SpeechMapper: Efficient training of speech LLMs

Marcely Zanon Boito

06/2026

Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS

About me

(2021) PhD in Computer Science at University Grenoble Alpes

- Low-resource speech processing
- Unsupervised word segmentation
- Speech discretization

(2021-2022) Postdoc at Avignon University

- Low-resource Speech Translation
- Self-Supervised Learning for Speech

NAVER LABS Europe

(Since 2022) Research Scientist at NAVER LABS Europe

- Multimodality and Speech Processing

This presentation is about end-to-end speech LLMs

1. Quick recap on speech LLMs
2. SpeechMapper **v1** & **v2**
3. 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

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

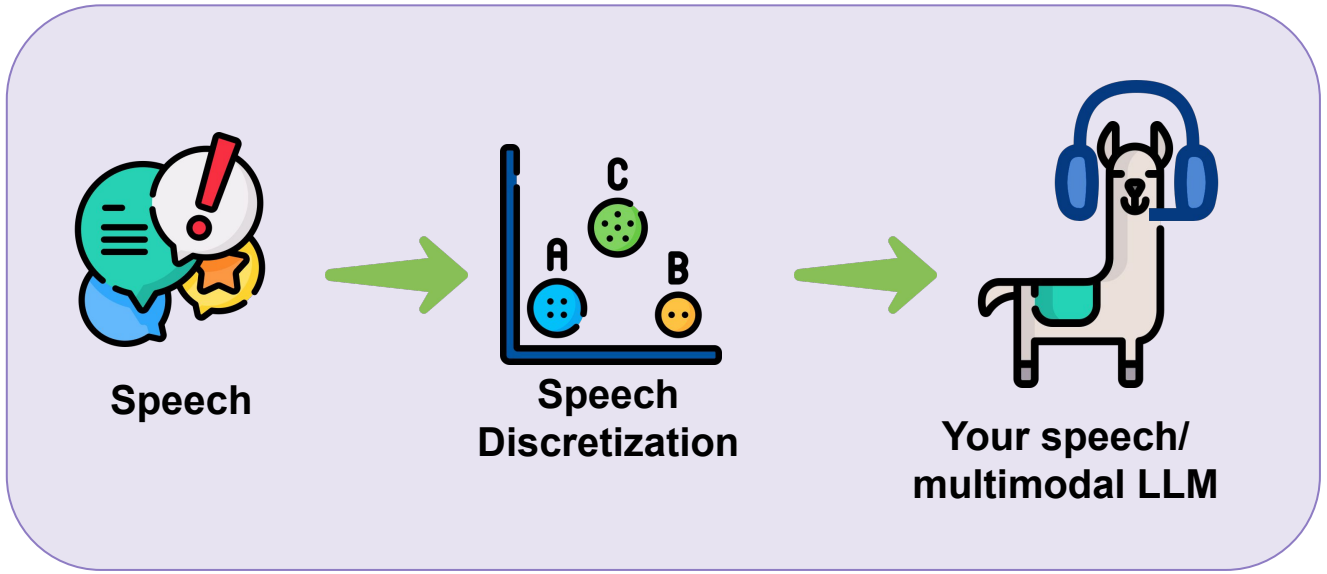
- ➔ LLM maintains its text capabilities
- ➔ Does not require training

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



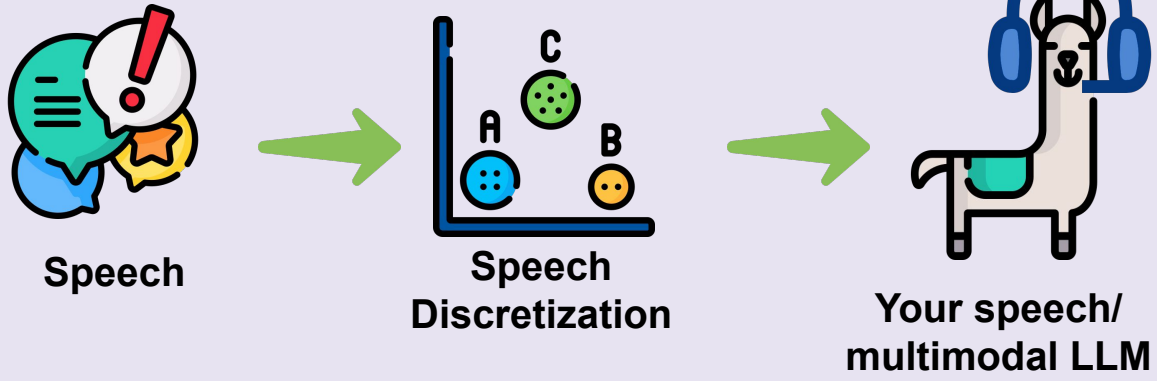
PROS

- ➔ 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

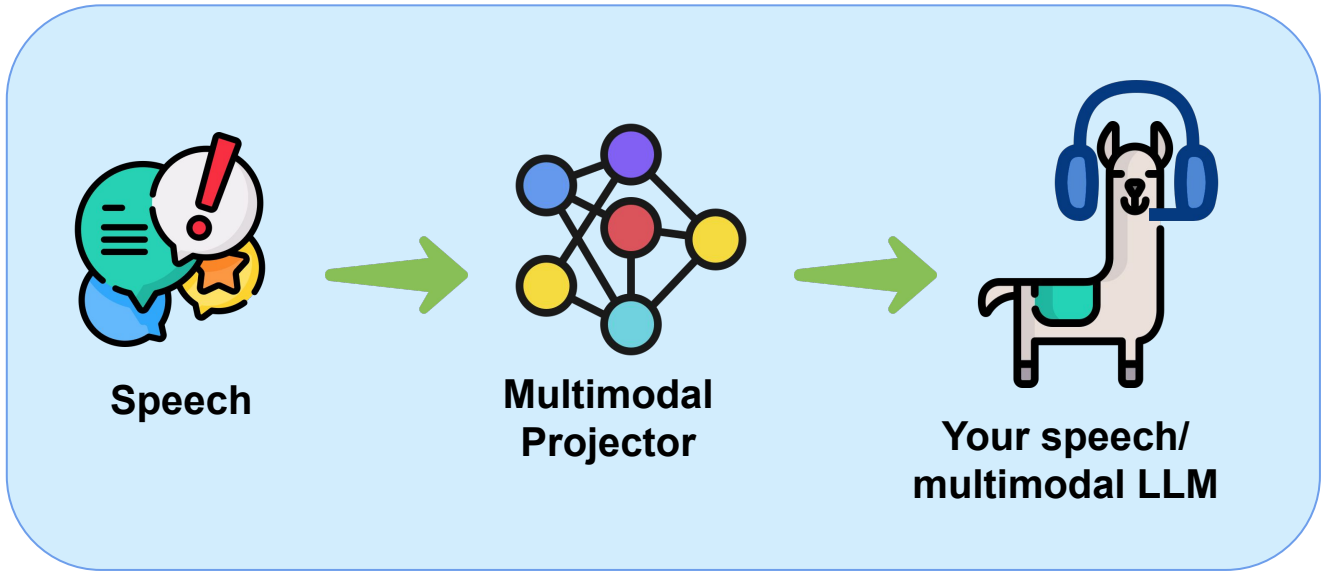
CONS

- ➔ 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?

