BSC's Submission to the Instruction Following Track of IWSLT 2026

Oriol Pareras[1], **Joan Lladó-Fuentes**[1], **Pol Buitrago**[1,2], **Marc Casals-Salvador**[1,2], **Federico Costa**[1,2], **Cristina España-Bonet**[1,3]
[1]Barcelona Supercomputing Center (BSC-CNS) [2]Universitat Politècnica de Catalunya (UPC)
[3]German Research Center for Artificial Intelligence (DFKI)
{oriol.pareras,joan.llado,pol.buitrago,marc.casals,federico.costa,cristina.espana}@bsc.es

Abstract

We present the Barcelona Supercomputing Center (BSC) submission to the Instruction Following (IF) track of IWSLT 2026, which evaluates unified spoken language systems capable of solving multiple tasks through natural language instructions. Our system consists of an end-to-end (E2E) architecture that combines a speech encoder with a translation-oriented Large Language Model. The model is trained on speech and text data, covering automatic speech recognition, translation, question answering, and instruction following. We investigate a Chain-of-Thought (CoT) generation strategy that explicitly decomposes tasks by producing an intermediate transcription before the final output, which enables effective reuse of text-only supervision and improves robustness across tasks. To further support generalization, we design diverse prompt formulations and align text-only and speech inputs under a shared inference pattern. Results on IWSLT 2025 evaluation data show that our approach achieves competitive and even state-of-the-art performance across tasks.

1 Introduction

This paper provides an overview of the system submitted by the BSC to the IF shared task of IWSLT 2026 (Adelani et al., 2026). This track aims to evaluate Spoken Language systems that integrate the general capabilities of LLMs. In addition to Automatic Speech Recognition (ASR), Speech-to-Text Translation (S2TT), and Spoken Question Answering (SQA) tasks, this year's edition introduces a surprisal task whose details remain undisclosed until submission time. This setting is designed to assess the capabilities of the systems to follow instructions through natural language prompts across all four tasks. The model must process English (*en*) audio, and follow instructions written in English, Chinese (*zh*), Italian (*it*), or German (*de*), generating the output in the corresponding language.

We participate in the unconstrained condition, which allows the use of any model and data, and adopt an E2E architecture based on a speech encoder coupled with an LLM. Our approach relies on an LLM tailored for Text-to-Text Translation (T2TT), which we use as the backbone and further fine-tune for both speech-based and text-only tasks.

Following our previous work (Pareras et al., 2025), we adopt a CoT generation strategy that forces the model to transcribe the audio before following the prompt's instruction. This approach has been shown to substantially improve S2TT performance (Hu et al., 2025). We expect all tasks to benefit from this strategy, as it decomposes the problem into two simpler sub-tasks while also enabling the use of text-only Question Answering (QA) and T2TT data to boost performance. In addition, we train the model using a wide variety of task-specific prompt templates to reduce prompt overfitting. Finally, we apply a post-editing stage to the model outputs to correct format inconsistencies.

Results on the IWSLT 2025 test set show that our best system achieves strong performance across the three evaluated tasks, in comparison with the systems of that edition. Specifically, compared to the Phi4-Multimodal (Abouelenin et al., 2025) baseline, our best system remains competitive for ASR, matches performance for S2TT and improves SQA performance on average across all the languages.

2 Data

To train our model, we have used the following datasets for each task:

- **ASR:** Common Voice 22.0 (Ardila et al., 2020) (*en*), VoxPopuli (Wang et al., 2021a) (*en*), a 6.7k hours subset of Multilingual LibriSpeech (Pratap et al., 2020) (*en*) and a 2.5k hours subset of YODAS–Granary (Rao Koluguri et al., 2025) (*en*).

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 171–182

- **S2TT:** EuroparlST (Iranzo-Sánchez et al., 2020) (*en→de*, *en→it*) and CoVoST2 (Wang et al., 2021b) (*en→de*, *en→zh*). Given the translation capabilities of the backbone LLM, we use limited data for this task.

- **T2TT:** Wikimedia (Tiedemann, 2012) (*en→de*, *en→it*) and Tatoeba (Tiedemann, 2020) (*en→de*, *en→it*). We use a subset of the Opus corpora (Zhang et al., 2020) (*en→zh*), the same that was used during the training of the backbone LLM (see Section 3.1).

