NAVER LABS Europe Submission to the Instruction-following Track

NAVER LABS Europe

Beomseok Lee[1,2,3,*], **Marcely Zanon Boito**[2,*], Laurent Besacier[2], Ioan Calapodescu[2]
[1]University of Trento, [2]NAVER LABS Europe, [3]Fondazione Bruno Kessler (*Equal contribution)

europe.naverlabs.com

IWSLT 2025 Instruction Following Constrained Short Track

Constrained Models

SFM: Seamless-m4t-v2-large

LLM: Llama-3.1-8B-Instruct

Constrained Data

EuroParl ST | CoVoST2 | SpokenSQuAD

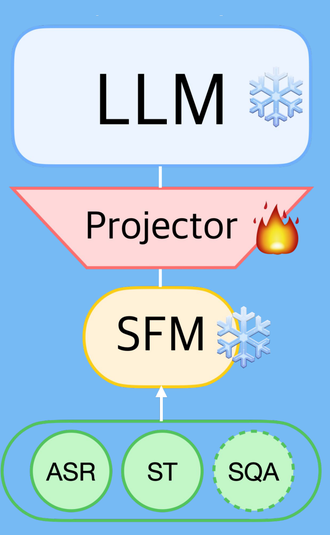
Target Task

Tasks	Input Audio	SQA Question	Output Text
ASR	🇬🇧		🇬🇧
ST	🇬🇧		🇮🇹 🇩🇪 🇨🇳
SQA	🇬🇧	🇬🇧 🇮🇹 🇩🇪 🇨🇳	🇬🇧 🇮🇹 🇩🇪 🇨🇳

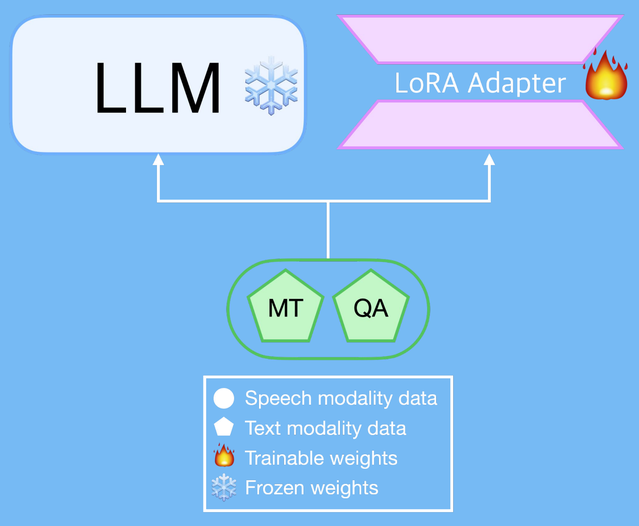
*Input audio: 15-30 seconds

Our approach: Multi-stage multimodal training

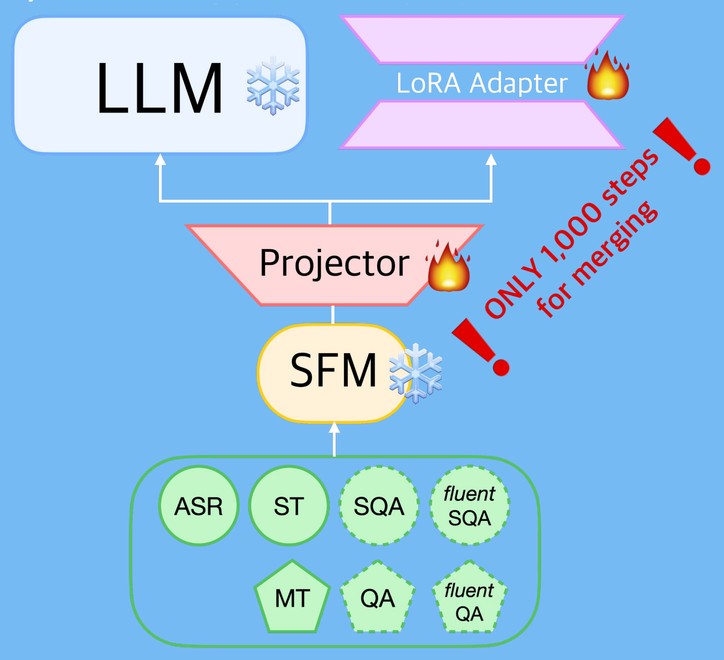
We separately train 1) a speech projector on ASR/ST/SQA **(A)** and 2) LoRA adapters on MT/QA **(B).**
We then merge them both via **multilingual and multimodal training (C)**.
All our models are multilingual.