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



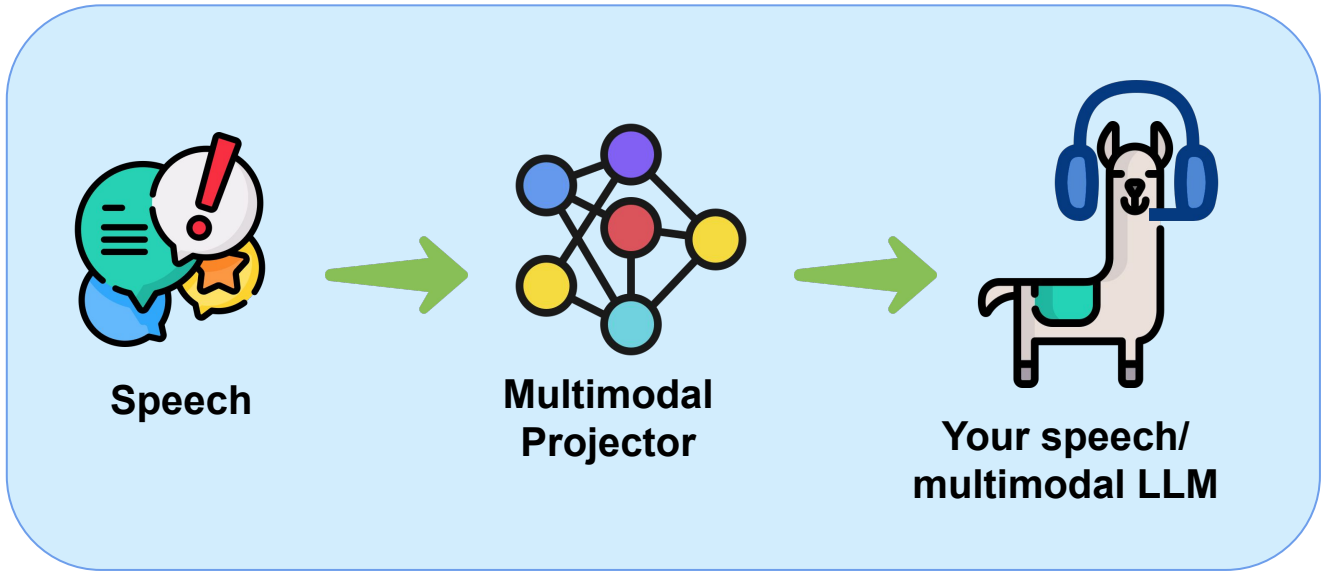
PROS

- ➔ 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

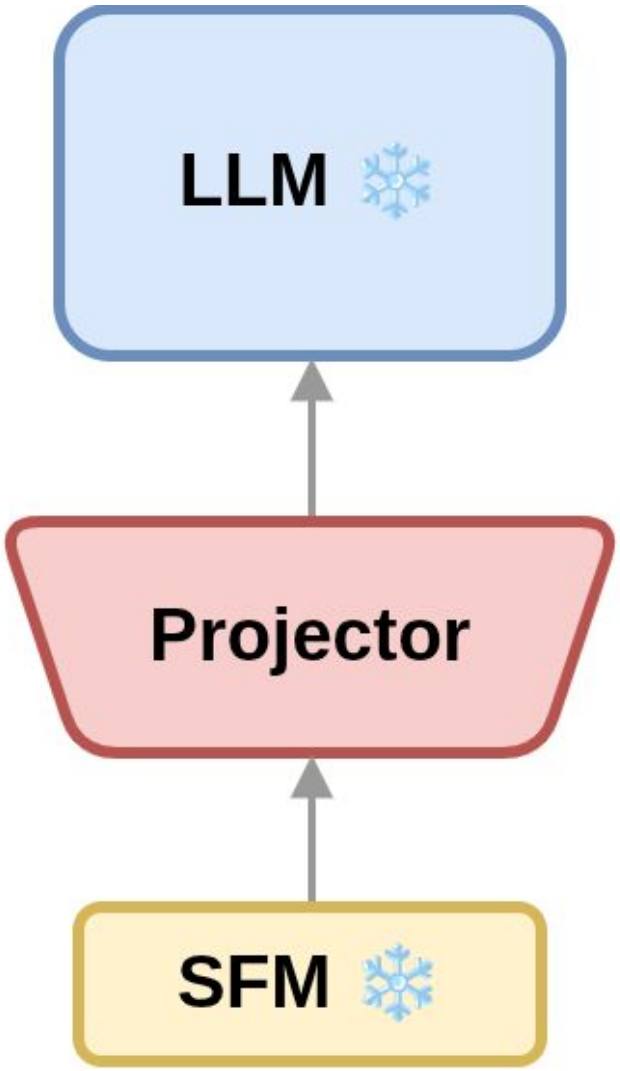
- ➔ No error propagation
- ➔ Acoustics ***potentially*** maintained
- ➔ Cheaper inference than cascading, potentially cheaper than discretizing

CONS

- ➔ Costly training for speech-to-text, even more costly for text-to-speech
  - ◆ in terms of **hardware** and **data**

Examples: WavLLM, SALMONN, Wav2Prompt

End-to-End Speech LLMs

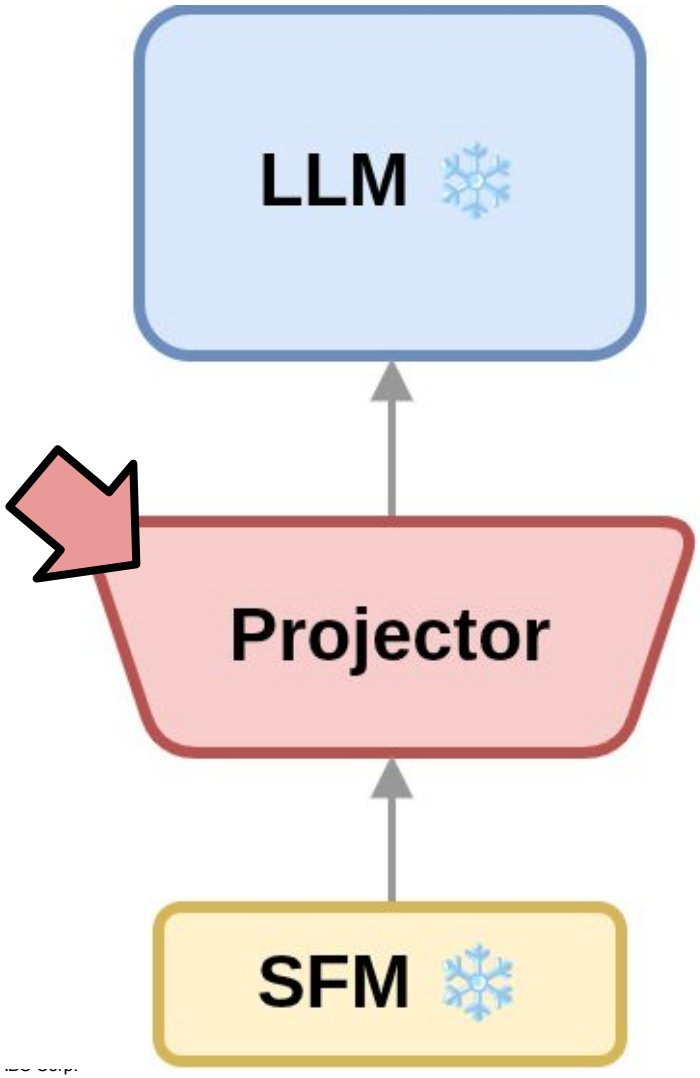


Bottlenecks:

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

End-to-End Speech LLMs

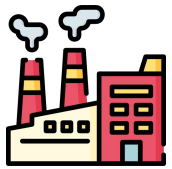


Bottlenecks:

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. **Projector might encode prompt and task as well, hurting generalization**

End-to-End Speech LLMs

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



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

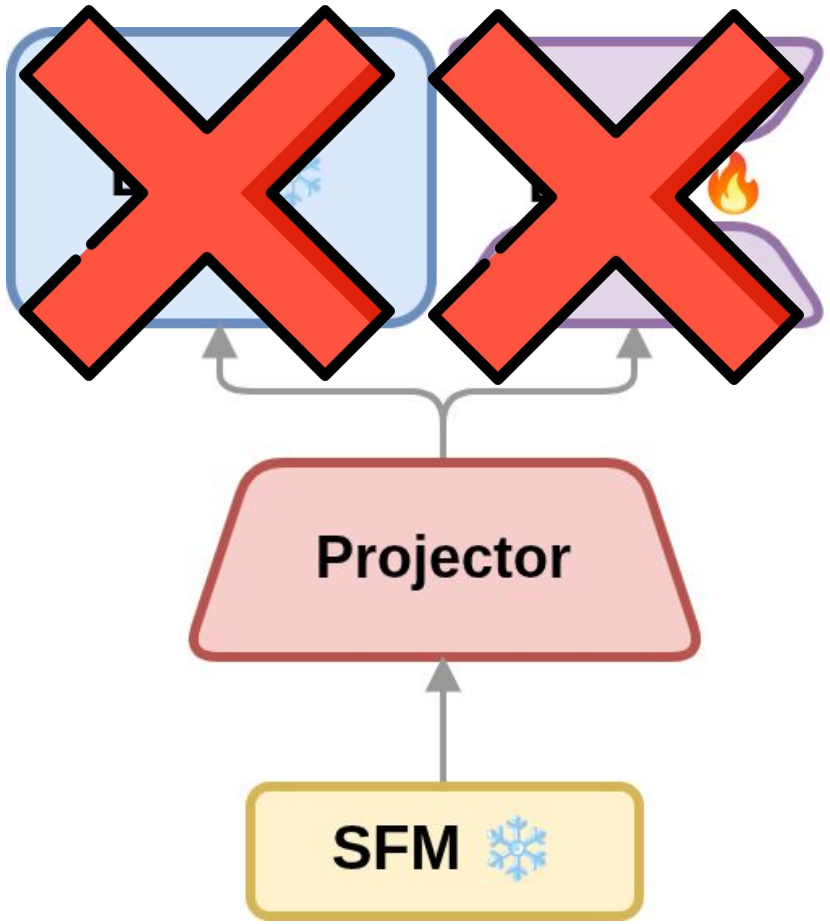
How to design a solution for easily switching between speech and text input?

SpeechMapper: Speech-to-text Embedding Projector for LLMs

ICASSP 2026

With
Biswesh Mohapatra
& Ioan Calapodescu

Then we asked ourselves…



Can we try to learn this projector without the LLM's forward pass?

Can we produce a projector output that *looks* like textual embeddings?

SpeechMapper: Reducing the dependency on the LLM forward pass

★ Propose a two stage approach

STAGE 1: Pre-Training (Speech-to-Embedding)

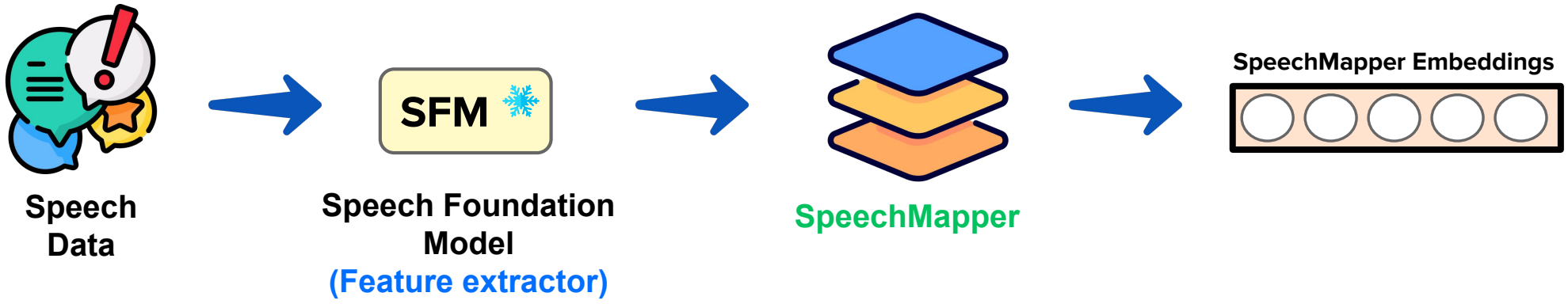
STAGE 2: Instruction Tuning (IT) Adaptation (Speech-to-Text)