- **SQA:** We use LibriSQA (Zhao et al., 2025) (*en*), which is adopted in the constrained condition of the shared task. We also add NMSQA (Lin et al., 2022) (*en*), which extends SQuAD (Rajpurkar et al., 2016) to speech inputs. For unanswerable questions we use SQuAD 2.0 (Rajpurkar et al., 2018) (*en*, *de*, *it* and *zh*) and also expand it to speech inputs (see §2.1).

- **QA:** We use CMRC2018 (Cui et al., 2019) (*zh*) to improve the SQA performance in *zh*, as the backbone LLM of the model we train has limited performance in this language on tasks beyond translation.

- **IF:** We use the "Precise IF" and "Chat" domains of Dolci-Instruct-SFT (Olmo et al., 2025) (*en*). For multilingual data, we use Aya Dataset (Singh et al., 2024) (*de*, *it*, *zh*).

For evaluation, we have used the IWSLT 2025 test set from MCIF (Papi et al., 2026).

2.1 Data Preprocessing

We apply the preprocessing pipelines described below to the T2TT and SQA data, and report the final data amounts in Table 1. More detailed statistics can be found in Appendix D.

T2TT Data As our backbone model is already finetuned for translation, we only used high quality T2TT samples to train our model, with the sole aim to preserve this capability during training. We use the Quality Estimation (QE) model[1] from CometKiwi (Rei et al., 2022) to select a subset of reliable samples. Thresholds were selected empirically based on the score distributions for each language pair while maintaining a sufficient number of training samples. We retain only samples with a score higher than 0.78 for *en →de*, 0.85 for *en→it* and 0.96 for *en→zh*.

Table 1: Amount of data per language and task.

Task	*en*	*de*	*it*	*zh*	Total
Hours					
ASR	11.5k	-	-	-	11.5k
S2TT	-	506	74	430	1k
SQA	842	362	411	411	2k
Number of Samples					
ASR	3.95M	-	-	-	3.95M
S2TT	-	321k	29k	289k	639k
T2TT	-	143k	114k	50k	307k
SQA	221k	39k	39k	39k	338k
QA	-	-	-	10k	10k
IF	340k	644	1k	6k	347k

SQA Data For the SQA task, we apply a translation pipeline to NMSQA and the non-answerable set of SQuAD 2.0 to cover the three remaining target languages. NMSQA's answers are usually short and context-dependent, while the answers on the evaluation data are longer and self-contained. To address this problem, we apply a similar strategy as in Lee et al.'s (2025), and rewrite them to better match the style of the test set. Finally, we extend SQuAD 2.0 to speech inputs and change its responses to the ones defined in the shared task. To avoid repeated samples during training, we partition the translated data such that each sample appears in only one target language. We describe each preprocessing method below.

- **Translation pipeline** We use Tower+ (Rei et al., 2025)[2] to translate the question and answer of each sample. The prompt used can be found in Appendix B.1.

- **Response rewriting** To rewrite the answers of NMSQA we use Gemma4 31B.[3] The prompt used can be found in Appendix B.2.

- **Speech expansion** For SQuAD 2.0, synthetic speech utterances are generated using the "context" field of each sample. Due to the length of the context of most samples, we summarize them using Gemma4 31B. The prompt used can be found in Appendix B.3. To synthesize speech we use Kyutai TTS (Zeghidour et al., 2025)[4]. The resulting speech utterances are no longer than 40 seconds long.

[1] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[2] https://huggingface.co/Unbabel/Tower-Plus-9B

