Prompting Speech Language Model for Speech Processing

Speaker: Kai-Wei Chang

kaiwei.chang.tw@gmail.com

- Name: Kai-Wei Chang
- Advisor: Prof. Hung-yi Lee
- National Taiwan University (NTU)

國立臺灣大學
National Taiwan University

Outline

- Pre-train, fine-tune paradigm vs.Prompting paradigm

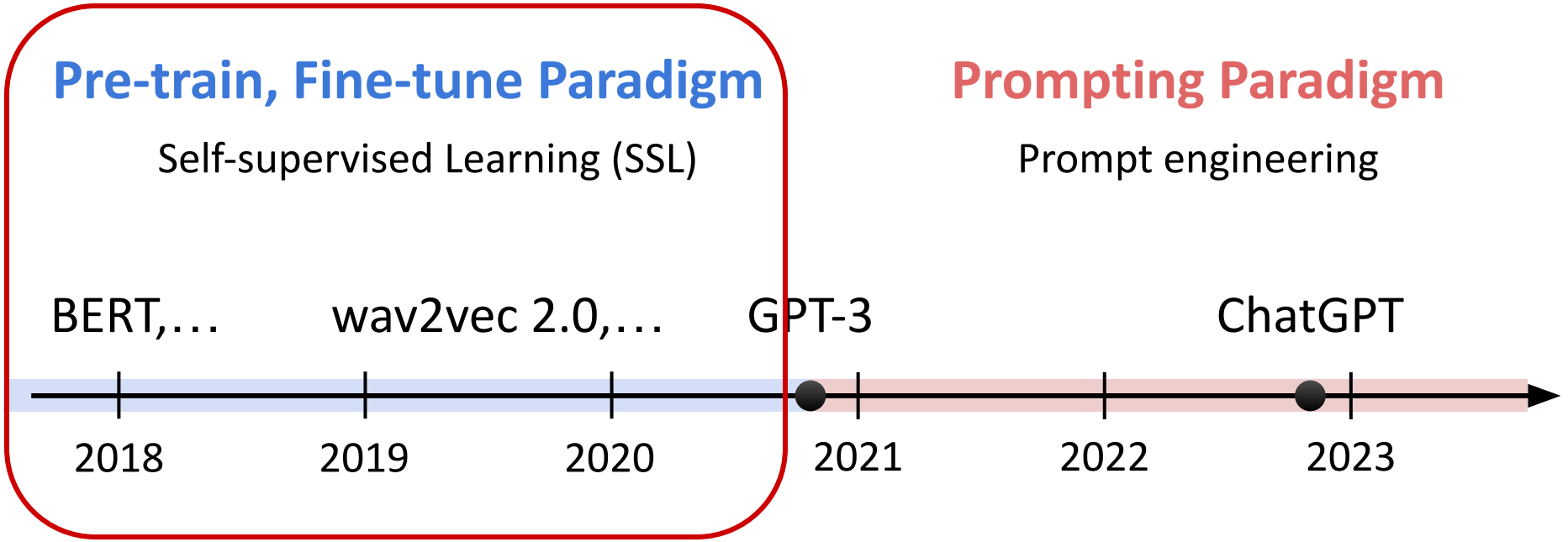
Outline

- Pre-train, fine-tune paradigm vs.Prompting paradigm
- SpeechPrompt: Prompting for speech processing
  - Textless speech LM
  - SpeechPrompt for various tasks
  - In-context Learning for Speech Language Model