(A) Speech Projector



(B) Text LoRA Adapters



(C) Multimodal Merging (A+B)

Methodology

Prompt Design

Speech content	Content: <speech>**[Speech embeddings]**</speech>
Text content	Content: <text>**[Speech transcription]**</text>
ASR	Question: Can you transcribe the Speech content into English text?\n
ST/MT (de)	Question: Können Sie den Inhalt der Rede in den deutschen Text übersetzen?\n
ST/MT (it)	Question: Puoi tradurre il contenuto del discorso in testo italiano?\n
ST/MT (zh)	Question: 你能把演讲内容翻译成中文吗?\n
SQA/QA	Question: **[QUESTION]**\n
Suffix	Your answer:

Task Prompt = Content + Task + Suffix

Data Augmentation

SpokenSQuAD audio regeneration
- Seamless-m4t-v2-large TTS

Multilingual SQA/QA
- Spoken-SQuAD question/answer translation
- Quality filtering with COMET-QA
- Seamless-m4t-v2-large MT

***fluent* SQA/QA**
- Span based answer → fluent answer
- LLaMA-3.1-8B-Instruct

Evaluation Metrics

Task	Metric
ASR	**MMS Norm + WER**
ST	**BLEU4, COMET**
SQA	**LLM-AS-A-JUDGE:** - EuroLLM-9B-Instruct - Gemma3-12/27B-Instruct - Llama-3.1-70B-Instruct

Experiment Results

	ASR (WER ⬇️)	ST/MT (BLEU ⬆️) / (COMET ⬆️)			SQA/QA (LLM-AS-A-JUDGE ⬆️)			
Model (fine-tuning tasks)	**en**	**en-de**	**en-it**	**en-zh**	**en-en**	**en-de**	**en-it**	**en-zh**
Text-only Models (MT/QA)								
Llama-3.1-8B-Instruct	-	23.88 / 0.779	35.51 / 0.806	45.89 / 0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69 / 0.838**	**48.31 / 0.863**	**53.65 / 0.867**	83.4%	75.7%	71.4%	69.5%
Speech-only Models (ASR/ST/SQA)								
Seamless-m4t-v2-large	**17.6**	**27.95** / 0.737	**43.54** / 0.788	33.58 / 0.753	-	-	-	-
A.1 Speech Projector (ASR/ST)	19.8	27.58 / **0.760**	36.30 / 0.796	40.62 / **0.793**	-	-	-	-
A.1 Speech Projector (ASR/ST/SQA)	19.9	27.20 / **0.760**	36.60 / **0.797**	**40.72** / 0.792	0.7%	0.5%	0.3%	0.6%
Multimodal Models (ASR/ST/SQA)								
A.1 + B (ASR/ST/MT/SQA/QA)	17.7	30.37 / 0.758	**41.22 / 0.791**	42.76 / 0.795	79.8%	71.9%	69.4%	65.5%
A.1 + B (ASR/ST/MT/*fluent*SQA/*fluent*QA)	18.6	**30.75** / 0.755	40.48 / 0.788	42.51 / 0.789	90.3%	85.2%	82.9%	76.4%
A.2 + B (ASR/ST/MT/SQA/QA)	18.2	29.91 / 0.759	38.13 / 0.786	43.12 / **0.799**	80.5%	74.9%	68.0%	66.7%
A.2 + B (ASR/ST/MT/*fluent*SQA/*fluent*QA)	18.7	29.68 / **0.763**	32.28 / 0.782	**43.38** / 0.798	**91.1%**	**87.3%**	**84.8%**	**78.0%**

Results Table: ASR and ST/MT scores are obtained using ACL 60-60 eval set, while SQA/QA scores are obtained using SpokenSQuAD test set. Text baseline (Llama-3.1-8B-Instruct) and Speech baseline (Seamless-m4t-v2-large) results are presented.

Submitted System

A.1 + B model trained with ASR/ST/MT/*fluent*SQA/fluentQA

en-en	ASR-WER ⬇️	13.0
	SQA-BERTScore ⬆️	0.50
en-de	ST-COMET ⬆️	0.71
	SQA-BERTScore ⬆️	0.38
en-it	ST-COMET ⬆️	0.75
	SQA-BERTScore ⬆️	0.42
en-zh	ST-COMET ⬆️	0.76
	SQA-BERTScore ⬆️	0.35

Official results from the organizers

READ OUR PAPER

Acknowledgements: This is an output of the **European Project UTTER** (Unified Transcription and Translation for Extended Reality) funded by European Union's Horizon Europe Research and Innovation programme under grant agreement number 101070631.