[3] https://huggingface.co/google/gemma-4-31B

[4] https://huggingface.co/kyutai/tts-1.6b-en_fr

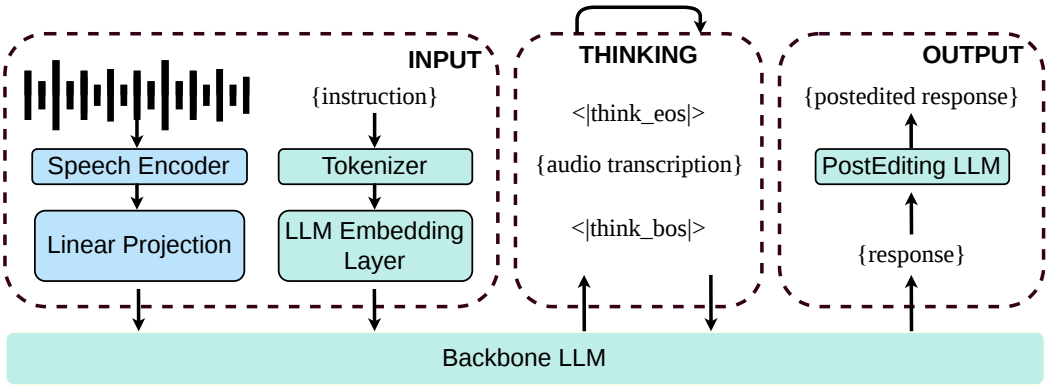

Figure 1: System Diagram

2.2 Prompt Templates

We apply task-specific prompt templates to format each training sample, incorporating: an instruction, a placeholder token to be filled with the audio speech embeddings, and the audio transcription enclosed within special `<|think_bos|>` and `<|think_eos|>` tokens (only in non-ASR speech tasks). This CoT formulation encourages the model to transcribe the audio before producing the final task output. Figure 2, located in Appendix A, illustrates the prompt formatting strategy alongside the instruction sets.

By mixing IF data with speech task data with varied instructions, we hypothesize that the model is encouraged to extend the instruction-following capabilities acquired from IF data to the speech domain. This could allow it to generalize better to tasks not explicitly seen during training, which is particularly relevant for the surprisal task.

To further maintain consistency with the CoT generation when training on text-only QA datasets, we simulate the transcription step. Concretely, the text context is injected into the reasoning block in the user turn, as a surrogate transcription.

3 System Description

Figure 1 shows a diagram of our system. Below we detail the model architecture, how it is trained and the strategy used for inference.

3.1 Speech LLM Architecture

We build a Speech LLM following a similar approach as Verdini et al.'s (2025). Each speech utterance is first encoded with a speech encoder and projected to the input embedding space of an LLM using a linear projection. We test two different speech encoders in different submissions: SeamlessM4T v2's (Barrault et al., 2023) speech encoder[5] and `mhubert-base-25hz` (Hassid et al., 2023).[6] The linear projection is randomly initialized at the start of the training. Only SEAMLESS encoder weights are frozen during training (mHuBERT encoder is fully fine-tuned). For the backbone LLM we use `SalamandraTA-7b-instruct-WMT25` (Garcia Gilabert et al., 2025),[7] a highly capable translation model in 40 languages.

3.2 Training Details

We train the model for one epoch, sampling each data file proportionally to its share of the total training samples to preserve the original distribution across sources and languages. We employ AdamW optimizer, a cosine learning rate scheduler, and gradient clipping with a maximum norm of 1.0. We use a maximum learning rate of $1 \cdot 10^{-5}$ and a minimum of $1 \cdot 10^{-6}$, a warm up during the first 10% of updates and set a maximum sequence length of 2048. To optimize memory usage, we apply mixed precision (`bfloat16`). We run the trainings on 32 GPUs with a global batch size of 256. All runs are executed on customized NVIDIA H100 GPUs with 64GB of VRAM.

3.3 Inference Strategy

For inference, we use beam-search decoding with five beams. Then, we postprocess the model's output. The first step is to remove the thinking sequence explained in §2.2. Consequently, we use an LLM to post-edit the model's responses. This allows us to correct formatting errors, especially for the surprisal task, because the model has not seen

[5] https://huggingface.co/facebook/seamless-m4t-v2-large

[6] https://huggingface.co/slprl/mhubert-base-25hz

[7] https://huggingface.co/LangTech-MT/salamandraTA-7b-instruct-WMT25

Table 2: MCIF IF short track results on IWSLT 2025 test set. Scores include ASR (WER), QA (BERTScore), and ST (COMET) across multiple languages. **Bold** values indicate the best result per task. U/C stand for UNCONSTRAINED / CONSTRAINED. *Submission* indicates the submission title of each model.

		ASR (↓)	SQA (↑)				S2TT (↑)		
Model	*Condition / Submission*	*en*	*en*	*de*	*it*	*zh*	*de*	*it*	*zh*
Phi4-Multimodal	U / PRIMARY	**0.07**	0.46	0.36	0.40	**0.37**	0.77	**0.81**	**0.81**
NLE	C / PRIMARY	0.13	**0.50**	0.38	**0.42**	0.35	0.71	0.75	0.76
CUNI-NL	U / PRIMARY	0.15	0.21	0.21	-	-	0.72	-	-
IST	U / PRIMARY	0.15	0.14	0.22	-	0.21	0.34	-	0.34
Ours (MHUBERT)	U / CONTRASTIVE 2	0.12	0.38	0.35	0.34	0.35	0.77	0.80	0.79
+ post-editing	U / CONTRASTIVE 1	0.12	0.38	0.37	0.35	0.35	0.77	0.80	0.79
Ours (SEAMLESS)	U / -	0.11	0.44	**0.42**	0.36	0.34	**0.78**	**0.81**	0.79
+ post-editing	U / PRIMARY	0.11	0.44	**0.42**	0.41	**0.37**	**0.78**	**0.81**	0.80

it during training. We use Gemma4 31B for this step, and the prompt template used can be found in Appendix C.