Outline

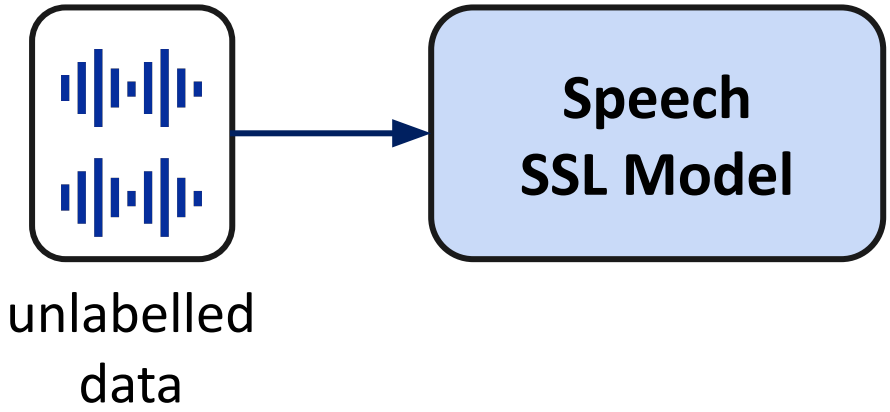
- Pre-train, fine-tune paradigm vs.Prompting paradigm
- SpeechPrompt: Prompting for speech processing
  - Textless speech LM
  - SpeechPrompt for various tasks
  - In-context Learning for Speech Language Model
- Related Works
- Conclusion
- Future Work

