMU	contrastive2	18.9	42.4	17.3	40.1
SYSTRAN	primary	19.2	36.0	17.5	33.9
MBZAI	primary	11.7	34.0	10.4	32.2
JHU	primary	8.2	30.4	6.8	27.6
JHU	contrastive1	30.7	42.8	7.3	27.9
JHU	contrastive2	28.6	42.4	5.5	26.1

Table 45: Automatic evaluation results for the Tunisian to English Speech Translation task, Unconstrained Condition. Primary systems are ordered in terms of the official metric BLEU on test23. We also report chrF score.

**Tunisian Arabic ASR (Unconstrained Condition)**

		test22		test23	
Team	System	WER	CER	WER	CER
GMU	primary	<b>38.0</b>	19.7	<b>39.9</b>	22.3
LIA	primary	38.6	<b>19.2</b>	40.0	<b>21.4</b>
LIA	contrastive1	39.2	44.4	40.3	22.5
AMIRBEK	primary	39.9	20.0	41.0	22.3
SYSTRAN	primary	40.6	21.0	41.8	23.3

Table 46: Word Error Rate (WER) and Character Error Rate (CER) of the ASR component of submitted cascaded systems on test22 and test23 after Arabic-specific normalization for e.g. Alif, Ya, Ta-Marbuta on the hypotheses and transcripts before computing WER/CER. Systems are ordered in terms of the official metric WER on test23.

## B.7 Instruction Following

en-en					
Model			Metric		
Organization	Condition	Role	ASR-WER ↓	SQA-BERTScore ↑	S2TSUM-BERTScore ↑
SHORT					
MICROSOFT-PHI MEETWEEN NLE	UNCONSTRAINED	BASELINE	<b>0.07</b>	0.46	-
	UNCONSTRAINED	PRIMARY	0.18	0.17	-
	CONSTRAINED	PRIMARY	0.13	<b>0.50</b>	-
CUNI-NL	UNCONSTRAINED	PRIMARY	0.15	0.21	-
		CONTRASTIVE	0.25	0.15	-
IST	UNCONSTRAINED	PRIMARY	0.15	0.14	-
LONG					
MICROSOFT-PHI KIT	UNCONSTRAINED	BASELINE	0.17	<b>0.42</b>	0.17
	CONSTRAINED	PRIMARY	<b>0.15</b>	0.41	<b>0.23</b>
en-de					
Model			Metric		
Organization	Condition	Role	S2TT-COMET ↑	SQA-BERTScore ↑	S2TSUM-BERTScore ↑
SHORT					
MICROSOFT-PHI NLE	UNCONSTRAINED	BASELINE	<b>0.77</b>	0.36	-
	CONSTRAINED	PRIMARY	0.71	<b>0.38</b>	-
CUNI-NL	UNCONSTRAINED	PRIMARY	0.72	0.21	-
		CONTRASTIVE	0.69	0.21	-
IST	UNCONSTRAINED	PRIMARY	0.34	0.22	-
LONG					
MICROSOFT-PHI KIT	UNCONSTRAINED	BASELINE	0.55	<b>0.35</b>	0.16
	CONSTRAINED	PRIMARY	<b>0.74</b>	<b>0.35</b>	<b>0.21</b>
en-it					
Model			Metric		
Organization	Condition	Role	S2TT-COMET ↑	SQA-BERTScore ↑	S2TSUM-BERTScore ↑
SHORT					
MICROSOFT-PHI NLE	UNCONSTRAINED	BASELINE	<b>0.81</b>	0.40	-
	CONSTRAINED	PRIMARY	0.75	<b>0.42</b>	-
LONG					
MICROSOFT-PHI KIT	UNCONSTRAINED	BASELINE	0.56	0.36	0.19
	CONSTRAINED	PRIMARY	<b>0.77</b>	<b>0.39</b>	<b>0.25</b>
en-zh					
Model			Metric		
Organization	Condition	Role	S2TT-COMET ↑	SQA-BERTScore ↑	S2TSUM-BERTScore ↑
SHORT					
MICROSOFT-PHI NLE IST	UNCONSTRAINED	BASELINE	<b>0.81</b>	0.33	-
	CONSTRAINED	PRIMARY	0.76	<b>0.35</b>	-
	UNCONSTRAINED	PRIMARY	0.34	0.21	-
LONG					
MICROSOFT-PHI KIT	UNCONSTRAINED	BASELINE	0.51	0.39	0.04
	CONSTRAINED	PRIMARY	<b>0.77</b>	<b>0.41</b>	<b>0.37</b>

Table 47: Complete results for the IF Task, including the BASELINE (Phi4-Multimodal). For each team, it is indicated whether the submission was under CONSTRAINED or unconstrained settings, and if it was PRIMARY or CONTRASTIVE. **Bold** indicates the best track-wise (SHORT and LONG) result per language direction, and underline indicates the overall best result among tracks.