SpeechMapper: Reducing the dependency on the LLM forward pass

★ **STAGE 1:** Speech-to-Embedding approach

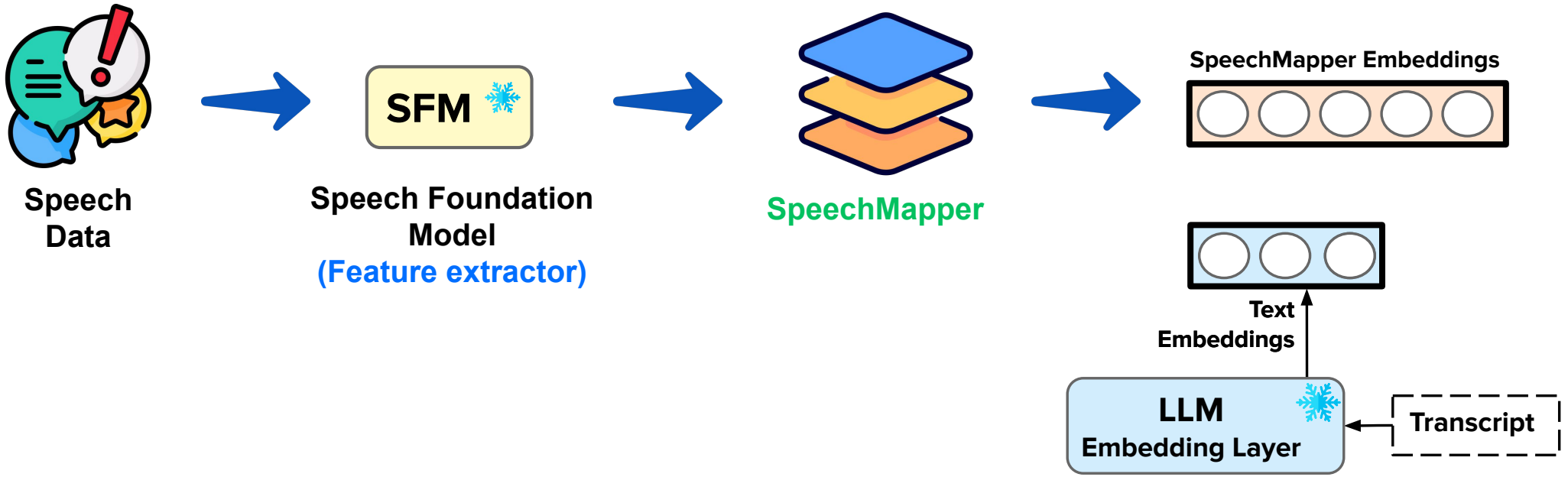
SpeechMapper: Reducing the dependency on the LLM forward pass

★ **STAGE 1:** Speech-to-Embedding approach



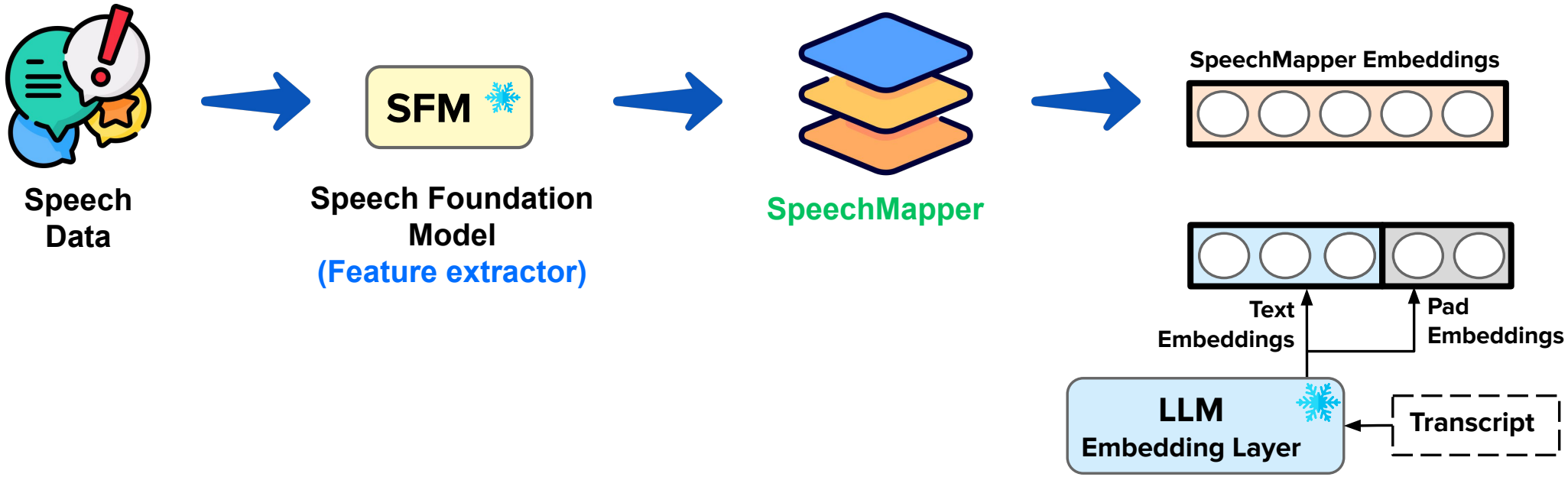
SpeechMapper: Reducing the dependency on the LLM forward pass

★ **STAGE 1:** Speech-to-Embedding approach



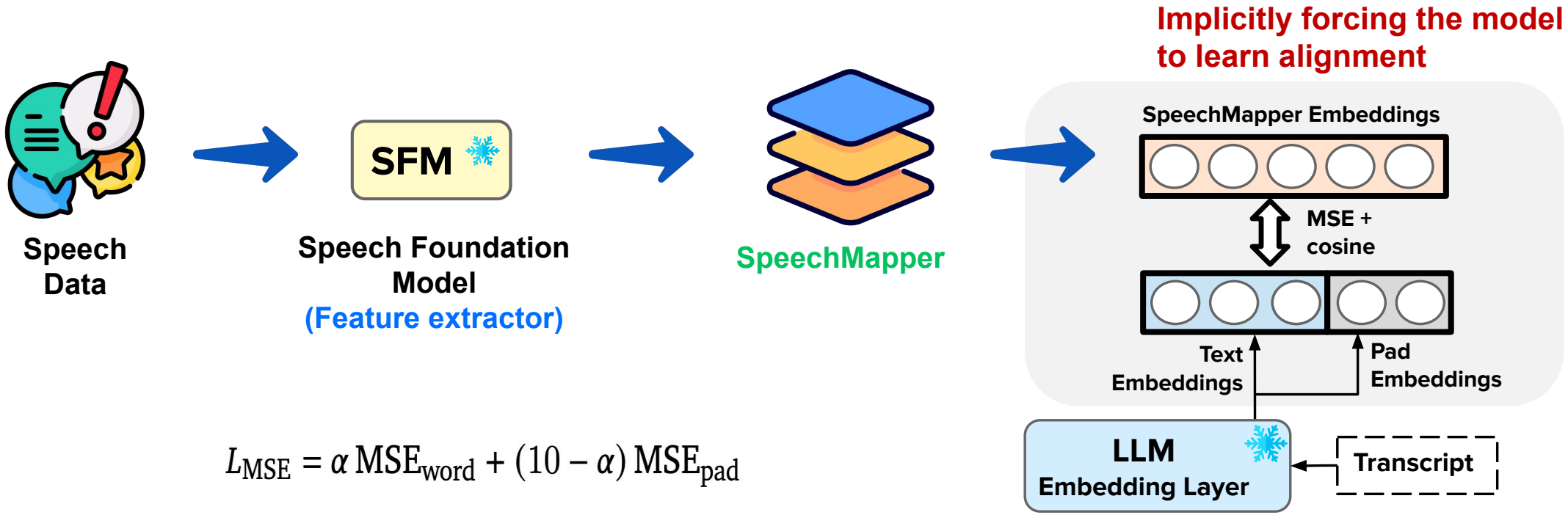
SpeechMapper: Reducing the dependency on the LLM forward pass

★ **STAGE 1:** Speech-to-Embedding approach



SpeechMapper: Reducing the dependency on the LLM forward pass

★ STAGE 1: Speech-to-Embedding approach

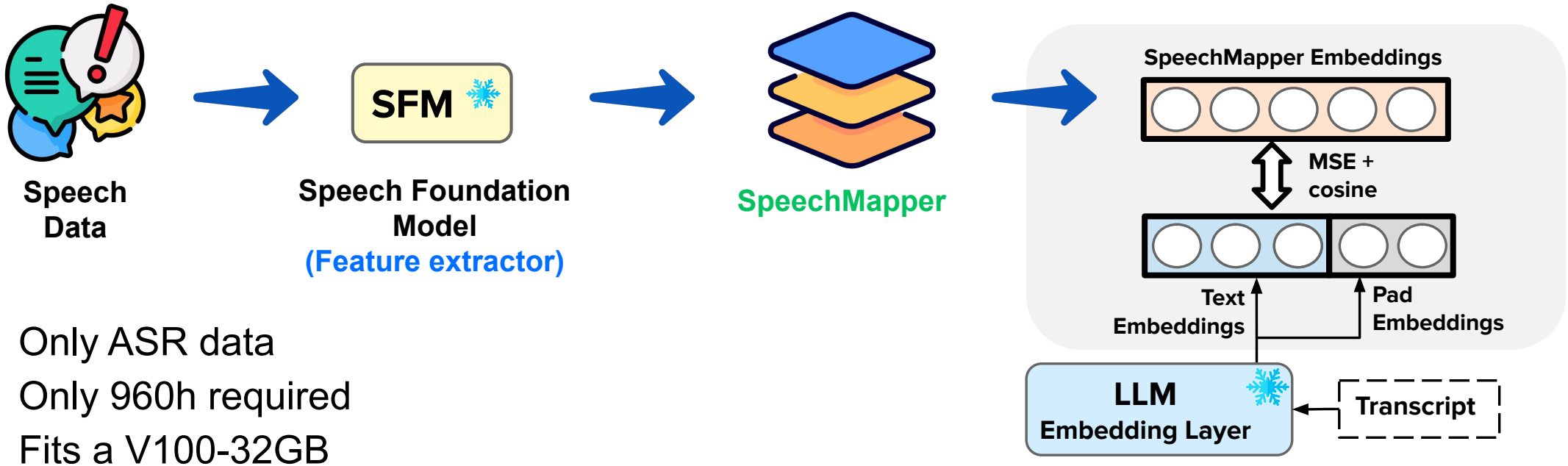


$$L_{\text{MSE}} = \alpha\,\text{MSE}_{\text{word}} + (10-\alpha)\,\text{MSE}_{\text{pad}}$$