Pre-train, fine-tune paradigm
vs.
Prompting paradigm

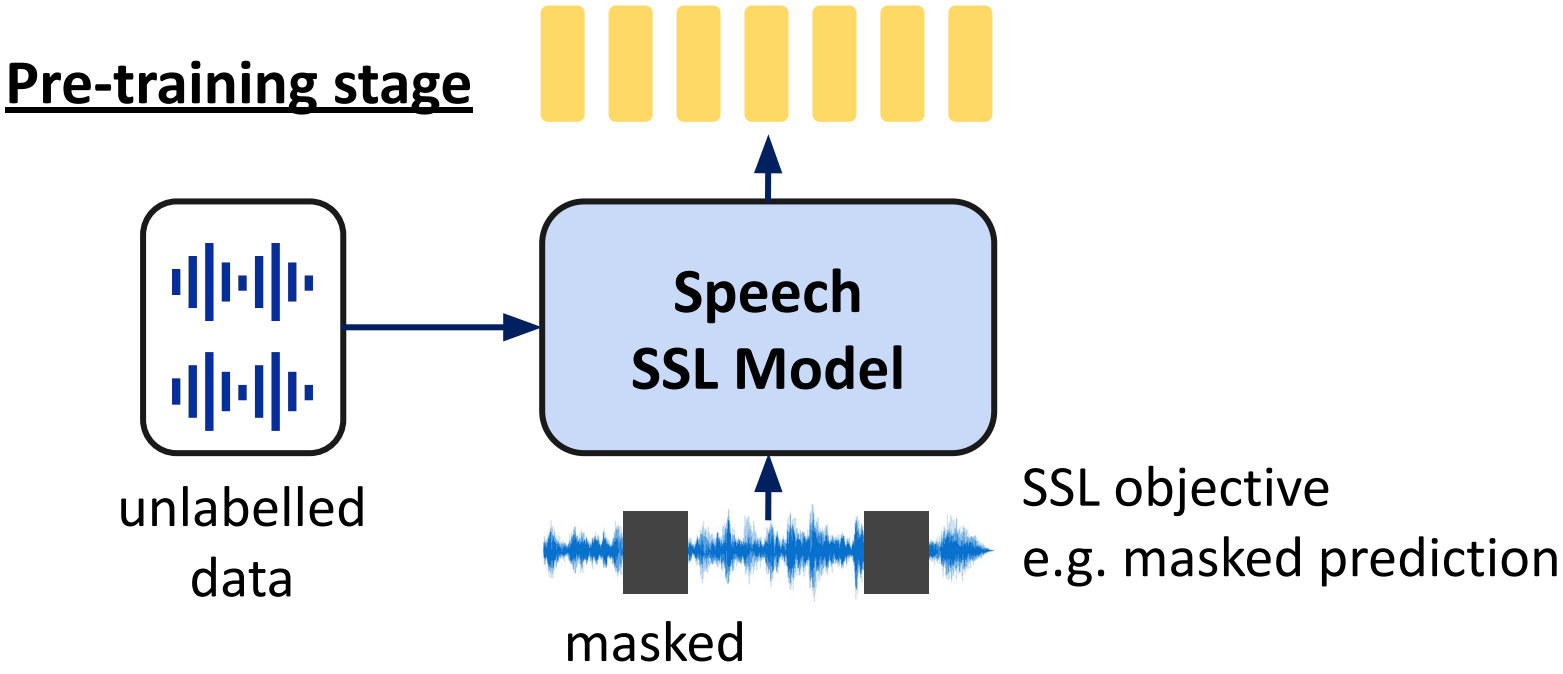

Pre-train, Fine-tune Paradigm
Self-supervised Learning (SSL)
Prompting Paradigm
Prompt engineering
BERT,…
wav2vec 2.0,…
GPT-3
ChatGPT
2018
2019
2020
2021
2022
2023

Pre-train, Fine-tune Paradigm

Pre-training stage

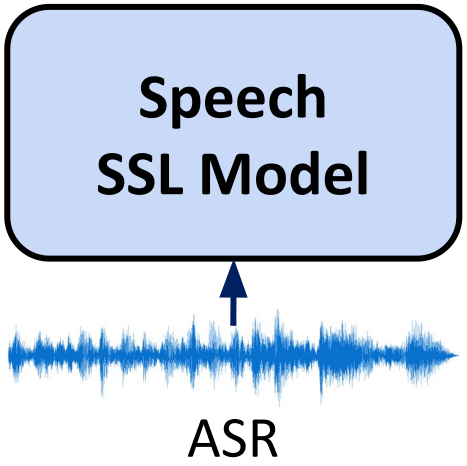


Pre-train, Fine-tune Paradigm



Pre-train, Fine-tune Paradigm

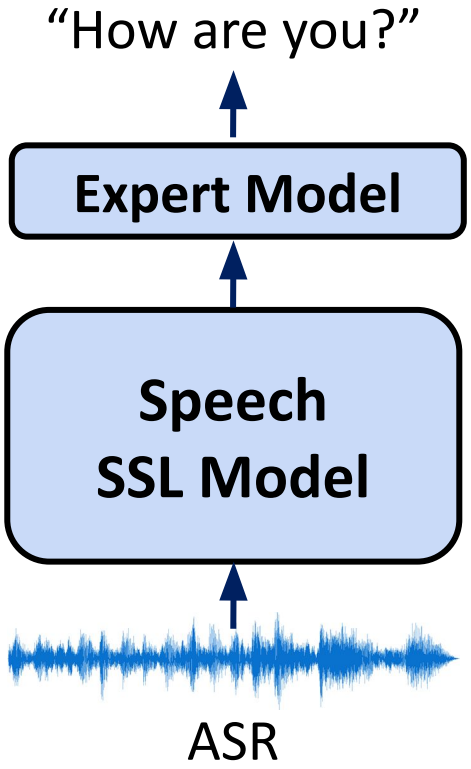
For a **downstream task**: **Fine-tuning stage**