4 Experiments and Results

We evaluate our models with the MCIF benchmark's framework[8]. ASR is evaluated with Word Error Rate (WER), SQA with BERTScore, and S2TT with COMET. We compare against four reference systems: Phi4-Multimodal (Abouelenin et al., 2025), which served as the baseline in IWSLT 2025; NLE (Lee et al., 2025), the top constrained submission; and IST (Attanasio et al., 2025) and CUNI-NL (Luu and Bojar, 2025), two unconstrained submissions.

We report results for two variants of our system, differing in the speech encoder used: MHUBERT-based and SEAMLESS-based. For each variant, we also report results after applying the post-editing stage described in §3.3. We note that the SEAMLESS-based model was only trained for approximately 0.8 epochs due to technical constraints; we hypothesize that completing the full training epoch would yield further improvements.

Table 2 reports the results of all systems across tasks and languages. Overall, our approach achieves competitive performance, matching or surpassing prior systems in several settings despite the training being incomplete for our strongest model.

ASR Although Phi4-Multimodal obtains the best ASR score with 0.07 WER, our best model (SEAMLESS-based + post-editing) reaches 0.11 WER, clearly outperforming NLE (0.13) and both CUNI-NL and IST (0.15).

S2TT Our best model (SEAMLESS-based + post-editing) achieves COMET scores that are overall on par with Phi4-Multimodal, obtaining the best score for *en→de* (0.78 vs. 0.77), tying for *en→it* (0.81), and remaining within 0.01 points for *en→zh* (0.80 vs. 0.81).

SQA Our best model (SEAMLESS-based + post-editing) achieves the best BERTScore for *de* (0.42) and ties with Phi4-Multimodal for *zh* (0.37). For *it*, it remains within 0.01 points of the best system (NLE, 0.42), reaching 0.41. For *en*, our model achieves 0.44, below Phi4-Multimodal (0.46) and NLE (0.50).

Post-editing has limited impact on most metrics, but yields notable SQA improvements for *it* (0.36 → 0.41) and *zh* (0.34 → 0.37) in the SEAMLESS-based model, which we attribute to the correction of output formatting errors rather than improvements in the model's underlying predictions. Qualitative inspection of the model's outputs on the IWSLT 2026 test set reveals that it helps enforce strict output formats, correcting frequent formatting errors in the raw model outputs. We hypothesize that this effect may be more relevant for the surprisal task.

5 Submitted models

We submitted three systems to the shared task. Our PRIMARY submission is the SEAMLESS-based model with post-editing. CONTRASTIVE 2 uses the MHUBERT encoder without post-editing, and CONTRASTIVE 1 is the same model with post-editing applied.

Table 3 reports the official results of our three submissions on the IWSLT 2026 benchmark. Overall, the submitted systems obtain similar ASR performance, which is consistent with the results observed in our experiments.

For S2TT, the scores are also in line with those obtained on the IWSLT 2025 benchmark. All sys-

[8] https://github.com/hlt-mt/mcif

Table 3: Official results for our submitted models in the IWSLT 2026 benchmark. The *zh* results for CONTRASTIVE 2 could not be evaluated due to a technical error. **Bold** values indicate the best result per task and language.

Submission	Lang.	TRANS-COMET (↑)	QA-BERTScore (↑)	QE-accuracy (↑)	QE-format-accuracy (↑)	ASR-WER (↓)
PRIMARY		—	**0.425**	—	—	0.134
CONTRASTIVE 1	*en*	—	0.383	—	—	**0.127**
CONTRASTIVE 2		—	0.420	—	—	**0.127**
PRIMARY		0.773	**0.454**	—	—	—
CONTRASTIVE 1	*it*	**0.787**	0.395	—	—	—
CONTRASTIVE 2		**0.787**	0.395	—	—	—
PRIMARY		**0.808**	**0.467**	0.785	0.953	—
CONTRASTIVE 1	*de*	0.798	0.425	**0.810**	**0.997**	—
CONTRASTIVE 2		0.799	0.383	0.796	0.712	—
PRIMARY		**0.782**	**0.413**	**0.929**	**0.950**	—
CONTRASTIVE 1	*zh*	0.750	0.398	0.872	0.663	—
CONTRASTIVE 2		—	—	—	—	—

tems achieve very similar translation performance, with the main difference appearing for *zh*, where the PRIMARY system outperforms the MHUBERT-based contrastive submissions.

