Training Speech LLMs: Insights and Lessons Learned

Marcely Zanon Boito

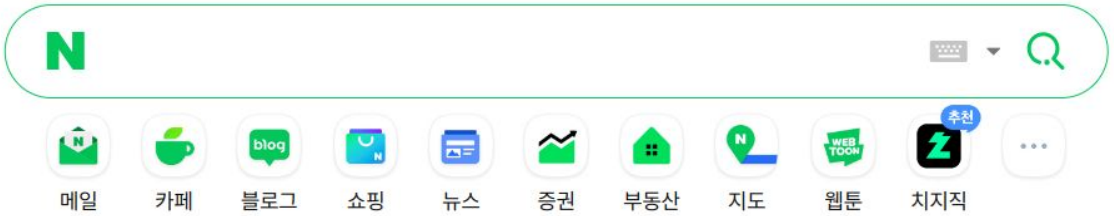
12/2025

Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS

- (2021) **PhD in Computer Science at University Grenoble Alpes**
  "Models and Resources for Attention-based Unsupervised Word Segmentation: an application to computational language documentation"
- (2021-2022) **Postdoc at Avignon University**
  Low-resource Speech Translation and Self-Supervised Learning for Speech
- (Since 2022) **Research Scientist at NAVER LABS Europe**
  Multimodality and Speech Processing

NAVER

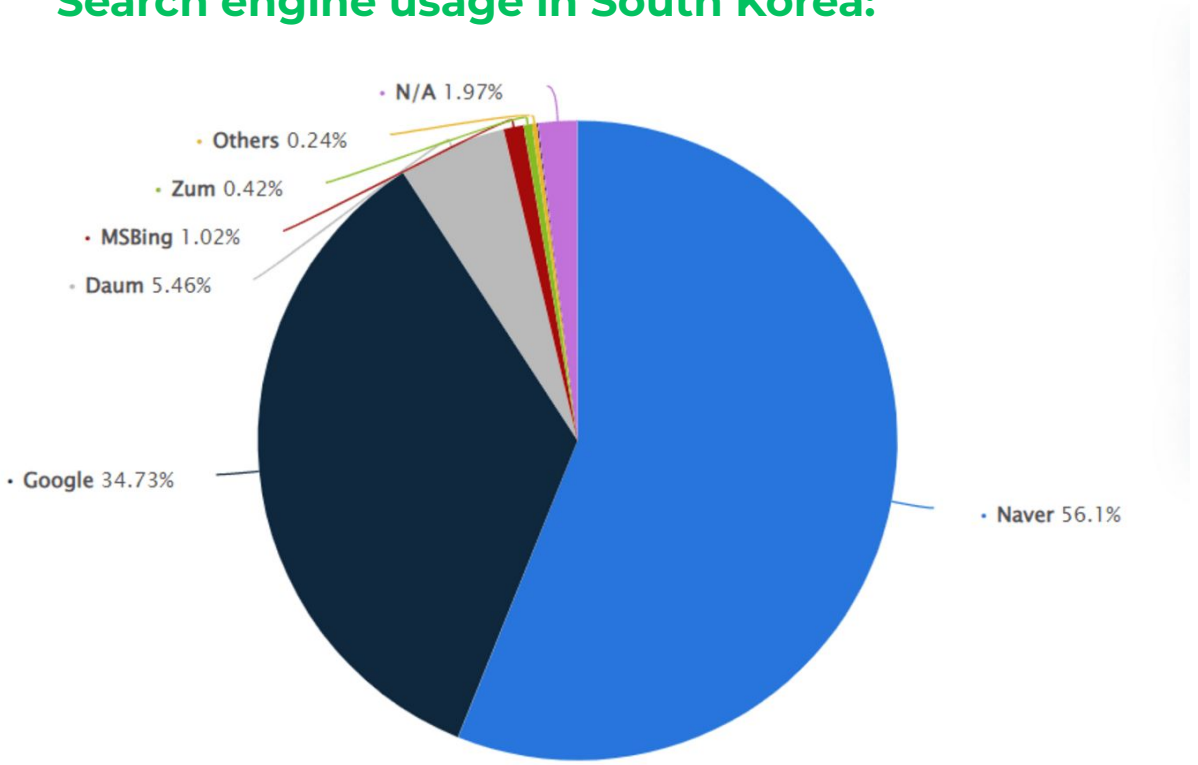


Huge collection of services.
Popular examples:

NAVER Cloud

LINE

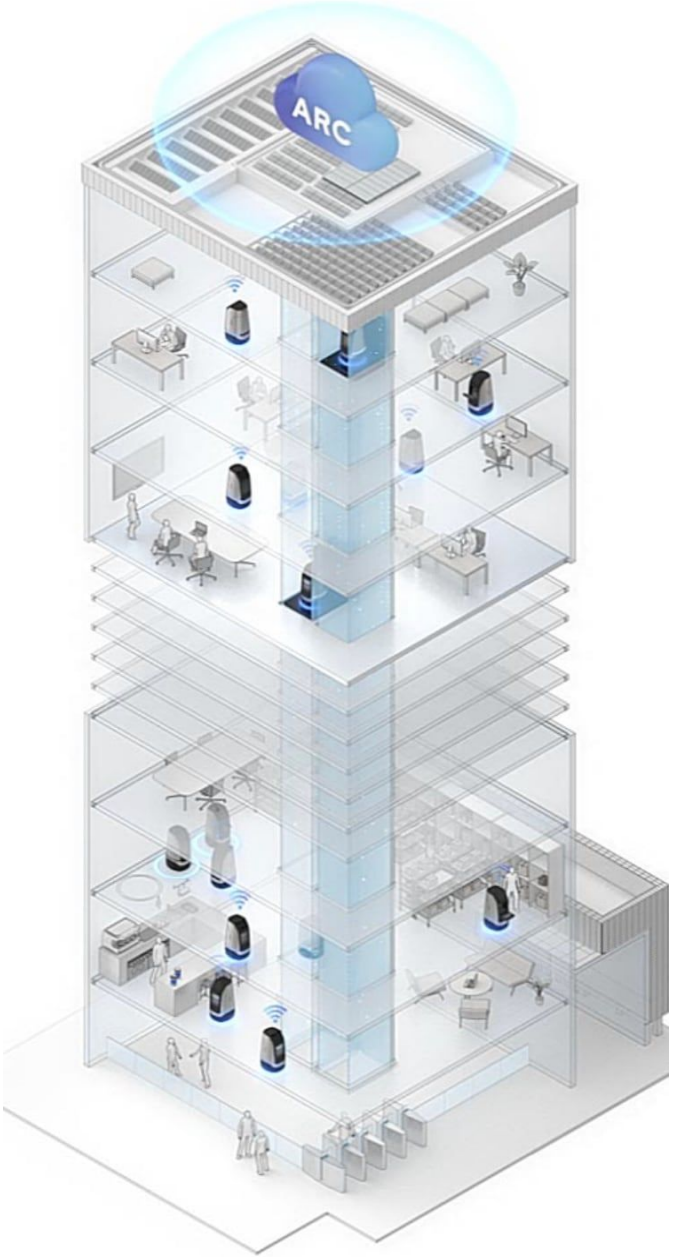
Search engine usage in South Korea:



2021, Source: https://www.link-assistant.com/news/naver-vs-google-in-korea.html

NAVER LABS

Adaptable robots for human environments



NAVER LABS ⇨ NAVER LABS Europe

- NAVER LABS Europe is a **fundamental research center**
- **Interactive Systems group** aims to equip robots with interaction (speech, text, gesture, etc)

This presentation is about (end-to-end) speech LLMs!

1. Quick recap on speech LLMs
2. **IWSLT 25 System**: best *short* instruction-following model
3. **SpeechMapper**: LLM-free speech projection training
4. Concluding remarks

A brief overview on Speech LLMs

Grounding LLMs in speech allows them to be more effective everyday assistants

For many applications, speech is more convenient than text:

- Robotics
- Home/Phone Assistants
- Embodied Systems

Speech is our instinctive communication channel:
when you fall downstairs, you scream, not text!

How can we add the speech modality to an LLM?

How can we add the speech modality to an LLM?

1. Cascading with an ASR module (no training required)

PROS

- ➔ LLM maintains its text capabilities
- ➔ Does not require training

How can we add the speech modality to an LLM?

1. Cascading with an ASR module (no training required)

PROS

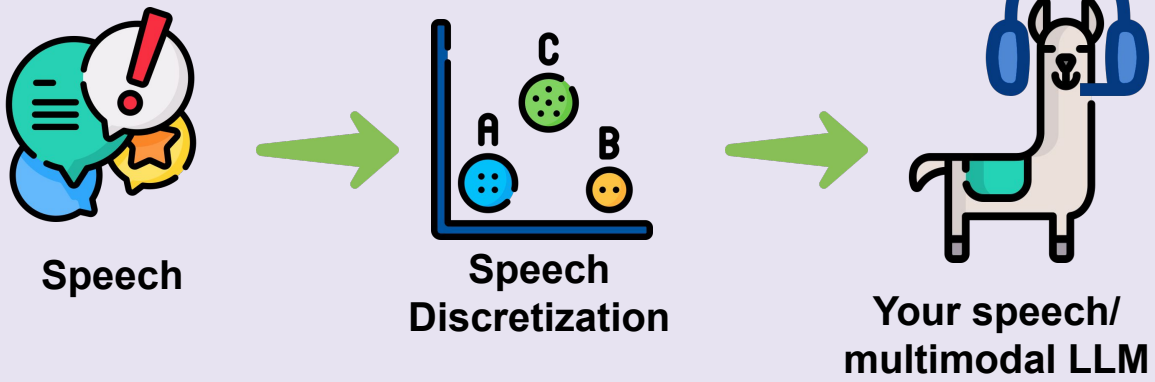
- ➔ LLM maintains its text capabilities
- ➔ Does not require training

CONS

- ➔ No acoustic information (e.g. emotion, speaker info)
- ➔ Error propagation
- ➔ Inference cost (ASR also requires an LM)

How can we add the speech modality to an LLM?