$$L_{\text{pretraining}} = L_{\text{MSE}} - \gamma\,L_{\text{cosine}}$$

SpeechMapper: Reducing the dependency on the LLM forward pass

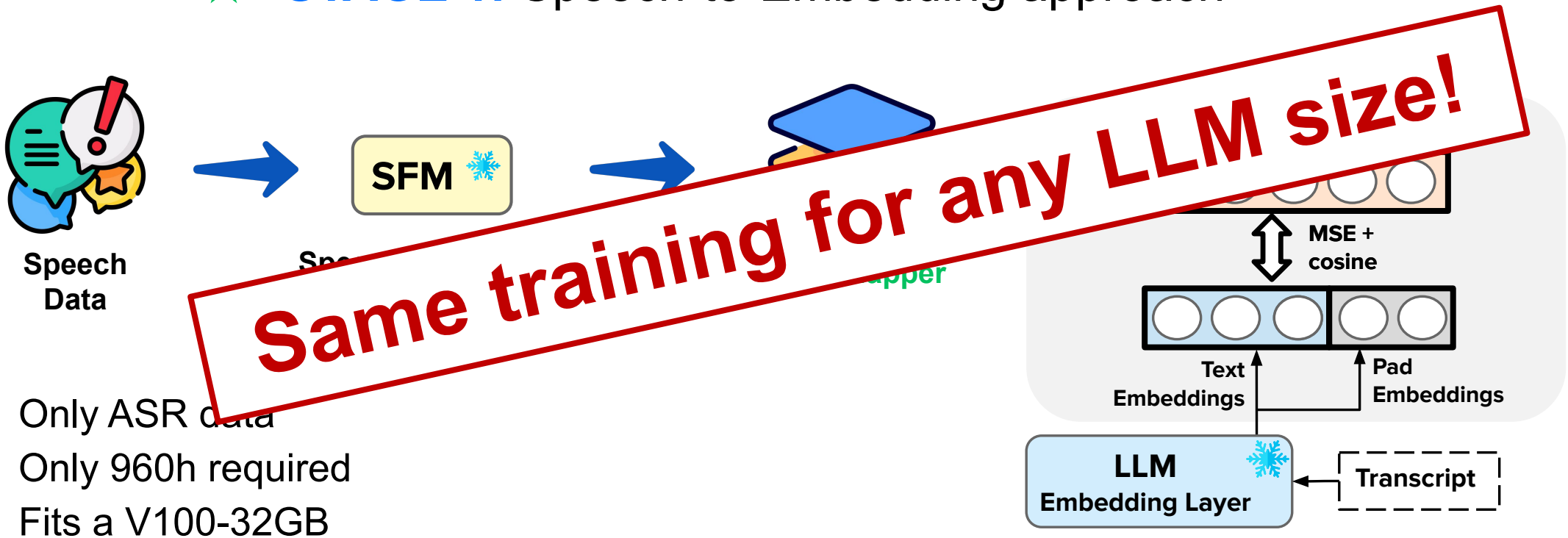
★ **STAGE 1:** Speech-to-Embedding approach



- ➔ Only ASR data
- ➔ Only 960h required
- ➔ Fits a V100-32GB
- ➔ Trains in 4 days on 4 GPUs

SpeechMapper: Reducing the dependency on the LLM forward pass

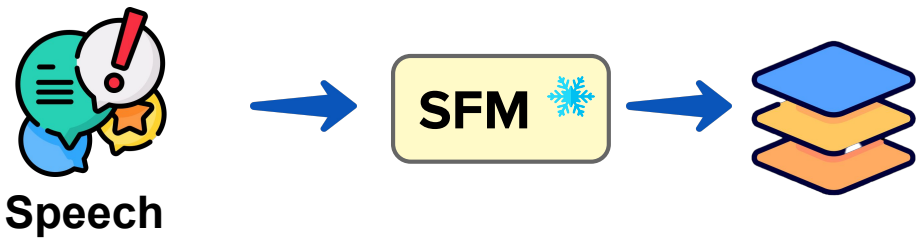
★ **STAGE 1:** Speech-to-Embedding approach



Same training for any LLM size!

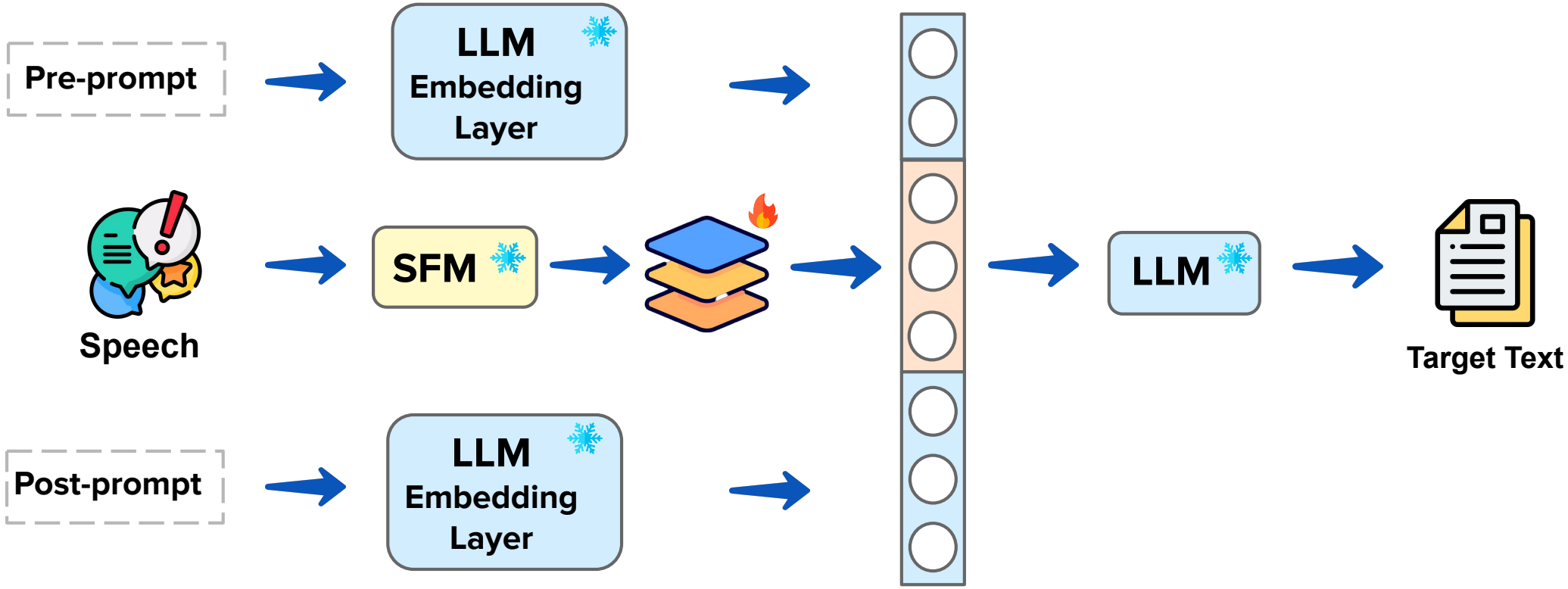
- → Only ASR data
- → Only 960h required
- → Fits a V100-32GB
- → Trains in 4 days on 4 GPUs