In SQA, the PRIMARY system achieves the best results across all languages despite being trained for only 0.8 epochs. Since our approach relies on the CoT generation with an intermediate transcription step, this may suggest that the SEAMLESS encoder enables to produce better transcriptions. Nevertheless, this advantage is not clearly reflected by ASR-WER alone.

Finally, comparing CONTRASTIVE 1 and CONTRASTIVE 2 shows that post-editing has mixed effects. It substantially improves the surprisal-task format accuracy for *de*, does not affect translation performance, and slightly hurts QA performance for *en*. This suggests that our post-editing strategy is mainly useful for enforcing output constraints, but may occasionally alter otherwise valid answers. We note that these effects are likely highly dependent on both the instruction-following capabilities of the post-editing model and the prompt used for post-editing.

6 Conclusion

In this paper, we presented BSC's submission to the Instruction Following track of IWSLT 2026. Our system combines a CoT transcription strategy with a translation-oriented LLM backbone, trained on a mixture of speech and text-only data. Results on both the IWSLT 2025 evaluation set and the official IWSLT 2026 benchmark show strong performance on S2TT, SQA, and the surprisal task for the proposed language pairs, despite the primary model not completing the full planned training schedule.

Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Modelos del Lenguaje. FC and CEB acknowledge their AI4S fellowship within the "Generación D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR. We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 as BSC, Spain.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Giuseppe Attanasio, Sonal Sannigrahi, Ben Peters, and André F. T. Martins. 2025. Instituto de telecomunicações at IWSLT 2025: Aligning small-scale speech and language models for speech-to-text learning. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 347–353, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889.

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

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis CONNEAU, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 63483–63501. Curran Associates, Inc.

Ke Hu, Zhehuai Chen, Chao-Han Huck Yang, Piotr Żelasko, Oleksii Hrinchuk, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2025. Chain-of-thought prompting for speech translation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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

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

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2022. DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering. In *Interspeech 2022*, pages 5165–5169.

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 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 282–288, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

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

Oriol Pareras, Gerard I Gállego, Federico Costa, Cristina Espana-Bonet, and Javier Hernando. 2025. Revisiting direct speech-to-text translation with

speech llms: Better scaling than cot prompting? *arXiv preprint arXiv:2510.03093*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nithin Rao Koluguri, Monica Sekoyan, George Zelenfroynd, Sasha Meister, Shuoyang Ding, Sofia Kostandian, He Huang, Nikolay Karpov, Jagadeesh Balam, Vitaly Lavrukhin, Yifan Peng, Sara Papi, Marco Gaido, Alessio Brutti, and Boris Ginsburg. 2025. Granary: Speech Recognition and Translation Dataset in 25 European Languages. In *Interspeech 2025*, pages 3923–3927.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and 1 others. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, and 1 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567.

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.

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

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sebastien Bratières, Paolo Merialdo, and Simone Scardapane. 2025. How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not. In *Interspeech 2025*, pages 1813–1817.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. 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.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. CoVoST 2 and Massively Multilingual Speech Translation. In *Interspeech 2021*, pages 2247–2251.

Neil Zeghidour, Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, and Alexandre Défossez. 2025. Streaming sequence-to-sequence learning with delayed streams modeling. Technical report, Kyutai.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Zihan Zhao, Yiyang Jiang, Heyang Liu, Yu Wang, and Yanfeng Wang. 2025. Librisqa: A novel dataset and framework for spoken question answering with large language models. *IEEE Transactions on Artificial Intelligence*, 6(11):2884–2895.

A Model's Prompt Templates

A.1 Base Template

Figure 2 illustrates the prompt formatting strategy described in §2.2, while Figure 3 shows the specific version for the ASR task.

Base Template

User's Turn
<|audio_bos|><|AUDIO|><|audio_eos|>
{instruction}

Assistant's Turn
<|think_bos|>
Transcription:
{audio transcription}
<|think_eos|>
{response}

Figure 2: Base template used to train the model. CoT process is omitted in ASR tasks.

Base Template (ASR)

User's Turn
<|audio_bos|><|AUDIO|><|audio_eos|>
{instruction}

Assistant's Turn
{response}

Figure 3: Base template used to train the model for ASR tasks

A.2 ASR Instructions

The following are the instructions tailored for the ASR task.