2. Discretization followed by multimodal training



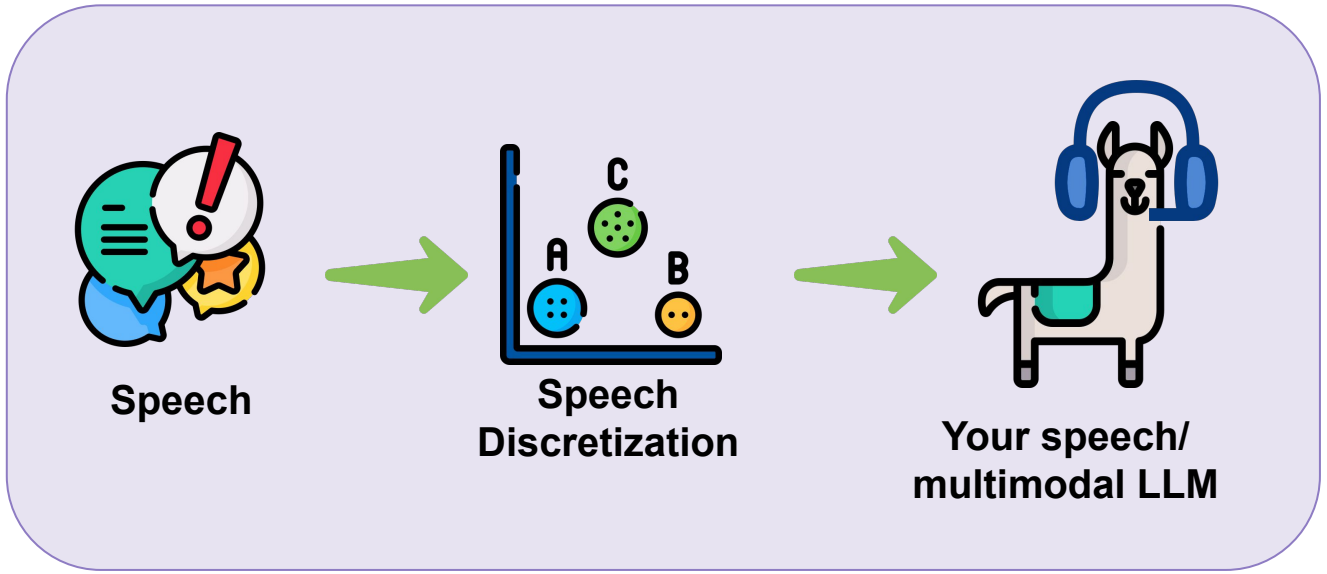
PROS

- ➔ Training on “text-like” input
- ➔ Speech encoding can be seen as **translation tasks**
- ➔ Acoustics ***potentially*** maintained

Examples: AudioPalm, SPIRIT LM, Moshi

How can we add the speech modality to an LLM?

2. Discretization followed by multimodal training



PROS

- ➔ Training on “text-like” input
- ➔ Speech encoding can be seen as **translation tasks**
- ➔ Acoustics ***potentially*** maintained

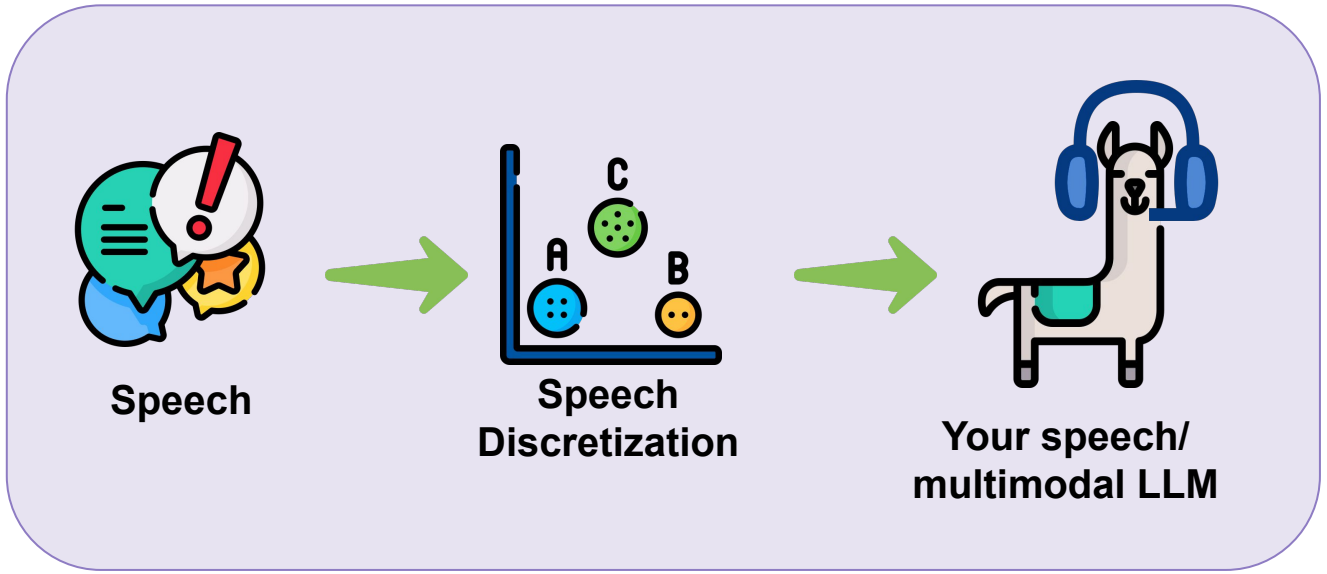
CONS

- ➔ Error propagation from discretizer
- ➔ Challenging to integrate speech modality without hurting text-based performance

Examples: AudioPalm, SPIRIT LM, Moshi

How can we add the speech modality to an LLM?

2. Discretization followed by multimodal training



PROS

- ➔ Training on “text-like” input
- ➔ Speech encoding can be seen as **translation tasks**
- ➔ Acoustics ***potentially*** maintained

CONS

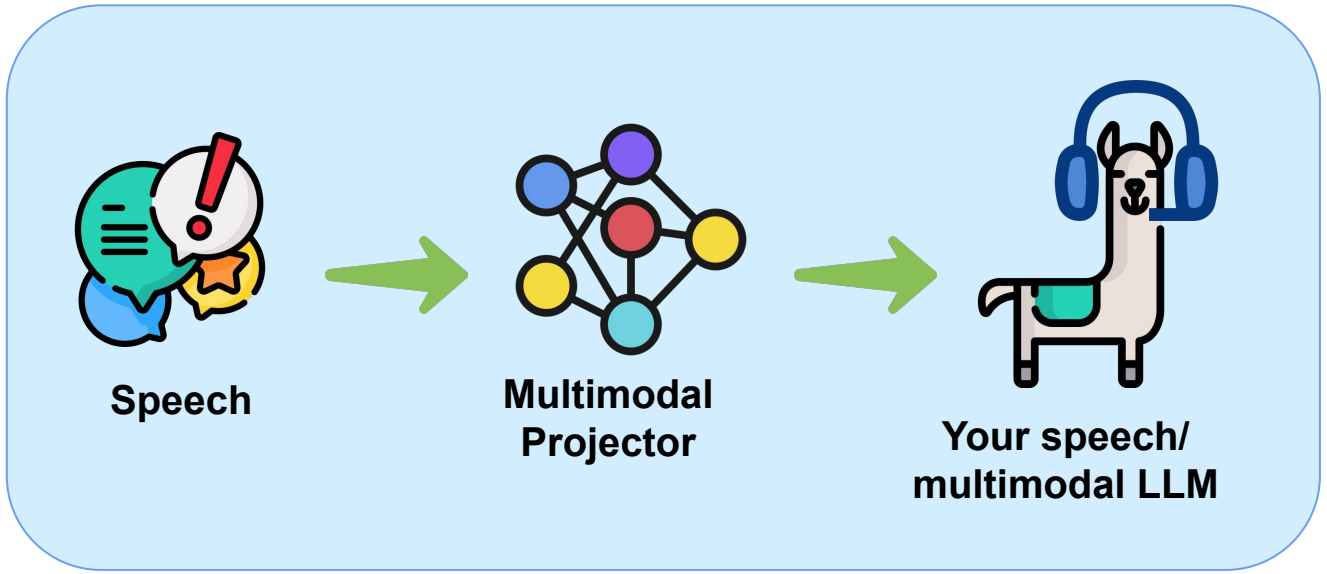
- ➔ Error propagation from discretizer
- ➔ Challenging to integrate speech modality without hurting text-based performance

Examples: AudioPalm, SPIRIT LM, Moshi

Check our work SPIRE: a from-English discrete speech LLM

How can we add the speech modality to an LLM?