★ STAGE 2: Instruction Tuning

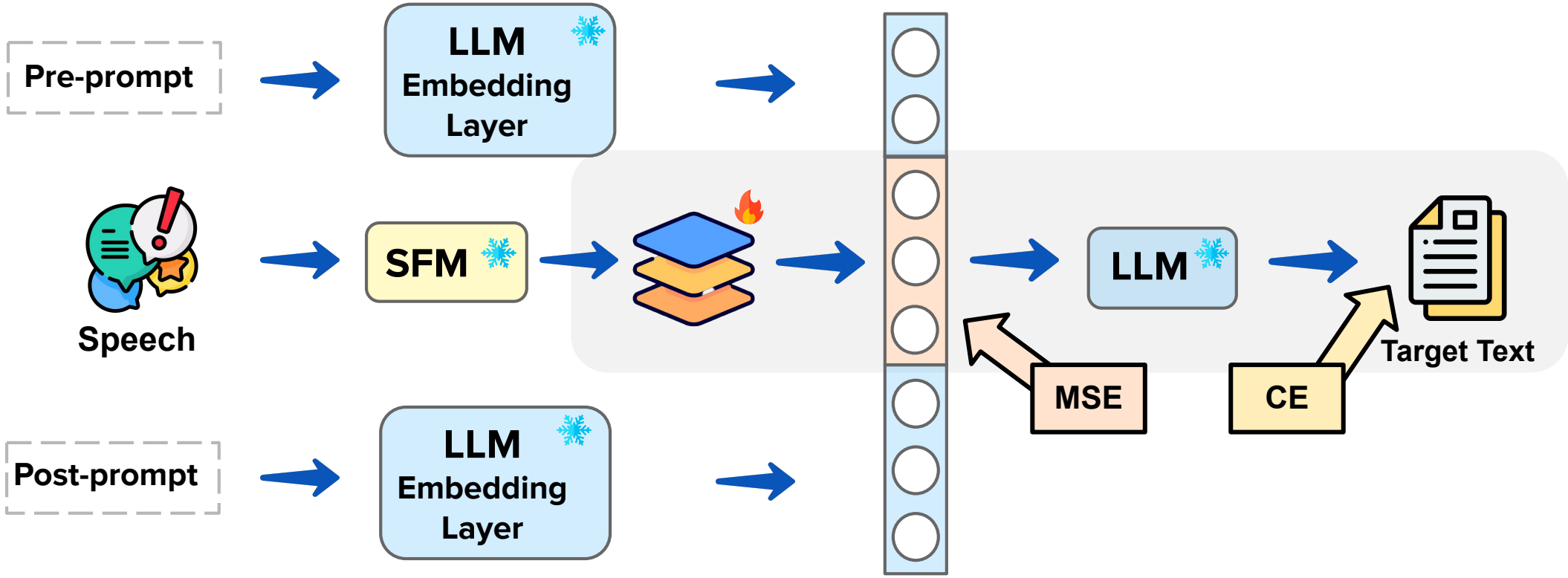


★ **Embeddings at this point can already be used, but they are not as sharp as we would like**

★ STAGE 2: IT adaptation

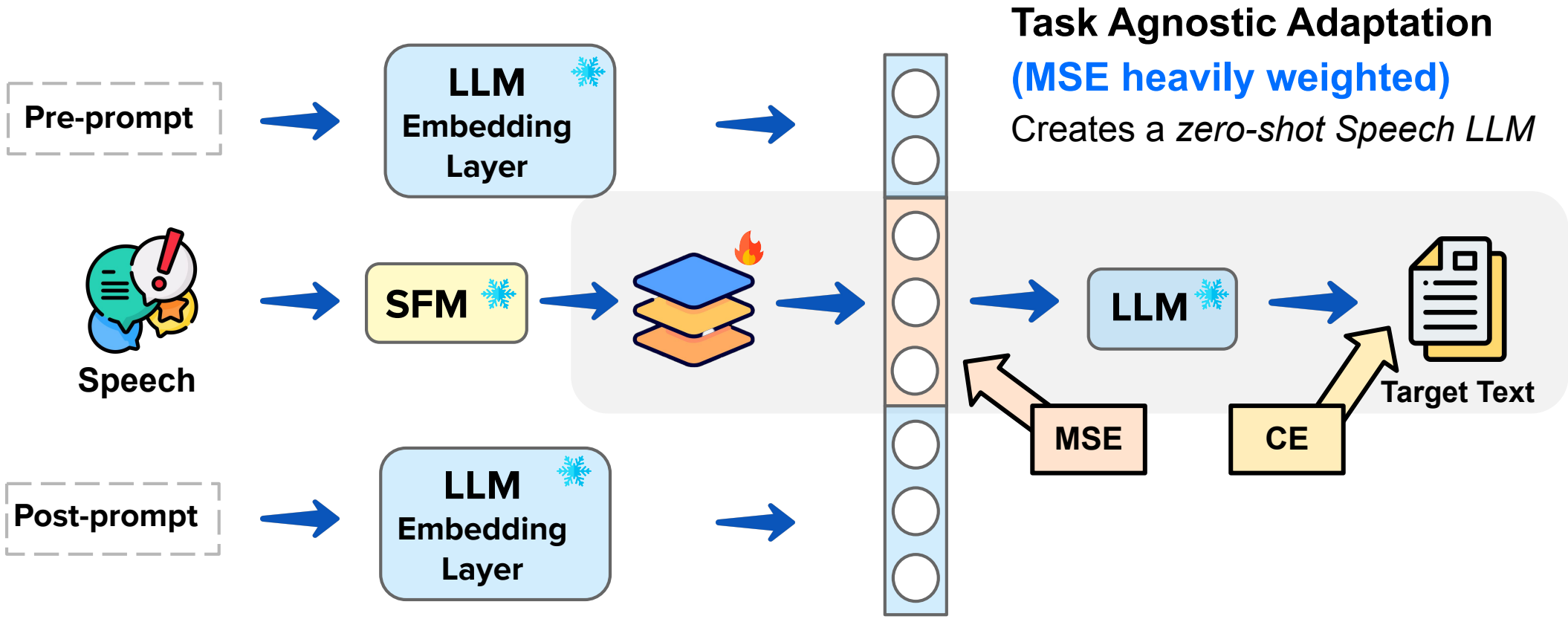


★ STAGE 2: IT adaptation



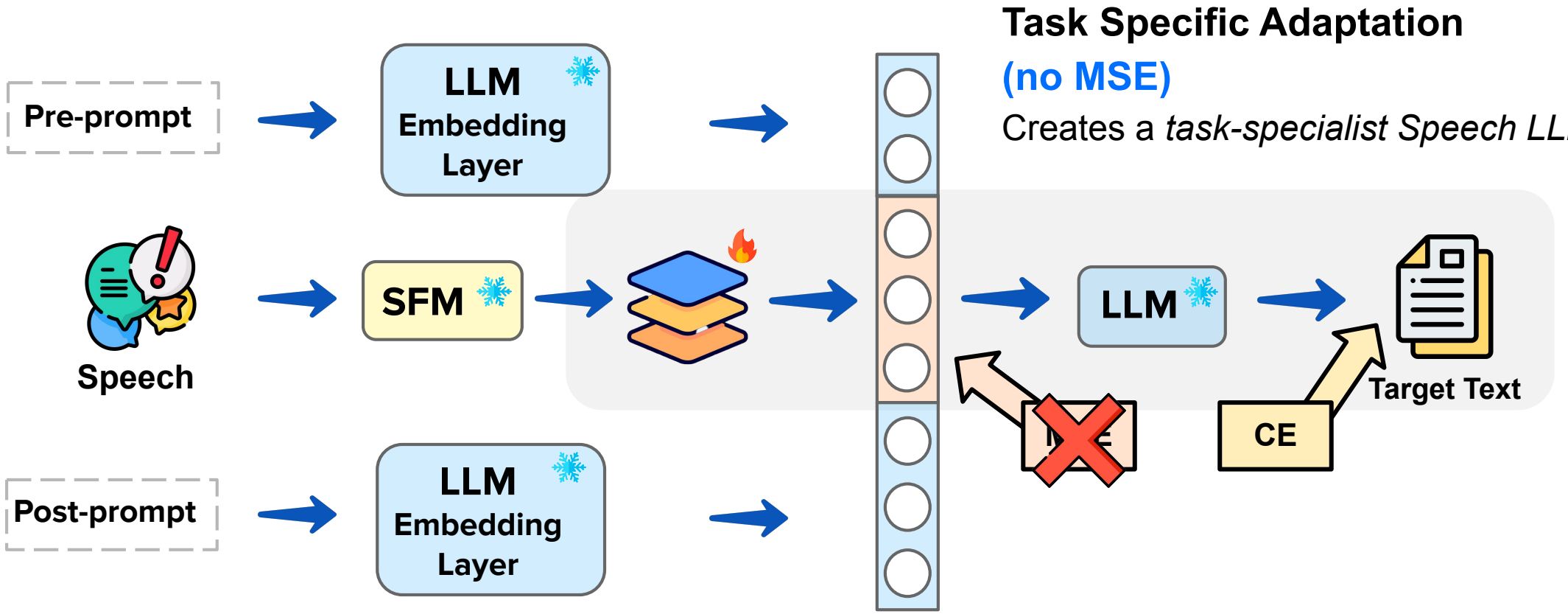
$$L_{IT} = (1 - \sigma) L_{CE} + \sigma L_{MSE}$$

★ STAGE 2: IT adaptation



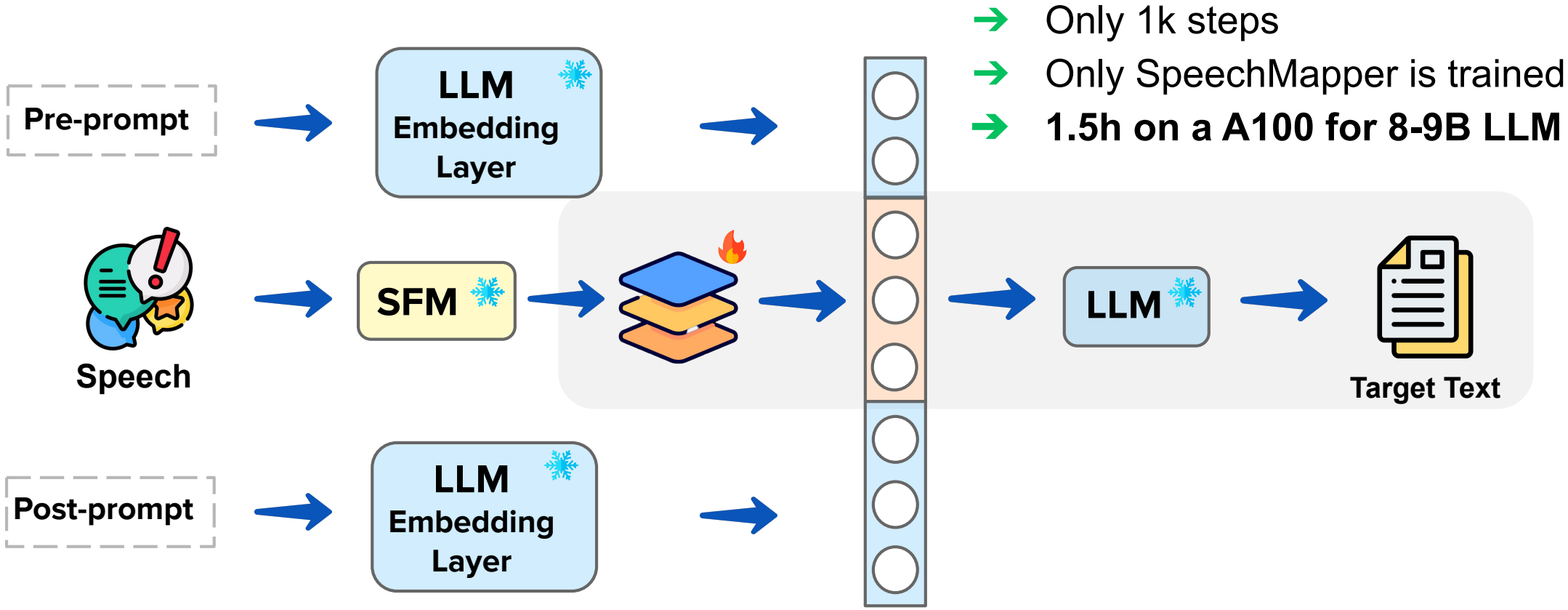
$$L_{IT} = (1 - \sigma) L_{CE} + \sigma L_{MSE}$$