English

1. Give me the transcription for this audio
2. Convert the speech in this recording to text
3. Could you transcribe this audio?
4. Listen and write down what is being said
5. Listen carefully and generate a word-for-word transcription
6. Transcribe the audio
7. Speech to text
8. Transcribe the following audio
9. What is being said in the audio?
10. What are the speakers saying? Write it out

German

11. Analysiere das akustische Signal und gib den passenden Text aus
12. Führe Spracherkennung auf dem folgenden Audio aus
13. Höre genau hin und erstelle eine wortgetreue Transkription
14. Höre zu und schreibe auf, was gesagt wird
15. Könntest du das Audio bitte transkribieren?
16. Sprache-zu-Text-Transkription
17. Transkribiere die folgende Audiodatei
18. Wandle die Sprache in dieser Aufnahme in Text um
19. Wandle diese gesprochene Äußerung in geschriebene Form um
20. Was sagen die Sprecher? Schreibe es als Text auf

Italian

21. Trascrivi l'audio.
22. Ascolta con attenzione e genera una trascrizione parola per parola
23. Ascolta e scrivi ciò che viene detto
24. Converti il audio di questa registrazione in testo
25. Cosa dicono? Scrivilo qui sotto.
26. Da audio a testo
27. Potresti trascrivere questa registrazione?
28. Trascrivi il seguente audio
29. Trasforma questo enunciato parlato nella sua forma scritta

Chinese

30. 分析这段语音并输出对应的文本内容
31. 语音转文字
32. 这段音频都说了什么？请把内容完整写出来
33. 请仔细聆听并生成逐字转写文本
34. 请先聆听这段语音，然后把说话内容完整写下来
35. 请对下面的音频进行语音识别并给出文字结果
36. 请将下面的音频转写成文字
37. 请将这段音频转换成文本
38. 请把这段录音中的语音转换成文字
39. 转写

A.3 S2TT Instructions

The following are the instructions tailored for the Speech to Text Translate task.

English

1. Can you translate this from [SOURCE LANG] to [TARGET LANG] please
2. Convert the speech from [SOURCE LANG] into written [TARGET LANG]
3. Could you transcribe the following audio, spoken in [SOURCE LANG], as [TARGET LANG]

4. Give the meaning of this [SOURCE LANG] audio in [TARGET LANG]
5. What is this [SOURCE LANG] audio about? Translate it into [TARGET LANG]
6. Listen to the [SOURCE LANG] speech and translate it into [TARGET LANG]
7. Produce a [TARGET LANG] translation of the [SOURCE LANG] recording
8. Translate this [SOURCE LANG] audio to [TARGET LANG]
9. Translate the following audio from [SOURCE LANG] into [TARGET LANG]
10. Translate this audio from [SOURCE LANG] to [TARGET LANG]

German

11. Bitte gib eine präzise Übersetzung dieses Audios in [SOURCE LANG] nach [TARGET LANG]
12. Bitte übersetze die folgende Audiodatei von [SOURCE LANG] nach [TARGET LANG]
13. Bitte übersetze diese in [SOURCE LANG] gesprochene Äußerung nach [TARGET LANG]
14. Erstelle eine flüssige Übersetzung der Äußerung von [SOURCE LANG] nach [TARGET LANG]
15. Formuliere die Bedeutung des Audios in [SOURCE LANG] auf [TARGET LANG]
16. Gib den Inhalt dieser Audioaufnahme in [SOURCE LANG] auf [TARGET LANG] wieder
17. Höre dir die Aufnahme in [SOURCE LANG] an und gib die Übersetzung auf [TARGET LANG] an
18. Wandle die folgende in [SOURCE LANG] gesprochene Sequenz in [TARGET LANG] um
19. Übertrage die gesprochene Passage aus [SOURCE LANG] in geschriebenes [TARGET LANG]

Italian

20. Ascolta l'audio in [SOURCE LANG] e fornisci la traduzione in [TARGET LANG]
21. Converti questo audio in [SOURCE LANG] in testo [TARGET LANG]
22. Esprimi il contenuto di questo audio in [SOURCE LANG] in lingua [TARGET LANG]
23. Interpreta la registrazione in [SOURCE LANG] e rispondi in [TARGET LANG]
24. Per favore traduci il seguente audio da [SOURCE LANG] a [TARGET LANG]
25. Produci una traduzione naturale in [TARGET LANG] per questo audio in [SOURCE LANG]
26. Traduci questa frase pronunciata in [SOURCE LANG] in [TARGET LANG]
27. Trasforma la seguente espressione orale in [SOURCE LANG] in [TARGET LANG]