3. End-to-end (continuous) training with masked multimodal input



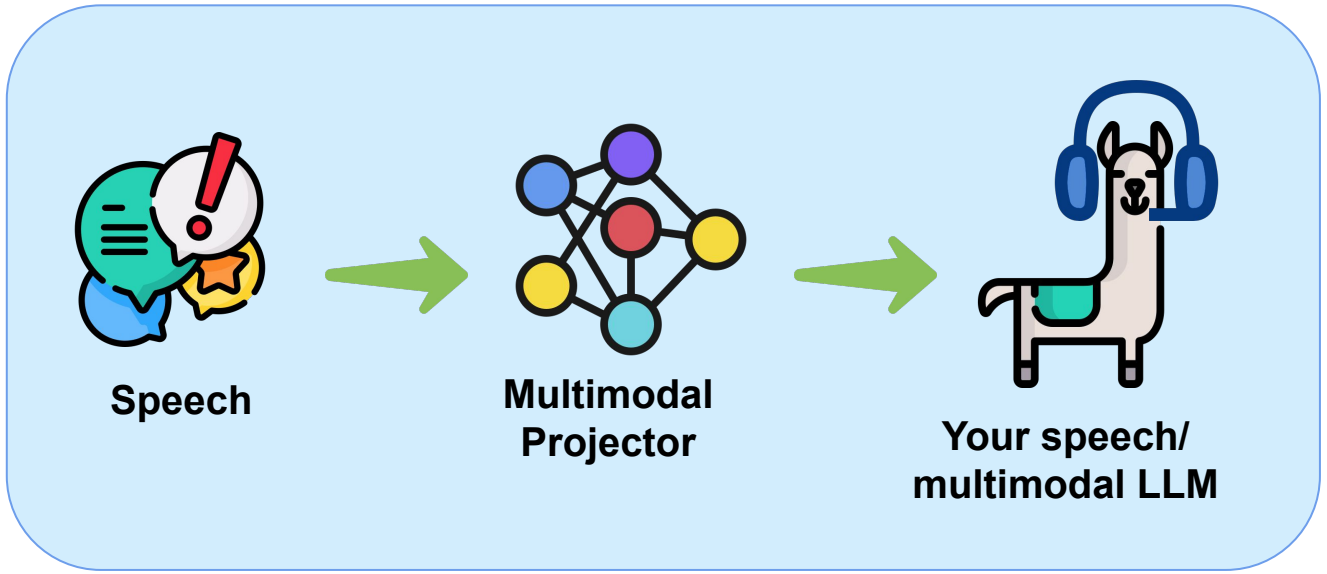
PROS

- ➔ No error propagation
- ➔ Acoustics ***potentially*** maintained
- ➔ Cheaper inference than cascading, potentially cheaper than discretizing

Examples: WavLLM, SALMONN, Wav2Prompt

How can we add the speech modality to an LLM?

3. End-to-end (continuous) training with masked multimodal input



PROS

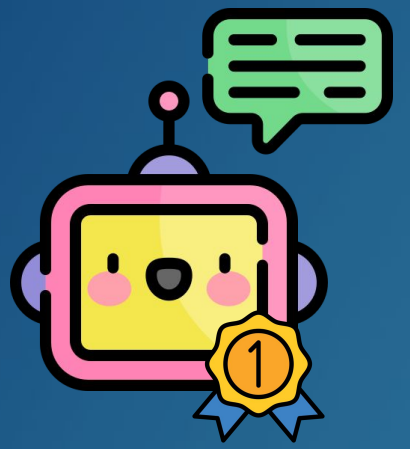
- ➔ No error propagation
- ➔ Acoustics ***potentially*** maintained
- ➔ Cheaper inference than cascading, potentially cheaper than discretizing

CONS

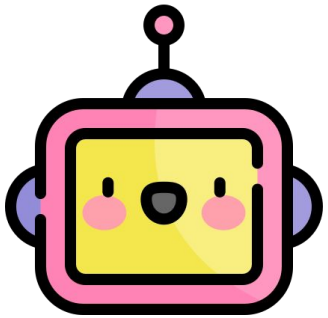
- ➔ Costly training for speech-to-text, even more costly for text-to-speech

Examples: WavLLM, SALMONN, Wav2Prompt

IWSLT 25: A multilingual continuous speech LLM



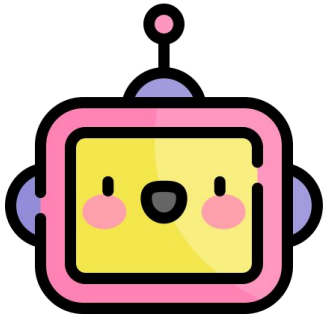
Instruction Following Challenge



Instruction Following Challenge

Task 1: Automatic Speech Transcription (ASR)

Speech Segment



Instruction: Can you transcribe the content in English text?

Output :
The town is also the site of a sausage festival.

Instruction Following Challenge

Task 2: Speech Translation (ST)

Speech Segment

Instruction: 你能把演讲内容翻译成中文吗?

Instruction: Können Sie den Inhalt der Rede in den deutschen Text übersetzen?

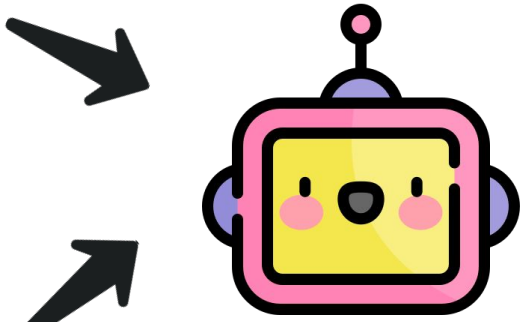
Output 1:
该镇也是香肠节的举办地。

Output 2:
Die Stadt ist auch der Austragungsort eines Würstchenfests.

Instruction Following Challenge

Task 3: Multilingual Spoken Question Answering (SQA)

Speech Segment



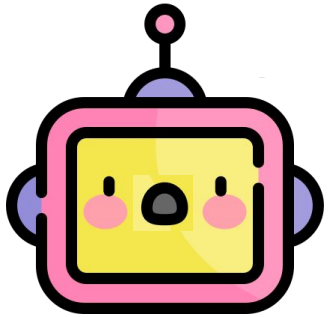
Instruction: ***Based on the speech segment***, can you answer the following question: Is this town mentioned the host of any particular events?

Output:
Yes. The town mentioned is the site of sausage festival.

Instruction Following Challenge

Task 3: Multilingual Spoken Question Answering (SQA)

Speech Segment



Instruction: ***Based on the speech segment***, can you answer the following question: How do I cook spaghetti?

Instruction: ***Based on the speech segment***, can you answer the following question: 我怎么做意大利面？

Output 1:
Not answerable.

Output 2:
无法回答。

Instruction Following Challenge

Why is this challenge ambitious?

- Requires the speech assistant to answer in the languages of the question
- Requires adaptation to the scientific domain and different English accents at test time (no in-domain data)
- **Controlled question answering setting:** specific answer in case of invalid questions

Instruction Following Challenge

Why is this challenge ambitious?

- Requires the speech assistant to answer in the languages of the question
- Requires adaptation to the scientific domain and different English accents at test time (no in-domain data)
- **Controlled question answering setting:** specific answer in case of invalid questions

Constrained setting: No multilingual SQA dataset provided for training, but backbones can be used to synthesize data

Creating a multilingual SQA training data

Complex work of data synthesis and filtering from existing English-only SpokenSQuAD SQA dataset:

- **Speech resynthesis** using Seamless with a random pull of speakers
  - Single TTS speaker models were changing behavior based on the voice
  - Fixed some training data misalignment

Creating a multilingual SQA training data

Complex work of data synthesis and filtering from existing English-only SpokenSQuAD SQA dataset:

- **Speech resynthesis** using Seamless with a random pull of speakers
  - Single TTS speaker models were changing behavior based on the voice
  - Fixed some training data misalignment

- **Answer rewriting** using Llama followed by LID (*fluent* SQA)
  - **Slot-based SQA is not the task we want to learn!**

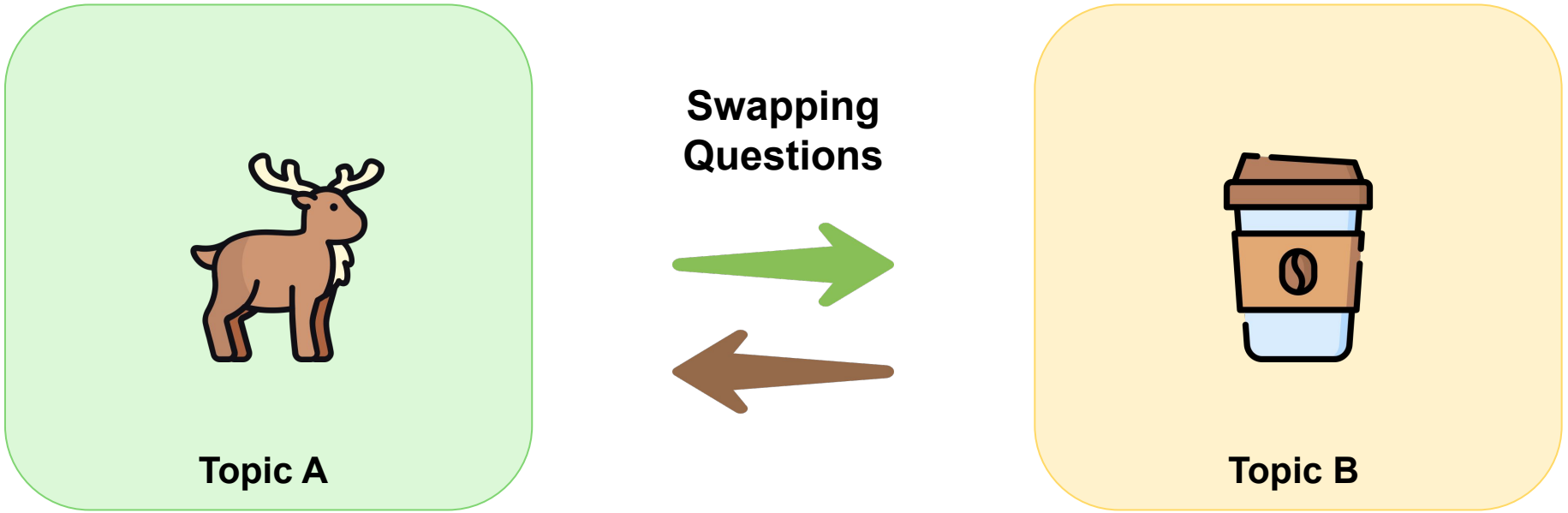
Creating a multilingual SQA training data

Complex work of data synthesis and filtering from existing English-only SpokenSQuAD SQA dataset:

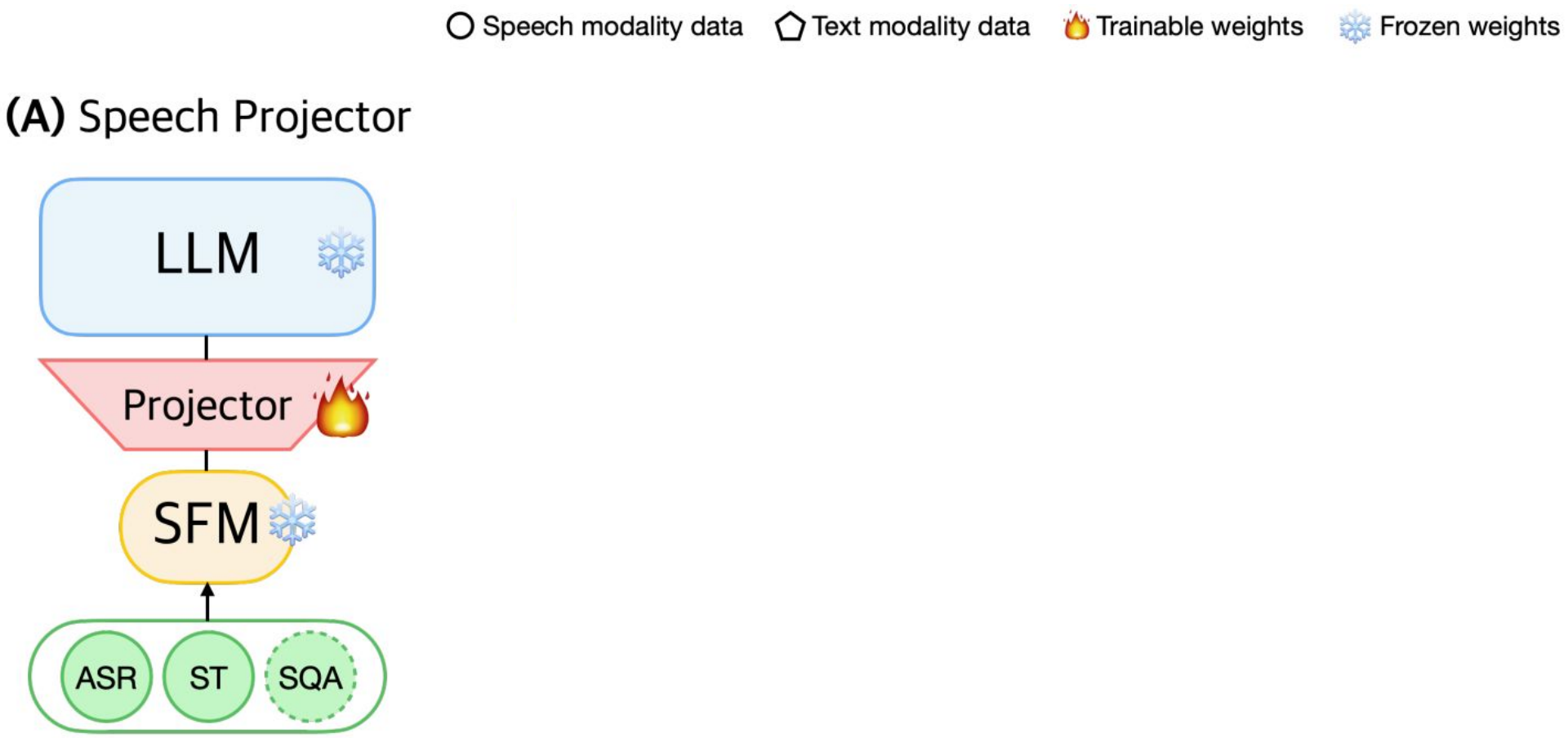
- **Speech resynthesis** using Seamless with a random pull of speakers
  - Single TTS speaker models were changing behavior based on the voice
  - Fixed some training data misalignment
- **Answer rewriting** using Llama followed by LID (*fluent* SQA)
  - **Slot-based SQA is not the task we want to learn!**
- **Question/Answering translation** followed by automatic translation quality filters using COMET

Creating a multilingual SQA training data

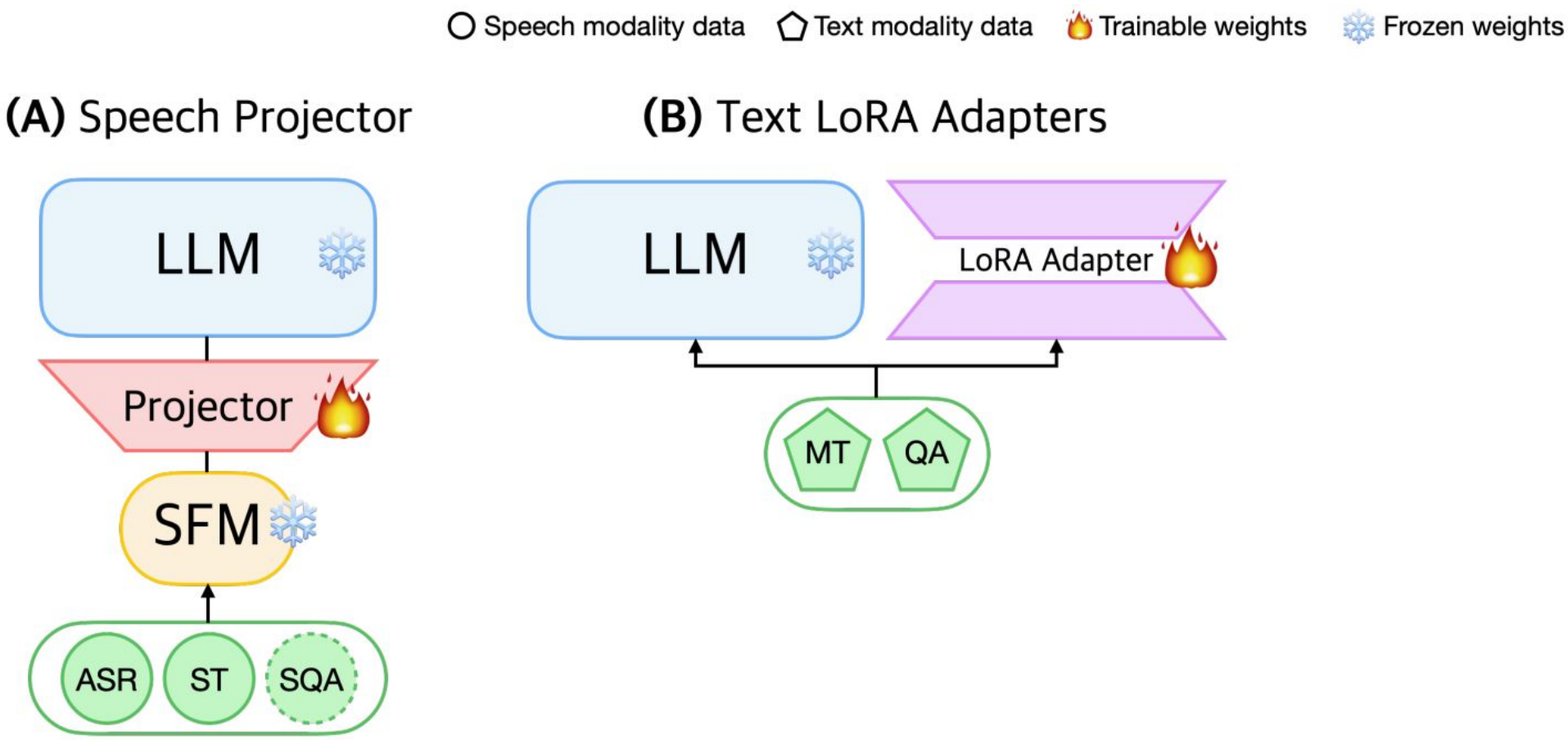
Creating **unanswerable examples** by swapping questions and changing the answer to “Not answerable”