Pre-train, Fine-tune Paradigm

For a **downstream task**: **<u>Fine-tuning stage</u>**

1. design a downstream expert model

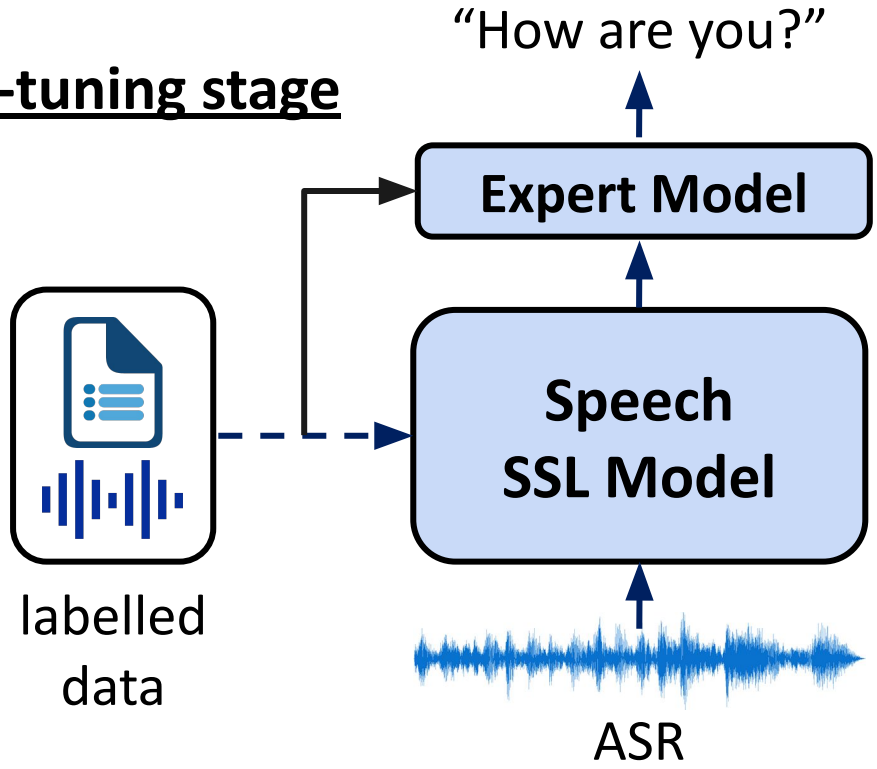


Pre-train, Fine-tune Paradigm

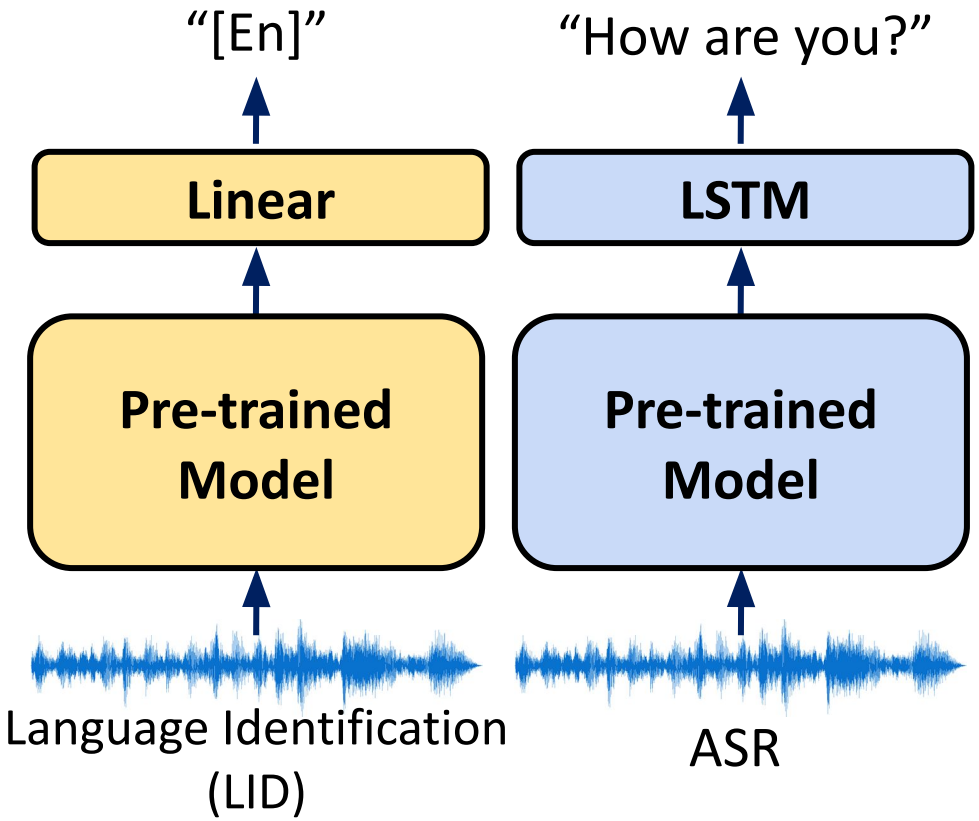
For a **downstream task**:

1. design a downstream expert model
2. **fine-tune** the head and the pre-trained model with a specific loss function (e.g. CTC loss)

Fine-tuning stage



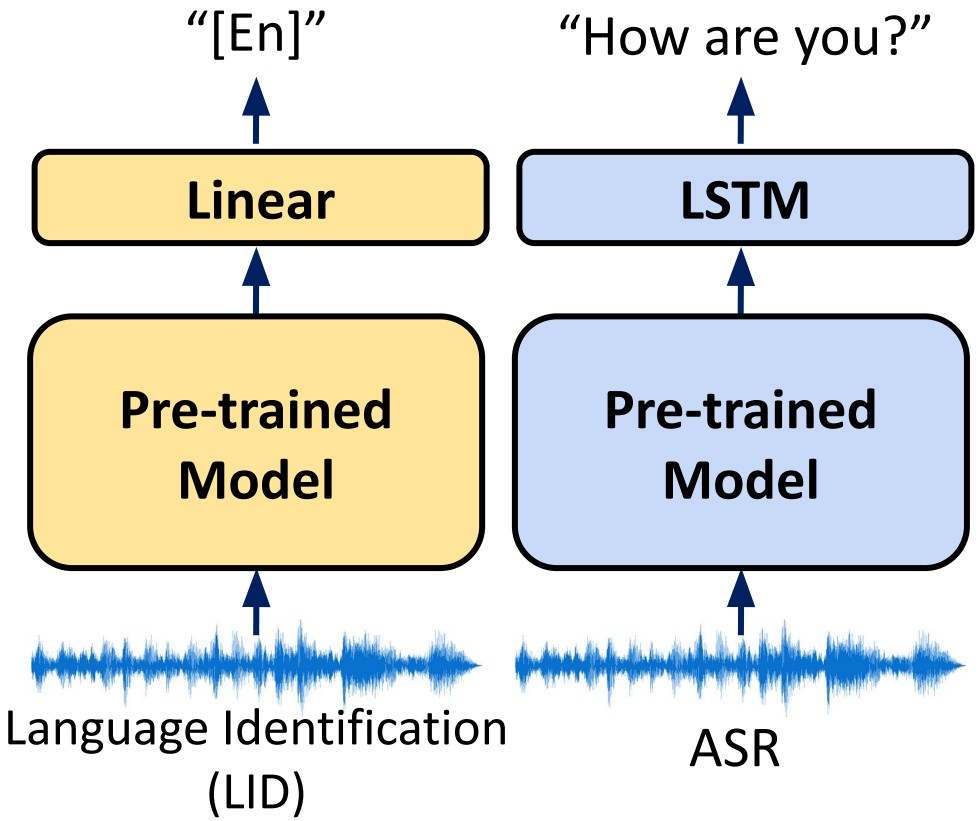
Pre-train, Fine-tune Paradigm



To perform a downstream task:

- design an expert model
- fine-tune the model
- save the parameters

Pre-train, Fine-tune Paradigm

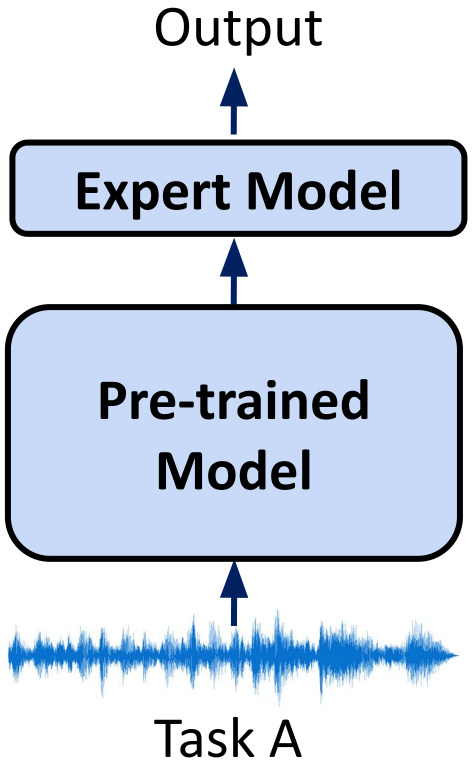


To perform a downstream task:

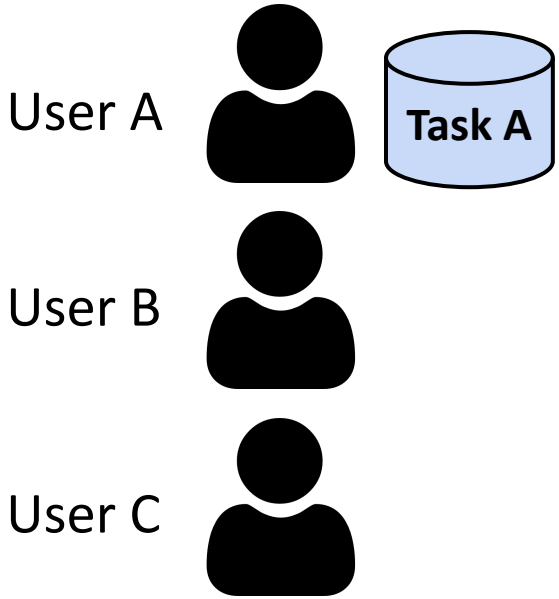
- design an expert model
- fine-tune the model
- save the parameters