★ STAGE 2: IT adaptation



$$L_{IT} = (1 - \sigma) L_{CE} + \sigma L_{MSE}$$

★ STAGE 2: IT adaptation



- ➔ Only 1k steps
- ➔ Only SpeechMapper is trained
- ➔ **1.5h on a A100 for 8-9B LLM**

Settings

Data

- Pretrained on 960 h LibriSpeech (with MMS normalization)
- Task Agnostic IT: Sampling from LibriSpeech
- Task-specific ST/SQA IT: Sampling from CoVoST2 and EuroParlST2 (ST) or SpokenSQuAD (SQA)

Backbones

- **SFM:** SeamlessM4T-large-v2
- **LLM:**
  - EuroLLM-9B-Instruct
  - Llama-3.1-8B-Instruct

Baseline

- **Best-IWSLT-IF:** Best IWSLT'25 instruction-following short system, trained in ASR, ST and SQA using the same datasets and backbones than us.

Speech Translation

Results: COMET metric (higher is better)

★ **Stage 1 achieves strong performance even surpassing Seamless ST on Europarl en-de, en-it**

		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
EuroLLM	Stage 1 (zero-shot)	73.5	76.0	74.1	75.8	64.2	64.8
Llama 3.1	Stage 1 (zero-shot)	76.4	73.9	72.3	76.8	67.1	69.3

Speech Translation

Results: COMET metric (higher is better)

- ★ **Stage 1 achieves strong performance even surpassing Seamless ST on Europarl en-de, en-it**
- ★ **CE+MSE yields reliable zero-shot behavior for Stage 2**

		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
EuroLLM	Stage 1 (zero-shot)	73.5	76.0	74.1	75.8	64.2	64.8
EuroLLM	Stage 2 [ASR CE] (zero-shot)	78.6±0.5	76.4±0.7	72.1±2.0	75.7±0.6	70.2±0.9	74.0±0.04
EuroLLM	Stage 2 [ASR CE+MSE] (zero-shot)	79.9±1.1	77.4±0.8	74.3±2.1	78.4±1.8	71.3±0.7	72.0±0.1
Llama 3.1	Stage 1 (zero-shot)	76.4	73.9	72.3	76.8	67.1	69.3
Llama 3.1	Stage 2 [ASR CE] (zero-shot)	70.4±2.9	69.4±2.2	60.5±7.5	63.9±6.8	63.4±5.4	62.2±8.2
Llama 3.1	Stage 2 [ASR CE+MSE] (zero-shot)	74.7±2.7	71.0±2.8	66.4±2.6	73.2±2.6	63.7±1.0	68.6±1.5

Speech Translation

Results: COMET metric (higher is better)

- ★ Stage 1 achieves strong performance even surpassing Seamless ST on Europarl en-de, en-it
- ★ CE+MSE yields reliable zero-shot behavior for Stage 2
- ★ In-domain ST boosts performance, narrowing the gap with BEST-IWSLT25-IF despite training for only 1K steps using ST data

		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
EuroLLM	Stage 1 (zero-shot)	73.5	76.0	74.1	75.8	64.2	64.8
EuroLLM	Stage 2 [ASR CE] (zero-shot)	78.6±0.5	76.4±0.7	72.1±2.0	75.7±0.6	70.2±0.9	74.0±0.04
EuroLLM	Stage 2 [ASR CE+MSE] (zero-shot)	79.9±1.1	77.4±0.8	74.3±2.1	78.4±1.8	71.3±0.7	72.0±0.1
EuroLLM	Stage 2 [ST CE] (in-domain)	**85.4**±0.4	**84.5**±0.5	**82.2**±0.3	**85.5**±0.6	**77.0**±0.1	**79.9**±0.02
Llama 3.1	Stage 1 (zero-shot)	76.4	73.9	72.3	76.8	67.1	69.3
Llama 3.1	Stage 2 [ASR CE] (zero-shot)	70.4±2.9	69.4±2.2	60.5±7.5	63.9±6.8	63.4±5.4	62.2±8.2
Llama 3.1	Stage 2 [ASR CE+MSE] (zero-shot)	74.7±2.7	71.0±2.8	66.4±2.6	73.2±2.6	63.7±1.0	68.6±1.5
Llama 3.1	Stage 2 [ST CE] (in-domain)	**84.5**±0.2	**82.4**±0.1	**80.9**±0.2	**84.5**±0.1	**75.5**±0.1	**78.6**±0.1

Spoken Question Answering

Results: Accuracy metric (higher is better)

- LLM-as-judge setup with average across 4 LLMs

★ **Best zero-shot model rivals BEST-IWSLT25-IF on LibriSQA**

		Spoken SQuAD	LibriSQA PartI	LibriSQA PartII
	Transcripts + EuroLLM 9B (topline)	91.1%±2.5	87.6%±5.1	73.4%±3.1
	Transcripts + Llama 3.1 8B (topline)	89.2%±2.4	85.1%±4.5	74.9%±3.5
	Seamless ASR + EuroLLM 9B (zero-shot)	89.2%±2.9	79.8%±6.5	73.5%±3.9
	Seamless ASR + Llama 3.1 8B (zero-shot)	85.6%±3.4	82.3%±5.7	74.7%±4.9
	BEST-IWSLT25-IF (in-domain)	87.4%±3.2	80.7%±6.7	62.5%±4.0
EuroLLM	Stage 1 (zero-shot)	61.9%±7.4	51.9%±15.6	60.3%±6.5
Llama 3.1	Stage 1 (zero-shot)	62.3%±5.1	70.7%±7.1	70.5%±3.7

Spoken Question Answering

Results: Accuracy metric (higher is better)

- LLM-as-judge setup with average across 4 LLMs

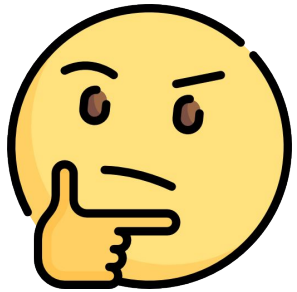
★ **Best zero-shot model rivals BEST-IWSLT25-IF on LibriSQA**

★ **In-domain IT reaches the same performance level as the strong pipeline based toplines**

		Spoken SQuAD	LibriSQA PartI	LibriSQA PartII
	Transcripts + EuroLLM 9B (topline)	91.1%±2.5	87.6%±5.1	73.4%±3.1
	Transcripts + Llama 3.1 8B (topline)	89.2%±2.4	85.1%±4.5	74.9%±3.5
	Seamless ASR + EuroLLM 9B (zero-shot)	89.2%±2.9	79.8%±6.5	73.5%±3.9
	Seamless ASR + Llama 3.1 8B (zero-shot)	85.6%±3.4	82.3%±5.7	74.7%±4.9
	BEST-IWSLT25-IF (in-domain)	87.4%±3.2	80.7%±6.7	62.5%±4.0
EuroLLM	Stage 1 (zero-shot)	61.9%±7.4	51.9%±15.6	60.3%±6.5
EuroLLM	Stage 2 [ASR CE+MSE] (zero-shot)	75.1%±9.5	79.3%±6.3	64.3%±4.8
EuroLLM	Stage 2 [ASR/SQA CE] (in-domain)	**87.4%**±3.2	**83.2%**±5.1	**68.1%**±2.3
Llama 3.1	Stage 1 (zero-shot)	62.3%±5.1	70.7%±7.1	70.5%±3.7
Llama 3.1	Stage 2 [ASR CE+MSE] (zero-shot)	72.3%±7.6	75.6%±7.1	68.9%±2.5
Llama 3.1	Stage 2 [ASR/SQA CE] (in-domain)	**87.9%**±3.5	**81.6%**±6.0	**72.5%**±1.4

- ★ **We managed to outperform our system from IWSLT'25, using less computing and data**
- ★ **Pre-trained block that can be very quickly adapted to excel in a given task**
- ★ ***Most* of the training is performed without the LLM**

How do we make training better?



$$L_{\text{MSE}} = \alpha\,\text{MSE}_{\text{word}} + (10 - \alpha)\,\text{MSE}_{\text{pad}}$$

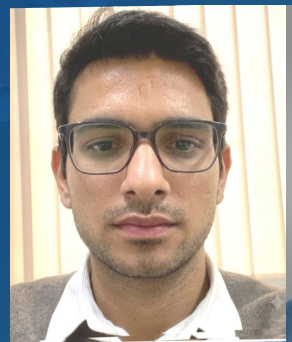
$$L_{\text{pretraining}} = L_{\text{MSE}} - \gamma\,L_{\text{cosine}}$$