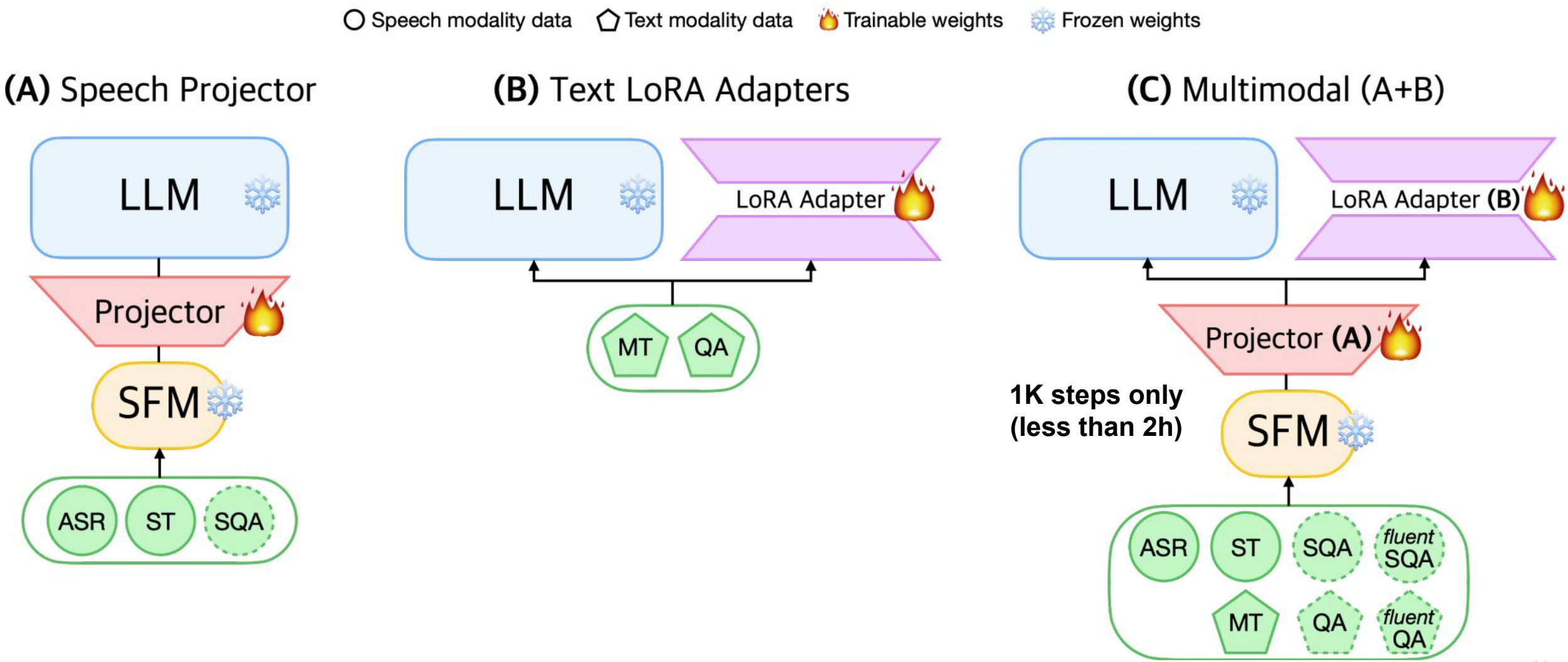
Training



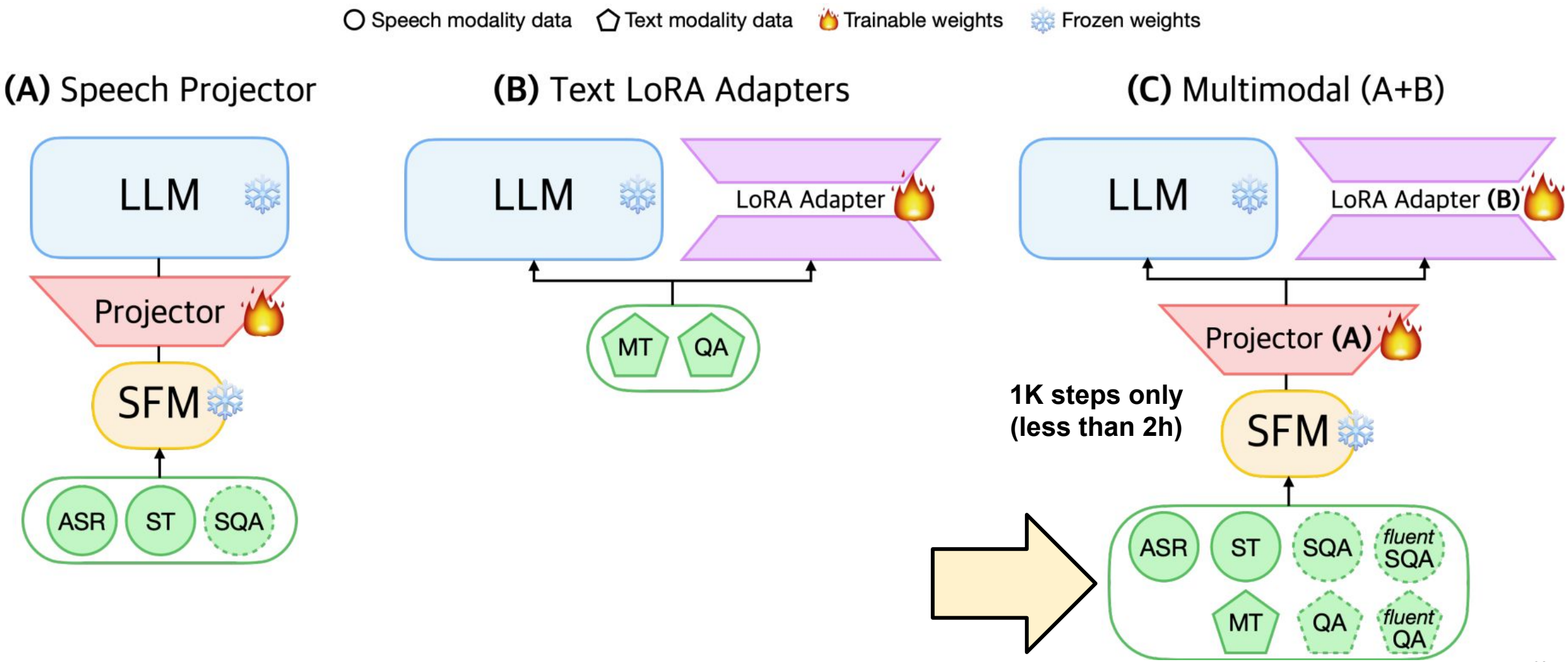
Training



Training



Training



Evaluation Metrics

BASELINES

- **Text:** Llama 3.1 8B
- **Speech:** SeamlessM4T large v2

DATASETS

- **ASR/ST:** ACL 60-60 (italian was automatically obtained)
- **SQA:** SpokenSQuAD
  - Original test-set for English
  - Automatically obtained (translation+COMET filter) for other languages
  - Non-answerable set for all languages (not shown here because accuracy is always around 99%)

METRICS

- **ASR:** WER
- **ST:** BLEU4/COMET
- **SQA:** LLM-as-judge (average over 4 models)

Results for the text toplines

Model (fine-tuning tasks)	ASR (WER)	ST/MT (BLEU)			ST/MT (COMET)			SQA/QA (LLM-AS-A-JUDGE)			
	en	en-de	en-it	en-zh	en-de	en-it	en-zh	en-en	en-de	en-it	en-zh
Text-only Models (MT/QA)											
`Llama-3.1-8B-Instruct` (zero-shot)	-	23.88	35.51	45.89	0.779	0.806	0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69**	**48.31**	**53.65**	**0.838**	**0.863**	**0.867**	83.4%	75.7%	71.4%	69.5%

It's more about format following than true performance gain

- MT performance increases because the model includes less rubbish in the answer
- QA performance decreases because the slot format is less natural and therefore penalized by the evaluation

Results for projector-only

Model (fine-tuning tasks)	ASR (WER) en	ST/MT (BLEU) en-de	ST/MT (BLEU) en-it	ST/MT (BLEU) en-zh	ST/MT (COMET) en-de	ST/MT (COMET) en-it	ST/MT (COMET) en-zh	SQA/QA (LLM-AS-A-JUDGE) en-en	SQA/QA (LLM-AS-A-JUDGE) en-de	SQA/QA (LLM-AS-A-JUDGE) en-it	SQA/QA (LLM-AS-A-JUDGE) en-zh
Text-only Models (MT/QA)											
`Llama-3.1-8B-Instruct` (zero-shot)	-	23.88	35.51	45.89	0.779	0.806	0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69**	**48.31**	**53.65**	**0.838**	**0.863**	**0.867**	83.4%	75.7%	71.4%	69.5%
Speech-only Models (ASR/ST/SQA)											
`SeamlessM4T-v2-large`	**17.6**	**27.95**	**43.54**	33.58	0.737	0.788	0.753	-	-	-	-
A.1 Speech Projector (ASR/ST)	19.8	27.58	36.30	40.62	**0.760**	0.796	**0.793**	-	-	-	-
A.2 Speech Projector (ASR/ST/SQA)	19.9	27.20	36.60	**40.72**	**0.760**	**0.797**	0.792	0.7%	0.5%	0.3%	0.6%

- WER of all models is high compared to their performance on training datasets (EuroParlST, CoVoST2).

Results for projector-only

Model (fine-tuning tasks)	ASR (WER) en	ST/MT (BLEU) en-de	ST/MT (BLEU) en-it	ST/MT (BLEU) en-zh	ST/MT (COMET) en-de	ST/MT (COMET) en-it	ST/MT (COMET) en-zh	SQA/QA (LLM-AS-A-JUDGE) en-en	SQA/QA (LLM-AS-A-JUDGE) en-de	SQA/QA (LLM-AS-A-JUDGE) en-it	SQA/QA (LLM-AS-A-JUDGE) en-zh
Text-only Models (MT/QA)											
`Llama-3.1-8B-Instruct` (zero-shot)	-	23.88	35.51	45.89	0.779	0.806	0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69**	**48.31**	**53.65**	**0.838**	**0.863**	**0.867**	83.4%	75.7%	71.4%	69.5%
Speech-only Models (ASR/ST/SQA)											
`SeamlessM4T-v2-large`	**17.6**	**27.95**	**43.54**	33.58	0.737	0.788	0.753	-	-	-	-
A.1 Speech Projector (ASR/ST)	19.8	27.58	36.30	40.62	**0.760**	0.796	**0.793**	-	-	-	-
A.2 Speech Projector (ASR/ST/SQA)	19.9	27.20	36.60	**40.72**	**0.760**	**0.797**	0.792	0.7%	0.5%	0.3%	0.6%

We investigated why:

- Audios not properly cropped
- style-shift in transcriptions
- challenge of NE
- LLM rephrasing

Results for projector-only solutions

Model (fine-tuning tasks)	ASR (WER) en	ST/MT (BLEU) en-de	ST/MT (BLEU) en-it	ST/MT (BLEU) en-zh	ST/MT (COMET) en-de	ST/MT (COMET) en-it	ST/MT (COMET) en-zh	SQA/QA (LLM-AS-A-JUDGE) en-en	SQA/QA (LLM-AS-A-JUDGE) en-de	SQA/QA (LLM-AS-A-JUDGE) en-it	SQA/QA (LLM-AS-A-JUDGE) en-zh
Text-only Models (MT/QA)											
`Llama-3.1-8B-Instruct` (zero-shot)	-	23.88	35.51	45.89	0.779	0.806	0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69**	**48.31**	**53.65**	**0.838**	**0.863**	**0.867**	83.4%	75.7%	71.4%	69.5%
Speech-only Models (ASR/ST/SQA)											
`SeamlessM4T-v2-large`	**17.6**	**27.95**	**43.54**	33.58	0.737	0.788	0.753	-	-	-	-
A.1 Speech Projector (ASR/ST)	19.8	27.58	36.30	40.62	**0.760**	0.796	**0.793**	-	-	-	-
A.2 Speech Projector (ASR/ST/SQA)	19.9	27.20	36.60	**40.72**	**0.760**	**0.797**	0.792	0.7%	0.5%	0.3%	0.6%

- ST performance on pair with Seamless for ACL 60-60
- **Models are not capable of slot-based SQA** (they only repeat training examples)