Chinese

28. 请将这段[SOURCE LANG]语音翻译成[TARGET LANG]
29. 这段[SOURCE LANG]音频的主要内容是什么？请用[TARGET LANG]写出来
30. 请先听懂这段[SOURCE LANG]语音，然后将其翻译成[TARGET LANG]
31. 请将这段[SOURCE LANG]的录音内容翻译成[TARGET LANG]
32. 请将这段[SOURCE LANG]音频转换成[TARGET LANG]文本
33. 请将这段[SOURCE LANG]语音翻译成[TARGET LANG]的书面文字
34. 请将这段[SOURCE LANG]语音翻译成[TARGET LANG]
35. 请将这段[SOURCE LANG]音频翻译成[TARGET LANG]
36. 请将这段[SOURCE LANG]语音翻译成[TARGET LANG]
37. 请将这段[SOURCE LANG]语音翻译成[TARGET LANG]

A.4 T2TT Instructions

The following are the instructions tailored for the T2TT task.

English

1. Translate the following text from [SOURCE LANG] to [TARGET LANG]
2. Could you please translate this from [SOURCE LANG] into [TARGET LANG]:
3. Translate from [SOURCE LANG] to [TARGET LANG]
4. Could you translate the text provided from [SOURCE LANG] to [TARGET LANG]
5. Could you translate this [SOURCE LANG] text into [TARGET LANG]
6. Give me the translation of this from [SOURCE LANG] to [TARGET LANG]
7. How do you say this in [TARGET LANG] starting from [SOURCE LANG]?
8. Please translate this text from [SOURCE LANG] to [TARGET LANG]

9. Provide the [TARGET LANG] version of this [SOURCE LANG] text.
10. Translation task: [SOURCE LANG] to [TARGET LANG].

German

11. Bitte übersetze diesen Text von [SOURCE LANG] nach [TARGET LANG]
12. Erstelle die Version in [TARGET LANG] für diesen Text in [SOURCE LANG].
13. Gib mir die Übersetzung davon von [SOURCE LANG] nach [TARGET LANG]
14. Wandle den folgenden Text von [SOURCE LANG] nach [TARGET LANG] um
15. Wie sagt man das auf [TARGET LANG], ausgehend von [SOURCE LANG]?
16. Übersetze diesen Text von [SOURCE LANG] nach [TARGET LANG]
17. Übersetze von [SOURCE LANG] nach [TARGET LANG]:
18. Übersetzungsaufgabe: [SOURCE LANG] nach [TARGET LANG].

Italian

19. Compito di traduzione: da [SOURCE LANG] a [TARGET LANG].
20. Converti il seguente testo da [SOURCE LANG] a [TARGET LANG]
21. Dammi la traduzione di questo da [SOURCE LANG] a [TARGET LANG]
22. Fornisci la versione in [TARGET LANG] di questo testo in [SOURCE LANG].
23. Per favore, traduci questo testo da [SOURCE LANG] a [TARGET LANG]
24. Potresti tradurre da [SOURCE LANG] in [TARGET LANG]:
25. Puoi tradurre il testo fornito da [SOURCE LANG] a [TARGET LANG]
26. Puoi tradurre quanto segue da [SOURCE LANG] a [TARGET LANG]
27. Traduci questo testo da [SOURCE LANG] a [TARGET LANG]

Chinese

28. 从[SOURCE LANG]翻译到[TARGET LANG]
29. 将提供的文本从[SOURCE LANG]翻译成[TARGET LANG]
30. 将文本从[SOURCE LANG]转换为[TARGET LANG]
31. 提供这段[SOURCE LANG]文本的[TARGET LANG]版本。
32. 给出这段文字从[SOURCE LANG]到[TARGET LANG]的翻译
33. 翻译以下[SOURCE LANG]文本到[TARGET LANG]
34. 翻译任务：[SOURCE LANG]到[TARGET LANG]。
35. 请将以下内容从[SOURCE LANG]翻译为[TARGET LANG]:

B Data Preprocessing Prompts

B.1 Translation Prompt

The following prompt is used to translate samples from NMSQA and SQuAD v2.0 from English to other languages, as explained in §2.1.

```
Translate the following English source
text to {language}:
English: {text}
{language}:
```