- **MSE loss is very unstable**
- **Very long warm-up required**
- **Training in f32 because of unstable gradients**
- **Resulting model is *“almost there”***

SpeechMapper v2: Efficient training of speech LLMs without instruction-following data

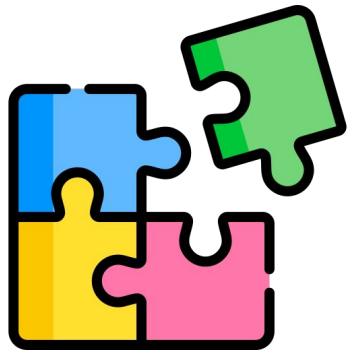
Mostly* unpublished

With
Hemant Yadav &
Jean-Luc Meunier &
Ioan Calapodescu



*Will appear at IWSLT'26

SpeechMapper v2: Single-stage Speech-to-Embedding Training



A lot of work in training dynamics optimization:

- Training optimizations:
  - fp16
  - Dynamic Batching
  - Torchtune implementation

SpeechMapper v2: Single-stage Speech-to-Embedding Training

$$Z_t = [z_t^{(1)}, \ldots, z_t^{(T')}, z_{\text{pad}}, \ldots, z_{\text{pad}}] \in \mathbb{R}^{T \times d}$$

1

$$\mathcal{L}_{\text{L1}} = \frac{1}{Td} \sum_{t=1}^{T} \left\| z_s^{(t)} - z_t^{(t)} \right\|$$

L1 replaces MSE

2

$$\mathcal{L}_{\cos} = \frac{1}{T} \sum_{t=1}^{T} \left(1 - \frac{z_s^{(t)} \cdot z_t^{(t)}}{\|z_s^{(t)}\|_2 \|z_t^{(t)}\|_2} \right)$$

We keep the cosine from before

3

$$\mathcal{L}_{\text{softmax}} = -\frac{1}{T} \sum_{t=1}^{T} y_t^{\top} \log(\text{softmax}(s_t))$$

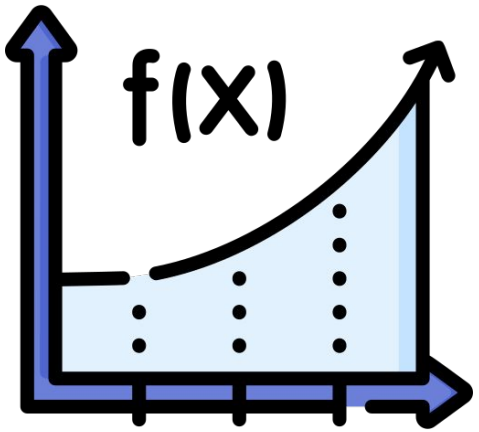
Contrastive loss is added

4

$$\mathcal{L} = \mathcal{L}_{\text{L1}} + \mathcal{L}_{\cos} + 0.1 * \mathcal{L}_{\text{softmax}} + \mathcal{L}_{\text{ctc}}$$

CTC loss is added to the mix

SpeechMapper v2: Single-stage Speech-to-Embedding Training



- ➔ **Training went from 4 days in 4xV100-32GB to 18h in the same hardware**
- ➔ **Training is now much more stable**

Settings

Baseline

- **Best-IWSLT-IF**
- **SpeechMapper v1** (stage 2 models)
- Diva (L1 loss + CE on continuation)
- **Qwen 2.5 Omni** (general purpose)

Backbones

- **SFM:** SeamlessM4T-large-v2
- **LLM:**
  - Llama-3-8B-Instruct
  - Llama-3.1-8B-Instruct
  - Qwen-2.5-7B-Instruct-2507

Speech Translation

COMET (higher the better)		EuroparlST			CoVoST2	
	en-de	en-es	en-fr	en-it	en-de	en-zh-CN
LLAMA 3.1 models						
Llama-3.1 (text topline)	82.2	86.4	83.9	85.2	83.2	85.7
BEST-IWSLT25-IF (task specialist)	**84.0**	83.5	81.1	**86.0**	78.9	80.7
SpeechMapper v1 stage 2 (zero-shot)	66.4	79.9	71.0	73.2	63.7	68.6
SpeechMapper v1 stage 2 (task specialist)	80.9	**84.5**	**82.4**	84.5	75.5	78.6
SpeechMapper v2 (zero-shot)	77.7	81.6	79.1	80.4	**80.1**	**82.2**

★ We beat SpeechMapper v1 stage 2 in zero-shot settings

★ SpeechMapper v2 only loses for instruction-tuning using in-domain data!

Speech Translation

COMET (higher the better)		EuroparlST			CoVoST2	
	en-de	en-es	en-fr	en-it	en-de	en-zh-CN
LLAMA 3.1 models						
Llama-3.1 (text topline)	82.2	86.4	83.9	85.2	83.2	85.7
BEST-IWSLT25-IF (task specialist)	**84.0**	83.5	81.1	**86.0**	78.9	80.7
SpeechMapper v1 stage 2 (zero-shot)	66.4	79.9	71.0	73.2	63.7	68.6
SpeechMapper v1 stage 2 (task specialist)	80.9	**84.5**	**82.4**	84.5	75.5	78.6
SpeechMapper v2 (zero-shot)	77.7	81.6	79.1	80.4	**80.1**	**82.2**
LLAMA 3 models						
Llama-3 (text topline)	81.1	85.9	83.1	84.3	82.2	82.9
Diva (zero-shot)	75.2	**80.4**	78.1	77.7	77.1	**77.7**
SpeechMapper v2 (zero-shot)	**76.5**	80.3	**78.5**	**79.6**	**78.9**	72.2

Speech Translation

COMET (higher the better)		EuroparlST			CoVoST2	
	en-de	en-es	en-fr	en-it	en-de	en-zh-CN
LLAMA 3.1 models						
Llama-3.1 (text topline)	82.2	86.4	83.9	85.2	83.2	85.7
BEST-IWSLT25-IF (task specialist)	**84.0**	83.5	81.1	**86.0**	78.9	80.7
SpeechMapper v1 stage 2 (zero-shot)	66.4	79.9	71.0	73.2	63.7	68.6
SpeechMapper v1 stage 2 (task specialist)	80.9	**84.5**	**82.4**	84.5	75.5	78.6
SpeechMapper v2 (zero-shot)	77.7	81.6	79.1	80.4	**80.1**	**82.2**
LLAMA 3 models						
Llama-3 (text topline)	81.1	85.9	83.1	84.3	82.2	82.9
Diva (zero-shot)	75.2	**80.4**	78.1	77.7	77.1	**77.7**
SpeechMapper v2 (zero-shot)	**76.5**	80.3	**78.5**	**79.6**	**78.9**	72.2
QWEN 2.5 models						
Qwen 2.5 (text topline)	80.0	84.7	82.1	83.0	80.8	86.1
Qwen 2.5 Omni (general purpose)	**77.9**	**82.9**	77.0	76.2	**78.4**	75.7
SpeechMapper v2 (zero-shot)	75.9	80.3	**78.0**	**78.7**	77.3	**81.8**

Spoken Question Answering

LLM-as-judge (higher the better)	SpokenSQuAD		LibriSQA part1		LibriSQA part2	
	acc	stdev	acc	stdev	acc	stdev
LLAMA 3.1 models						
Llama-3.1 (text topline)	89.3	4.2	86.6	6.8	70.8	4.7
BEST-IWSLT25-IF (task specialist)	87.5	3.9	82.0	7.5	63.0	4.7
SpeechMapper v1 stage 2 (zero-shot)	72.1	4.2	76.6	6.6	68.9	2.2
SpeechMapper v1 stage 2 (task specialist)	**89.4**	3.0	**82.5**	6.1	73.1	2.2
SpeechMapper v2 (zero-shot)	85.9	3.1	81.8	6.3	**75.0**	4.0
LLAMA 3 models						
Llama-3 (text topline)	89.9	4.2	87.0	5.9	71.3	3.7
Diva (zero-shot)	72.8	9.1	67.8	14.5	66.6	2.0
SpeechMapper v2 (zero-shot)	**83.5**	2.5	**82.7**	5.6	**74.0**	3.5
QWEN 2.5 models						
Qwen 2.5 (text topline)	88.7	2.0	87.9	4.4	75.0	1.6
Qwen 2.5 Omni (general purpose)	**83.7**	1.6	72.9	3.5	68.8	3.3
SpeechMapper v2 (zero-shot)	80.7	1.6	**84.3**	5.4	**74.8**	2.4

- ★ **We got rid of stage 2 for zero-shot training**
  - ○ **Even if we decide to perform this IT step, it will cost much less time and resources**
- ★ **No instruction-data, no LLM loaded in memory**
- ★ **Competitive with Qwen-2.5-Omni and better than Diva**
- ★ **We now can train a projector for any size of an LLM in less than one day**

Concluding

SpeechMapper

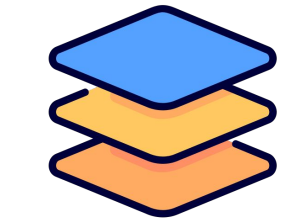
- ★ Training a speech projector for semantic tasks without loading the LLM has many advantages (hardware, no catastrophic forgetting, etc)
- ★ The tradeoff is more parameters: SpeechMapper is larger than regular projectors one could use when training with LLM+CE
- ★ SpeechMapper success depends on the LLM capability to deal with embedding noise