Results for projector-only solutions

Model (fine-tuning tasks)	ASR (WER)	ST/MT (BLEU)			ST/MT (COMET)			SQA/QA (LLM-AS-A-JUDGE)			
	en	en-de	en-it	en-zh	en-de	en-it	en-zh	en-en	en-de	en-it	en-zh
Text-only Models (MT/QA)											
`Llama-3.1-8B-Instruct` (zero-shot)	-	23.88	35.51	45.89	0.779	0.806	0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69**	**48.31**	**53.65**	**0.838**	**0.863**	**0.867**	83.4%	75.7%	71.4%	69.5%
Speech-only Models (ASR/ST/SQA)											
`SeamlessM4T-v2-large`	**17.6**	**27.95**	**43.54**	33.58	0.737	0.788	0.753	-	-	-	-
A.1 Speech Projector (ASR/ST)	19.8	27.58	36.30	40.62	**0.760**	0.796	**0.793**	-	-	-	-
A.2 Speech Projector (ASR/ST/SQA)	19.9	27.20	36.60	**40.72**	**0.760**	**0.797**	0.792	0.7%	0.5%	0.3%	0.6%

- **Hypothesis 1:** SQA has poor synergy with ASR/ST due to the task requiring a different model behavior where the prompt is actually relevant
- **Hypothesis 2:** SQA cannot be properly learned using projection-only (no LoRA)

Results for projector-only solutions

Model (fine-tuning tasks)	ASR (WER) en	ST/MT (BLEU) en-de	ST/MT (BLEU) en-it	ST/MT (BLEU) en-zh	ST/MT (COMET) en-de	ST/MT (COMET) en-it	ST/MT (COMET) en-zh	SQA/QA (LLM-AS-A-JUDGE) en-en	SQA/QA (LLM-AS-A-JUDGE) en-de	SQA/QA (LLM-AS-A-JUDGE) en-it	SQA/QA (LLM-AS-A-JUDGE) en-zh
Text-only Models (MT/QA)											
`Llama-3.1-8B-Instruct` (zero-shot)	-	23.88	35.51	45.89	0.779	0.806	0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69**	**48.31**	**53.65**	**0.838**	**0.863**	**0.867**	83.4%	75.7%	71.4%	69.5%
Speech-only Models (ASR/ST/SQA)											
`SeamlessM4T-v2-large`	**17.6**	**27.95**	**43.54**	33.58	0.737	0.788	0.753	-	-	-	-
A.1 Speech Projector (ASR/ST)	19.8	27.58	36.30	40.62	**0.760**	0.796	**0.793**	-	-	-	-
A.2 Speech Projector (ASR/ST/SQA)	19.9	27.20	36.60	**40.72**	**0.760**	**0.797**	0.792	0.7%	0.5%	0.3%	0.6%

- ~~**Hypothesis 1:**~~ SQA has poor synergy with ASR/ST due to the task requiring a different model behavior where the prompt is actually relevant
- ~~**Hypothesis 2:**~~ SQA cannot be properly learned using projection-only (no LoRA)

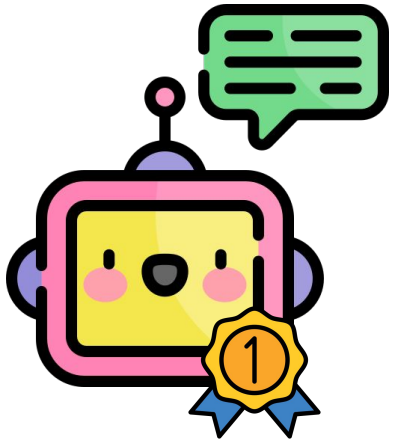
- **Hypothesis now:** Difficult task to learn from scratch! It's all about task upsampling and diversity!

Results for multimodal training

Model (fine-tuning tasks)	ASR (WER) en	ST/MT (BLEU) en-de	ST/MT (BLEU) en-it	ST/MT (BLEU) en-zh	ST/MT (COMET) en-de	ST/MT (COMET) en-it	ST/MT (COMET) en-zh	SQA/QA (LLM-AS-A-JUDGE) en-en	SQA/QA (LLM-AS-A-JUDGE) en-de	SQA/QA (LLM-AS-A-JUDGE) en-it	SQA/QA (LLM-AS-A-JUDGE) en-zh
Text-only Models (MT/QA)											
`Llama-3.1-8B-Instruct` (zero-shot)	-	23.88	35.51	45.89	0.779	0.806	0.809	**91.8%**	**92.0%**	**88.6%**	**84.6%**
B. Text-only LoRA (MT/QA)	-	**41.69**	**48.31**	**53.65**	**0.838**	**0.863**	**0.867**	83.4%	75.7%	71.4%	69.5%
Speech-only Models (ASR/ST/SQA)											
`SeamlessM4T-v2-large`	**17.6**	**27.95**	**43.54**	33.58	0.737	0.788	0.753	-	-	-	-
A.1 Speech Projector (ASR/ST)	19.8	27.58	36.30	40.62	**0.760**	0.796	**0.793**	-	-	-	-
A.2 Speech Projector (ASR/ST/SQA)	19.9	27.20	36.60	**40.72**	**0.760**	**0.797**	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	**41.22**	42.76	0.758	**0.791**	0.795	79.8%	71.9%	69.4%	65.5%
A.1 + B (ASR/ST/MT/*fluent*SQA/*fluent*QA)	18.6	**30.75**	40.48	42.51	0.755	0.788	0.789	90.3%	85.2%	82.9%	76.4%
A.2 + B (ASR/ST/MT/SQA/QA)	18.2	29.91	38.13	43.12	0.759	0.786	**0.799**	80.5%	74.9%	68.0%	66.7%
A.2 + B (ASR/ST/MT/*fluent*SQA/*fluent*QA)	18.7	29.68	32.28	**43.38**	**0.763**	0.782	0.798	**91.1%**	**87.3%**	**84.8%**	**78.0%**

- Better ASR scores, equivalent ST scores
- **Last phase seems to be relevant mainly for SQA**

Challenge Results Overview[1]

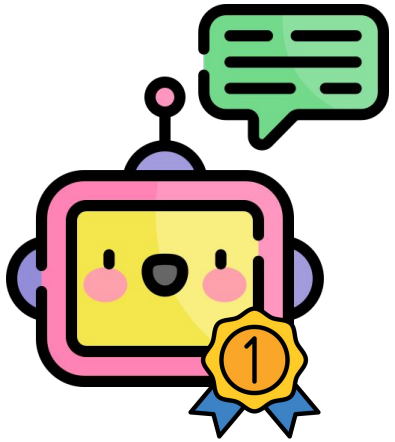


***Baseline* system: Microsoft-Phi** (trained in unknown amounts of data)

1. **ASR:** No system was able to beat the baseline. **NLE's submission was the best of the submitted models.**
2. **ST:** No system was able to beat the baseline. **NLE submission statistically tied with other systems for two languages (de, zh), best submission for Italian.**
3. **SQA: NLE system beats even Microsoft-Phi.**

[1]For full results, check SHORT table at page 70 of https://aclanthology.org/2025.iwslt-1.44.pdf

Challenge Results Overview[1]



Remarkably, we achieved these results as the only constrained submission.

Our system held its ground against models leveraging more powerful backbones and far larger training data.

[1]For full results, check SHORT table at page 70 of https://aclanthology.org/2025.iwslt-1.44.pdf

Does this model generalize?

Speech LLM papers talk a lot about *task and prompt overfitting*

Our experience with that was that:

➔ **ASR-only** speech LLMs were unable to perform different tasks, no performance degradation changing the prompt since task-overfitted

Does this model generalize?

Speech LLM papers talk a lot about *task and prompt overfitting*

Our experience with that was that:

- ➔ **ASR-only** speech LLMs were unable to perform different tasks, no performance degradation changing the prompt since task-overfitted
- ➔ **ASR+ST** speech LLMs were unable to perform SQA, prompt confusion existed, ***limited*** performance degradation when changing the prompt **(changing the prompt language helped)**

Does this model generalize?

Speech LLM papers talk a lot about *task and prompt overfitting*

Our experience with that was that:

- ➔ **ASR-only** speech LLMs were unable to perform different tasks, no performance degradation changing the prompt since task-overfitted
- ➔ **ASR+ST** speech LLMs were unable to perform SQA, prompt confusion existed, ***limited*** performance degradation when changing the prompt **(changing the prompt language helped)**
- ➔ ASR+ST+SQA speech LLMs **were able to generalize to new prompt formats and even languages**

Does this model generalize?

Speech LLM papers talk a lot about *task and prompt overfitting*

SETTING: instruction in the target language
METRIC: COMET

	EuroParl				CoVoST2	
	en-es	en-fr	en-de	en-it	en-de	en-zh
Transcripts + EuroLLM 9B (topline)	85.9	85.0	82.5	86.0	78.3	80.0
Transcripts + Llama 3.1 8B (topline)	82.8	81.0	81.2	84.1	82.0	77.0
Seamless ST (in-domain)	80.4	74.8	70.0	76.0	83.0	82.0
BEST-IWST25-IF (in-domain)	83.5	81.1	84.0	86.0	78.9	80.7

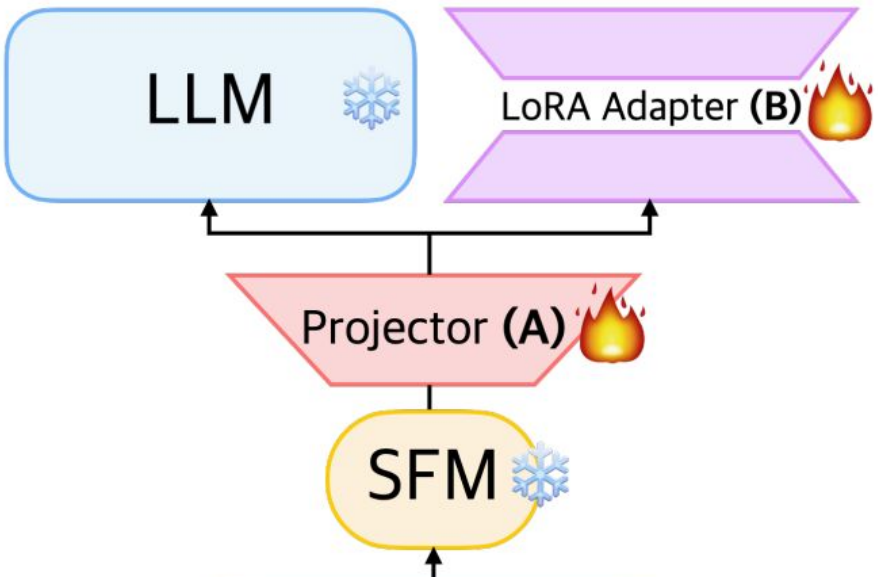
My take on this: if the prompt is not diverse or "interesting enough", the model will encode the task on the projected representation, instead of relying on it!