B.2 Response Rewriting Prompt

The following prompt is used to extend the answers provided by the SQuAD v2.0 dataset, as explained in §2.1.

```
I want you to extend a text to a longer
one. You will receive a text for context,
for question, and for answer. The text
in answer is the one you need to extend
following the content in the context and
the question you are asked, and place it
to the Extended Answer field. The
context will be always in English, but
the question and the answer can be in
English, Italian, Chinese or German. The
Extended Answer must be written in the
same language as the Question and Answer
fields. The tone of the Extended Answer
should maintain the style of the
original answer and question. Do not add
external information. Use only the facts
provided in the Context.

Provide only the text for the Extended
Answer field, without any additional
comments or introductions. I will give
you a few examples so you can see how it
must be done:

_

Context: Hi, I'm Mira, and today I'll be
talking about our paper, Marked Pronas,
using natural language prompts to
measure stereotypes in language models.
```

```
This work is done in collaboration with
Sinder Mooch and Dandarovski.
Question: How many authors are involved
in the paper?
Answer: three
Extended Answer: Three authors are
involved in the paper.

—

Context: We have fine tuned two
different models. We have fine tuned a
model of long mBART to produce document
level simplifications and we also fine
tuned the normal based long mBART to
produce sentence level simplifications.
Question: Which models were investigated
during the experiments?
Answer: two different models
Extended Answer: long-mBART and normal
base mBART were used for the
experiments.

—

Context: So going back to the question
that we raised in the title of our
paper, do CoNLL-2003 taggers still work
in 2023? And we found that the answer is
actually a resounding yes.
Question: Do CoNLL-2003 taggers still
work?
Answer: yes
Extended Answer: Yes, Transformer-based
models are able to generalize even if
trained on older data.

—

Now do it for this sample:

Context: {context}
Question: {question}
Answer: {answer}
Extended Answer:
```

B.3 Summarization Prompt

The following prompt is used to summarize text for the non-answerable subset of SQuAD v2.0, as explained in §2.1.

```
Summarize the following text: {text}

Respond in a very short and concise manner
in plain text
```

C Post-Editing Prompt

The following prompt is used to post-edit the responses generated by the Speech-LLM, as explained in §3.3.

```
You are a post-editor for an
instruction-following system.

Your task is to correct the model output
only when necessary so that it complies
with the user's instruction, focusing on
formatting and instruction-following.

You are given:
1. The original instruction (prompt)
2. The model output

Important:
- You do NOT have access to the original
input (e.g., audio). Do NOT attempt to
infer or reconstruct missing
information.
- Do NOT re-solve the task. Only edit
the existing output if required.
- Do NOT reason about the model output.
If you doubt, return the original
output.

Guidelines:
- If the instruction requires a fixed
format, enforce it exactly.
- If already correct, return it
unchanged.
- Preserve the original content and
wording as much as possible.
- Do not add new content or guess
missing information.
- If the instruction is transcription or
translation, do not modify the content
in any way.
- Only fix clear formatting or
instruction-following issues.

Instruction:
{prompt}

Model output:
{output}

Final corrected output:
```

D Data Statistics

Table 4 presents the statistics of the data used to train the Speech-LLM. Note that the amounts reported correspond to the samples selected for training, not all the data (e.g. in NMSQA, a 30 / 30 / 30 / 10% distribution was selected for *de* / *it* / *zh* / *en*).

Table 4: Detailed preprocessed data statistics

Task	Dataset	Language	# Samples	Duration (h)
ASR	Common Voice 22.0	*en*	1,138,761	1,800
	VoxPopuli	*en*	177,020	501
	LibriSpeech	*en*	1,621,206	6,694
	Yodas (Granary)	*en*	1,010,031	2,469
S2TT	EuroparlST	*en*→*de*	32,629	77
		en→*it*	29,553	74
	CoVoST2	*en*→*de*	289,159	429
		en→*zh*	289,341	430
T2TT	Wikimedia	*en*→*de*	43,159	—
		en→*it*	64,627	—
	Tatoeba	*en*→*de*	100,000	—
		en→*it*	50,000	—
	Opus	*en*→*zh*	50,000	—
SQA	LibriSQA	en	208,030	727
	NMSQA/SQuAD2.0	*en*	12,978	115
		de	38,933	362
		it	38,933	411
		zh	38,933	411
QA	CMRC2018	*zh*	10,150	—
IF	Dolci Instruct SFT	*en*	339,966	—
		de	403	—
		it	279	—
		zh	1,254	—
	Aya Dataset	*de*	241	—
		it	738	—
		zh	4,909	—