SpeechMapper

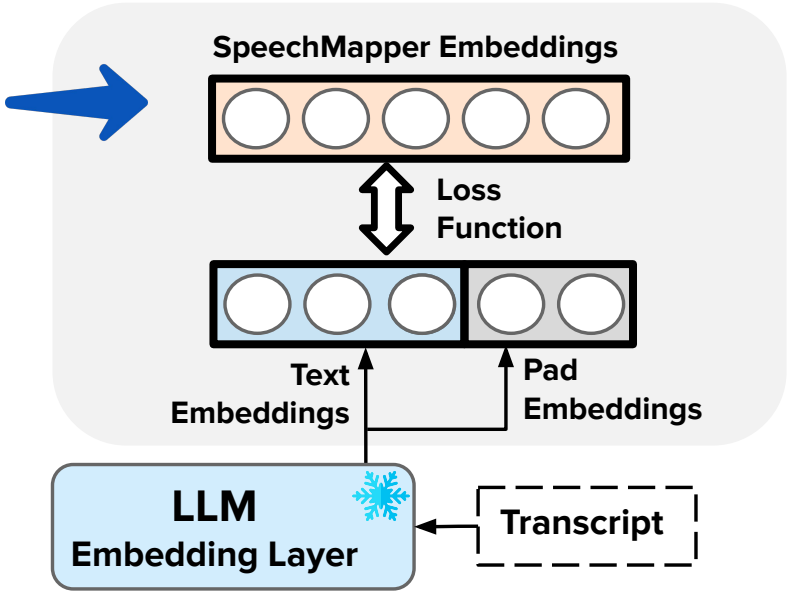
- ★ Training a speech projector for semantic tasks without loading the LLM has many advantages (hardware, no catastrophic forgetting, etc)
- ★ The tradeoff is more parameters: SpeechMapper is larger than regular projectors one could use when training with LLM+CE
- ★ SpeechMapper success depends on the LLM capability to deal with embedding noise
- ★ One could ask: **why not using a SOTA ASR instead?**

Paralinguistically-aware SpeechMapper

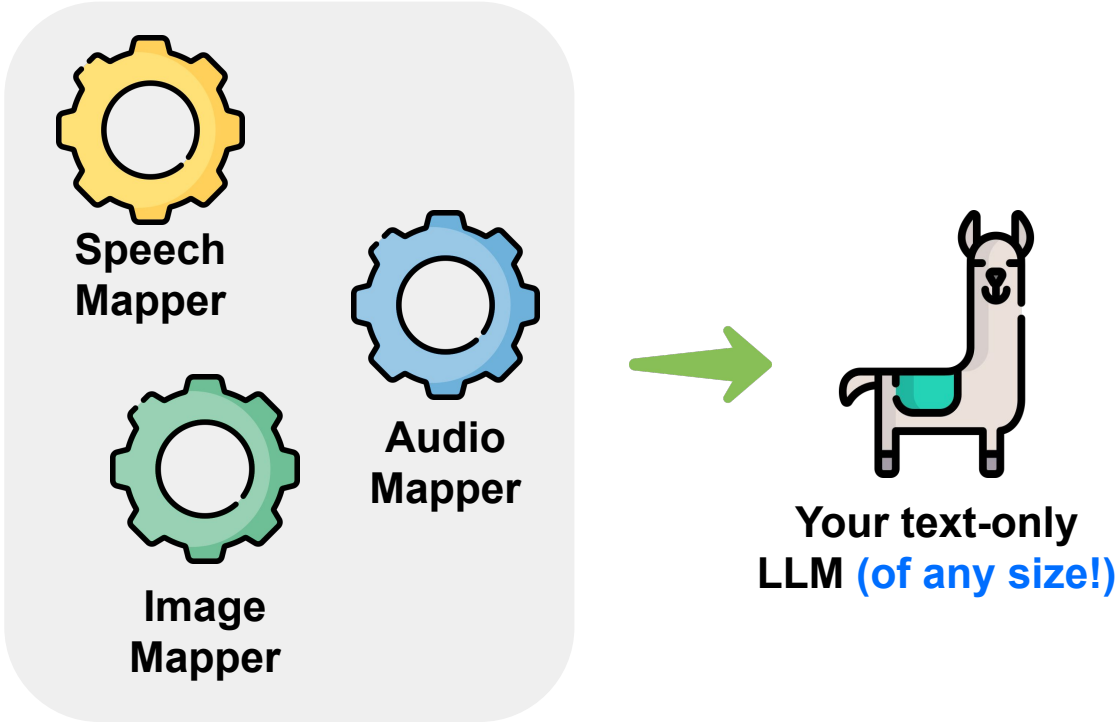
- ★ Our ambition for SpeechMapper was never to have simply a content projector
- ★ We would like to exploit the *"empty space"* in the sequences we are padding, overloading the pad tokens with acoustic information



SpeechMapper



Our ambition: Modular Multimodality



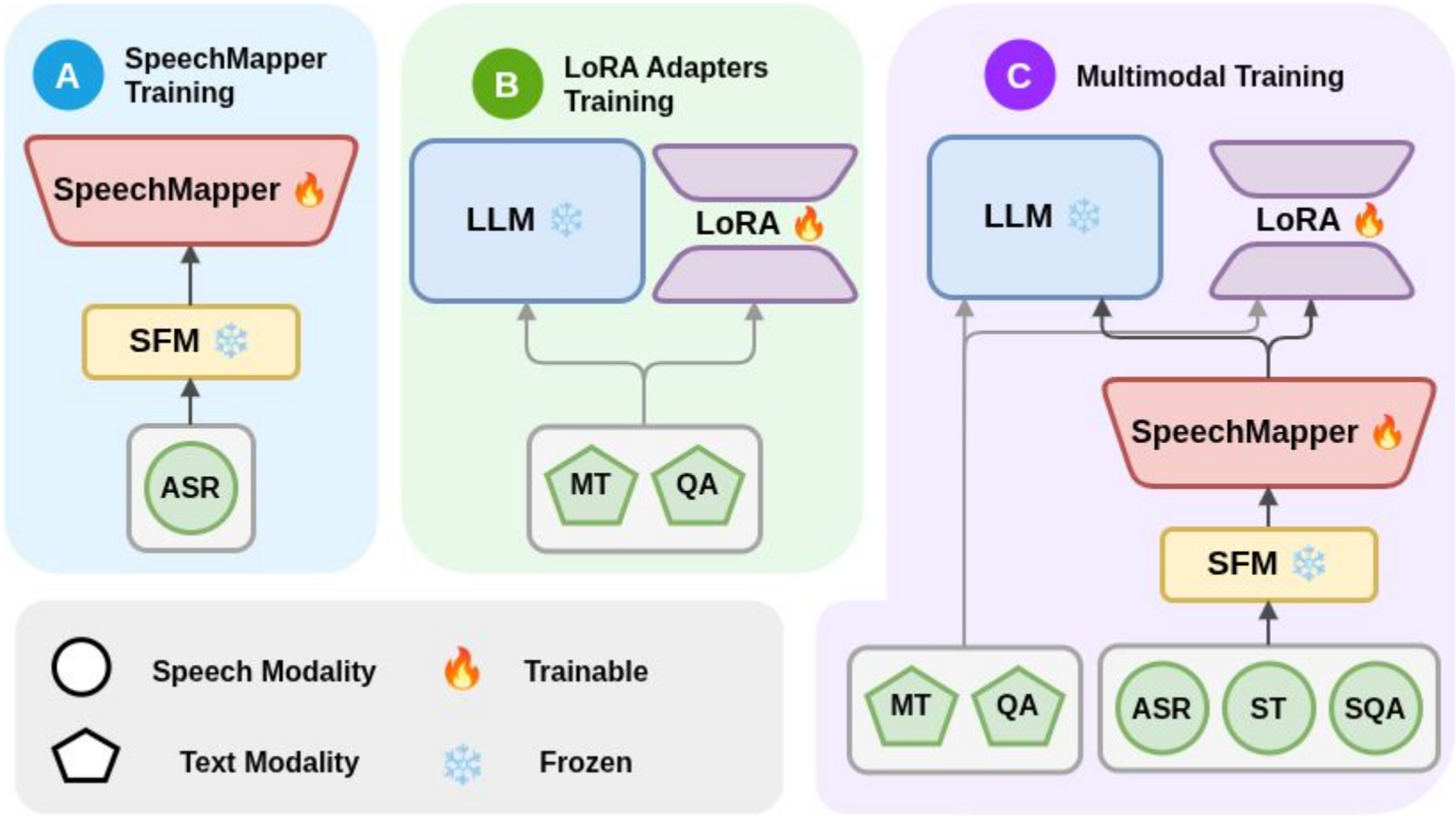
Thanks for listening!

12/2025

Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS

IWSLT'26



IWSLT'26

Models	Lang	TRANS-COMET	QA-BERTScore	QE-accuracy	QE-format-accuracy	ASR-WER
NLE_IWSLT26_IF_SHORT_CONSTRAINED_PRIMARY	en	—	0.531	—	—	0.136
NLE_IWSLT26_IF_SHORT_UNCONSTRAINED_CONSTRASTIVE	en	—	0.501	—	—	0.134
NLE_IWSLT26_IF_SHORT_CONSTRAINED_PRIMARY	it	0.763	0.456	—	—	—
NLE_IWSLT26_IF_SHORT_UNCONSTRAINED_CONSTRASTIVE	it	0.733	0.514	—	—	—
NLE_IWSLT26_IF_SHORT_CONSTRAINED_PRIMARY	de	0.765	0.470	0.786	0.997	—
NLE_IWSLT26_IF_SHORT_UNCONSTRAINED_CONSTRASTIVE	de	0.749	0.462	0.333	0.005	—
NLE_IWSLT26_IF_SHORT_CONSTRAINED_PRIMARY	zh	0.794	0.487	0.894	1.000	—
NLE_IWSLT26_IF_SHORT_UNCONSTRAINED_CONSTRASTIVE	zh	0.755	0.466	0.500	0.014	—

★ **We managed to outperform our solution from last year, despite working on a weaker backbone (half of the parameters)**