+ Hemant
+ Biswesh

Currently under-review at ICASSP'26

SpeechMapper: LLM-free speech projection training

Speech LLMs Training

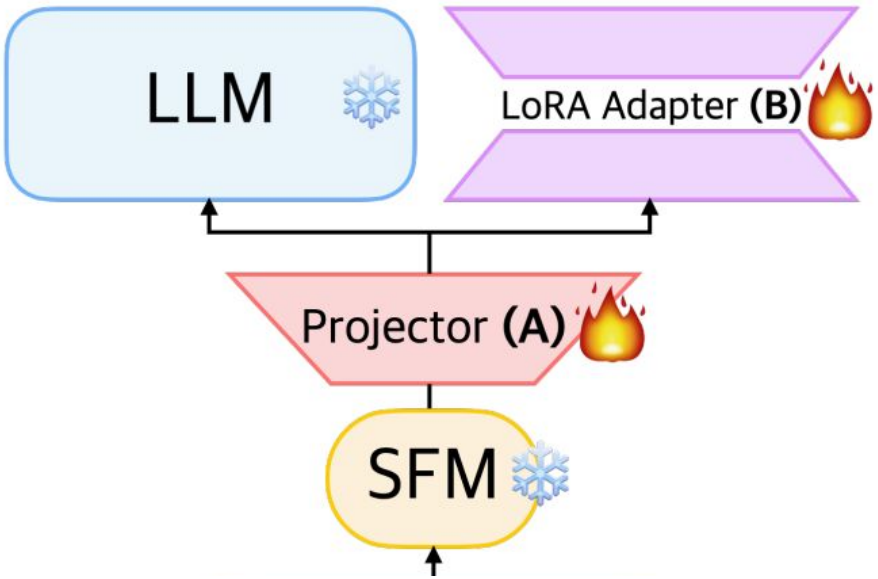


Backbones: Usually frozen because very slow to tune

LoRa adapters: can be included in both SFM and LLMs

Projector: Mandatory (at least an MLP)

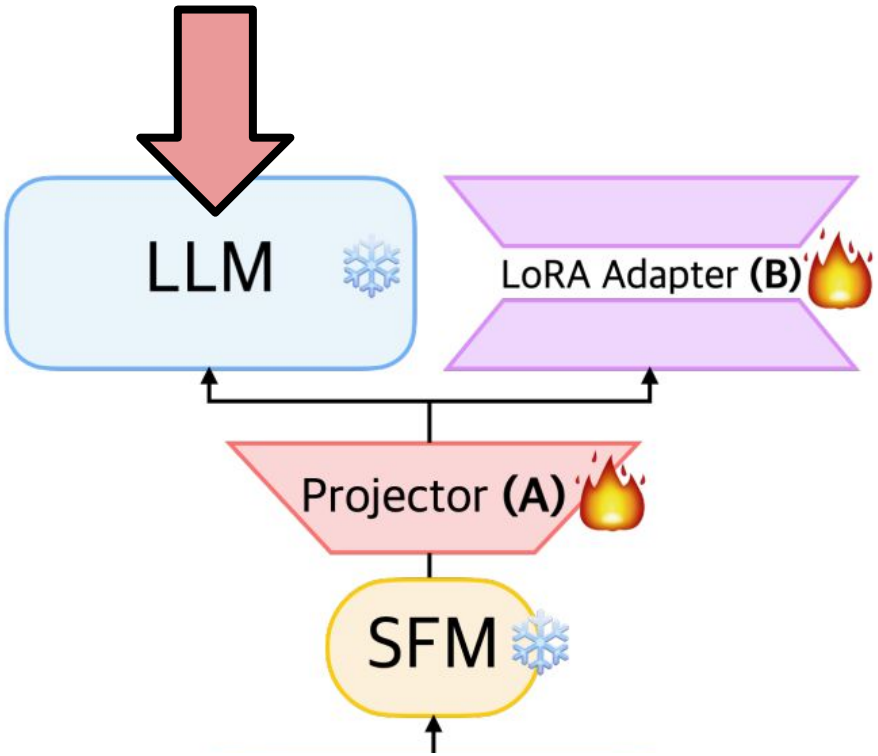
Speech LLMs Training



Bottlenecks:

1. Slow to train due to the long audio sequences + deep forward pass
2. Limit on the LLM size: very complex to train larger-than-8B speech LLMs
3. CE causes prompt and task overfitting

Speech LLMs Training



Bottlenecks:

1. Slow to train due to the long audio sequences + deep forward pass
2. Limit on the LLM size: very complex to train larger-than-8B speech LLMs
3. **CE causes prompt and task overfitting**

Speech LLMs Training

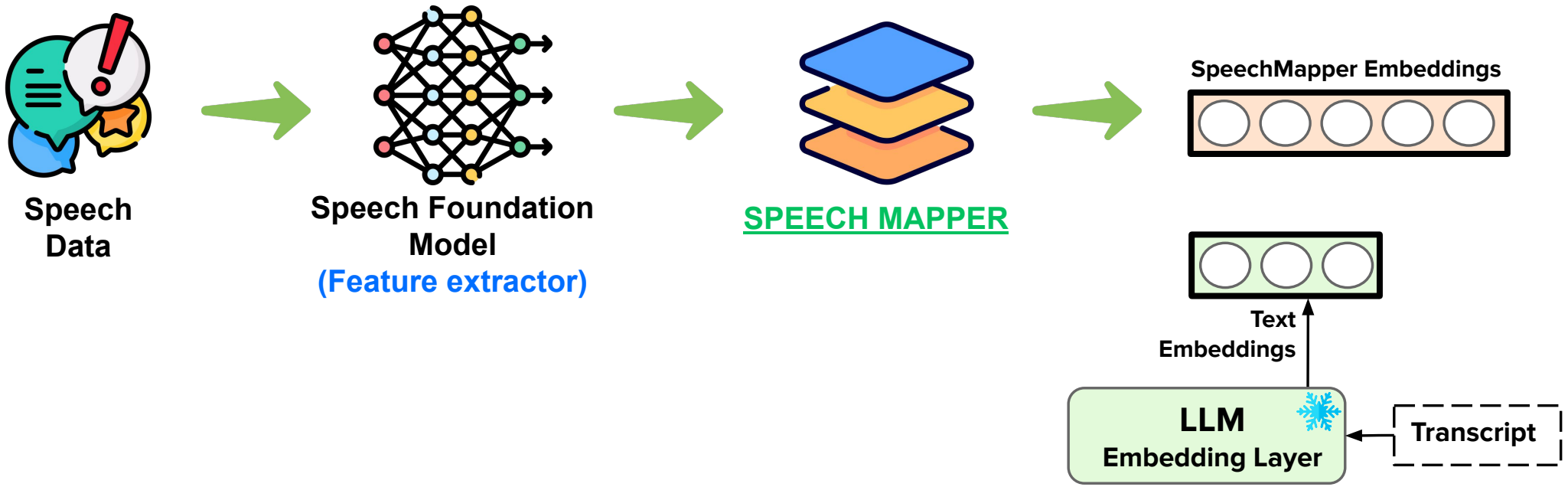
How to make a task- and prompt-agnostic speech projector for LLMs?

How to reduce hardware and data limitations for training theses models?

How to design a solution for easily switching between speech and text input?

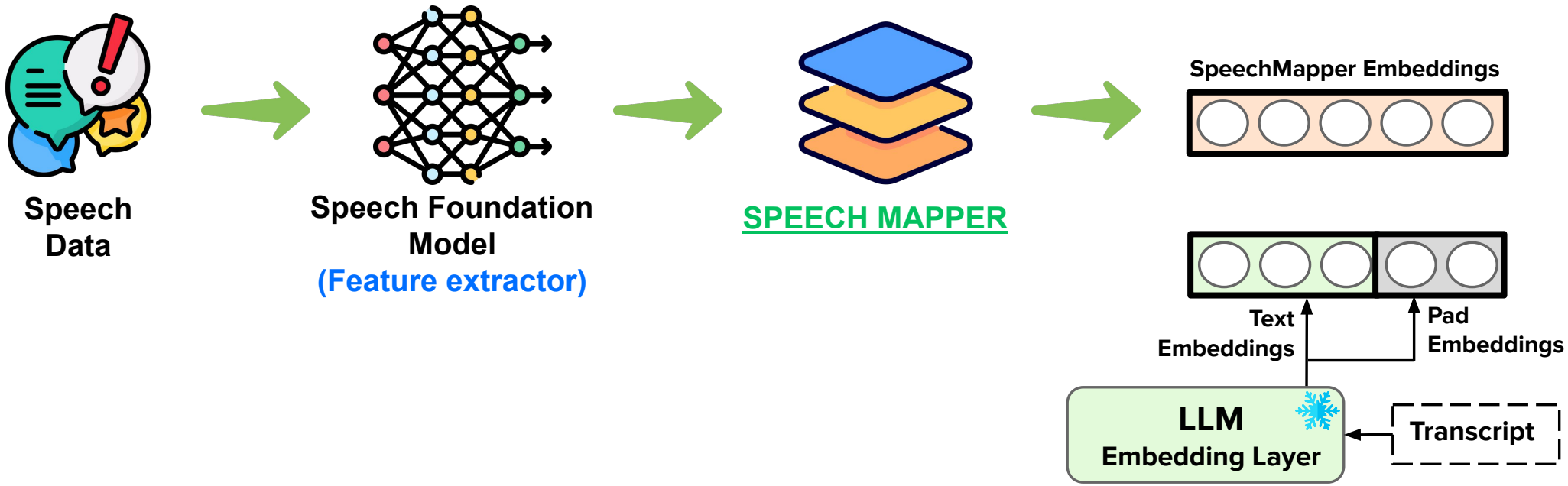
SpeechMapper: Removing the dependency on the LLM forward pass

★ Speech-to-Embedding approach



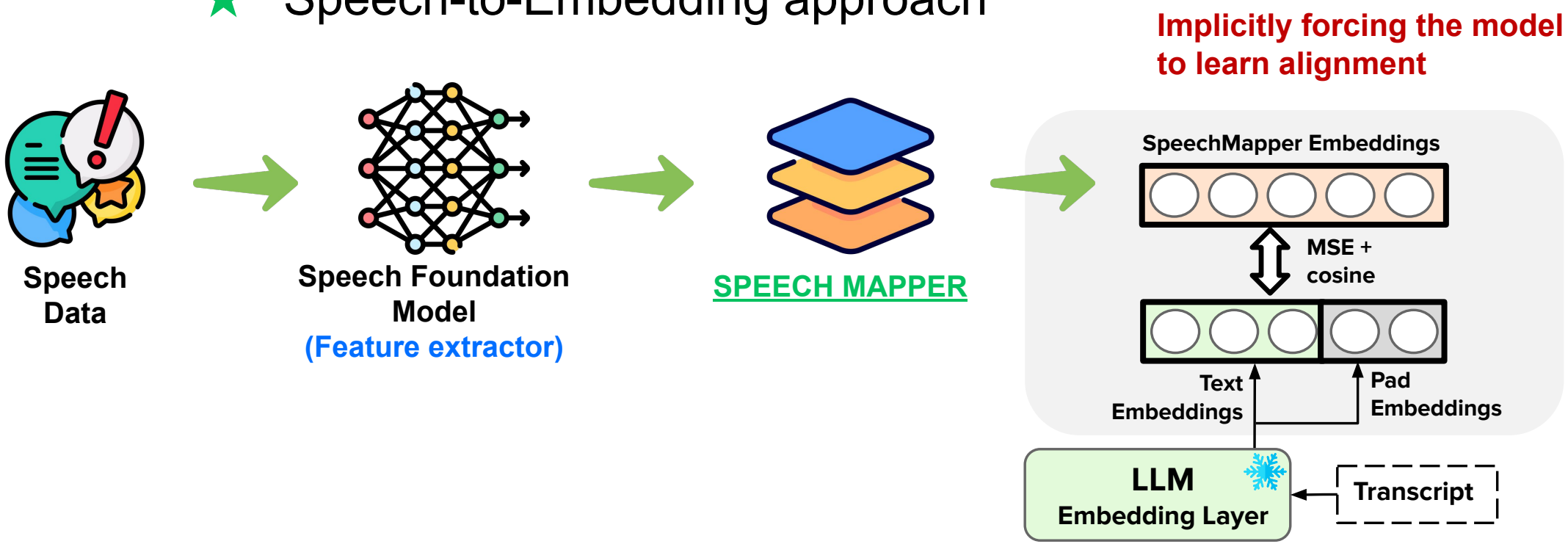
SpeechMapper: Removing the dependency on the LLM forward pass

★ Speech-to-Embedding approach



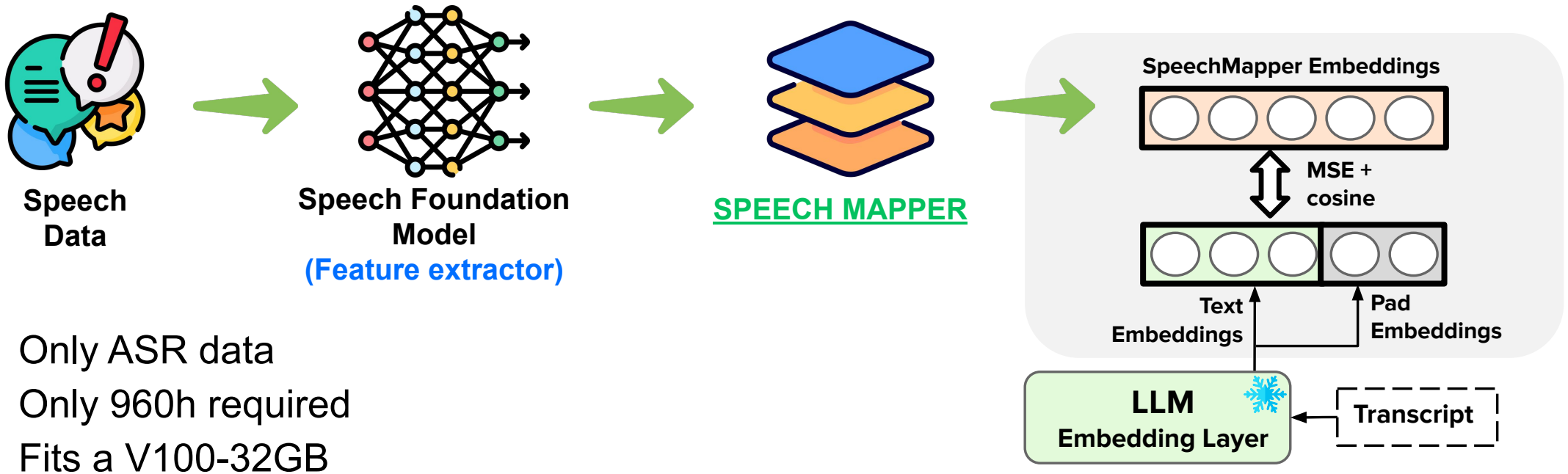
SpeechMapper: Removing the dependency on the LLM forward pass

★ Speech-to-Embedding approach



SpeechMapper: Removing the dependency on the LLM forward pass

★ Speech-to-Embedding approach



- ➔ Only ASR data
- ➔ Only 960h required
- ➔ Fits a V100-32GB
- ➔ Trains in 4 days (V1) or 18h (V2)

SpeechMapper: Removing the dependency on the LLM forward pass

★ Speech-to-Embedding approach

Same training for any LLM size!

- → Only ASR data
- → Only 960h required
- → Fits a V100-32GB
- → Trains in 4 days (V1) or 18h (V2)

IWSLT'25 versus SpeechMapper

- SQA task, accuracy metric (higher is better)
- LLM-as-judge setup with average across 4 LLMs
- All models share the same backbone

	SpokenSQuAD	LibriSQA-I	LibriSQA-II
IWSLT'25 (SQA fine-tuned)	87.4	80.7	62.5
SpeechMapper v1 (zero-shot)	72.9	75.6	68.5
SpeechMapper v1 (SQA fine-tuned)	**89.0**	**82.5**	72.9
SpeechMapper v2 WIP (zero-shot)	85.6	80.8	**76.2**

IWSLT'25 versus SpeechMapper

- ST task, COMET metric (higher is better)
- All models share the same backbone

	EuroParl ST				CoVoST2	
	en-es	en-fr	en-de	en-it	en-de	en-zh
IWSLT'25 (ST fine-tuned)	83.5	81.1	**84.0**	**86.0**	**78.9**	**80.7**
SpeechMapper v1 (zero-shot)	74.6	72.1	70.0	73.1	61.7	66.1
SpeechMapper v1 (ST fine-tuned)	**84.7**	**82.3**	80.7	84.4	75.5	78.5
SpeechMapper v2 WIP (zero-shot)	83.0	80.0	78.6	81.8	75.3	77.9

A push for simplicity

- Even less training data (960h instead of 2k)
- Simpler training regime, no prompt or task overfitting
- And still maintaining the backbone’s text-based performance
- Scalable to larger LLMs sizes

Concluding

This talk focused on continuous Speech LLMs for semantic tasks

★ First part focused on sharing our bests tricks from IWSLT'25

★ Second part briefly covered a *smarter* projection architecture for semantic tasks called SpeechMapper

★ But this is only about speech semantics! Acoustics missing!

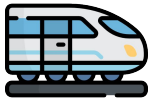
This talk focused on continuous Speech LLMs for semantic tasks

★ First part focused on sharing our bests tricks from IWSLT'25

★ Second part briefly covered a *smarter* projection architecture for semantic tasks called SpeechMapper

★ But this is only about speech semantics! Acoustics missing!

You should probably not train a huge LLM just to replace a 1B ASR!



Where should we go from here?

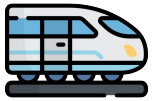
We need more multimodal benchmarks! We need complex instructions!

PunchBench: Benchmarking MLLMs in Multimodal Punchline Comprehension

Kun Ouyang[†‡], Yuanxin Liu[†], Shicheng Li[†],
Yi Liu[†], Hao Zhou[‡], Fandong Meng[‡], Jie Zhou[‡], Xu Sun[†*]

MULTIMODAL CROSSLINGUAL INSTRUCTION-FOLLOWING BENCHMARK FROM SCIENTIFIC TALKS

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu,
Ioannis Douros, Luisa Bentivogli, Jan Niehues



Where should we go from here?

It’s not very creative but… **let’s get even more multimodal!**

- ➔ Acoustics
- ➔ Images
- ➔ Videos
- ➔ 3D information
- ➔ Pose estimation

Thanks for listening!

Happy holidays!

12/2025

Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS

Results for projector-only

Model (fin	JUDGE) t	en-zh
Llama-3.	%	84.6%
B. Text-onl	%	69.5%
Seamless		-
A.1 Speech		-
A.2 Speech	%	0.6%

An example from the test set:

Audios not properly cropped, style-shift in transcriptions, challenge of NE, rephrasing of LLMs

Reference:
"So we further investigate the results on **SVAMP**."
"And this dataset is challenging because the author tried to manually **ah adding** something to confuse the NLP model **like such as** adding irrelevant information and extra quantities."

Generated:
"So we further investigate the results on." **(audio cropped, SVAMP in the next segment)**
"**swamp**, and this dataset is challenging because the author tried to manually **add** something to confuse the NLB model **like** adding environmental information and extra quantities."