Achieve good performance

Pre-train, Fine-tune Paradigm

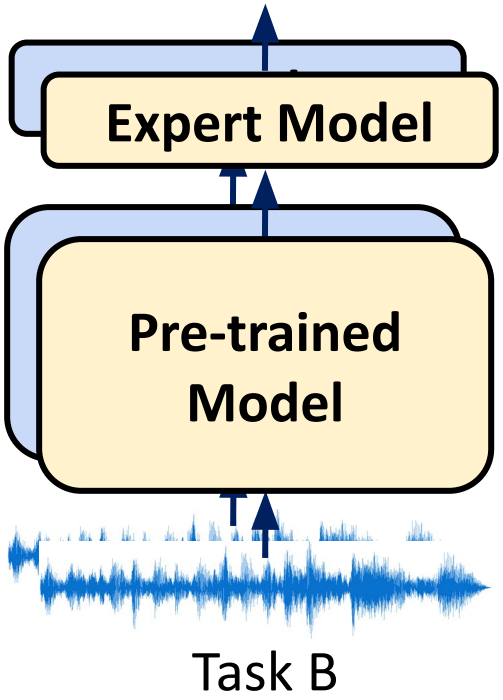


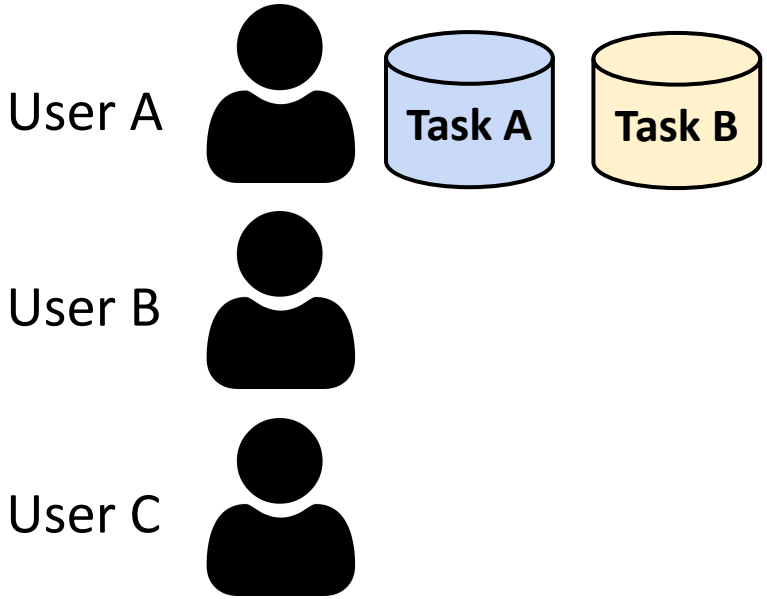
If you want to serve lots of users…



Pre-train, Fine-tune Paradigm

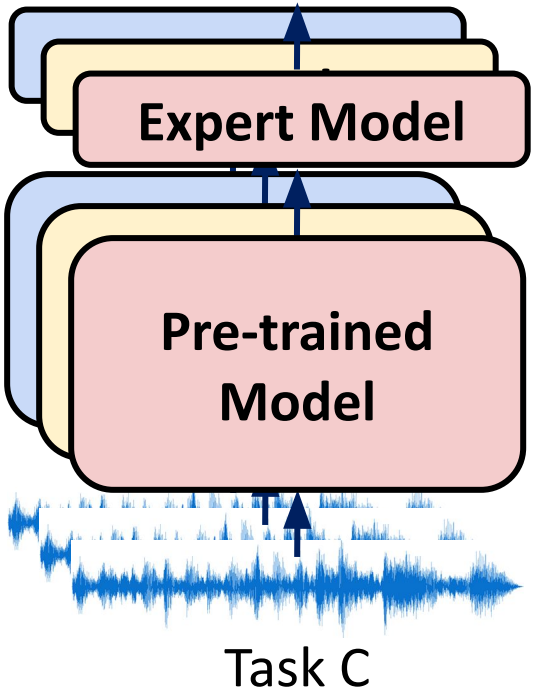
If you want to serve lots of users…

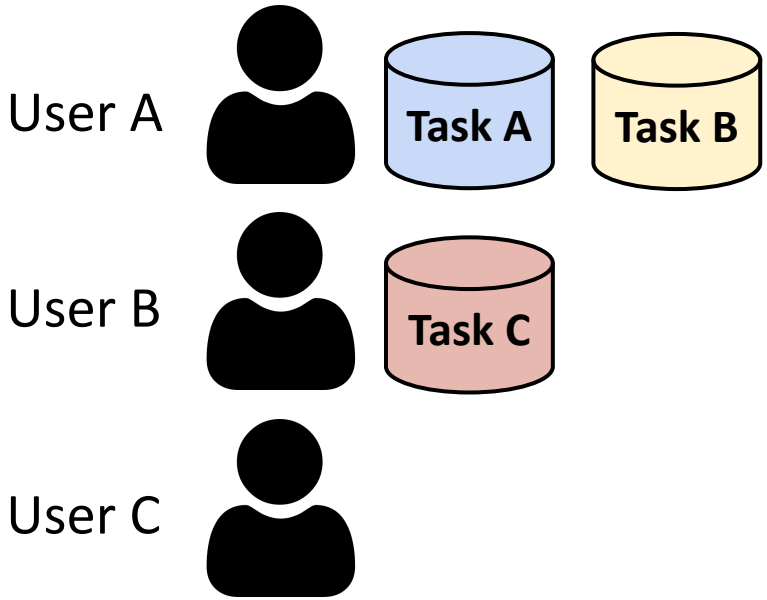




Pre-train, Fine-tune Paradigm

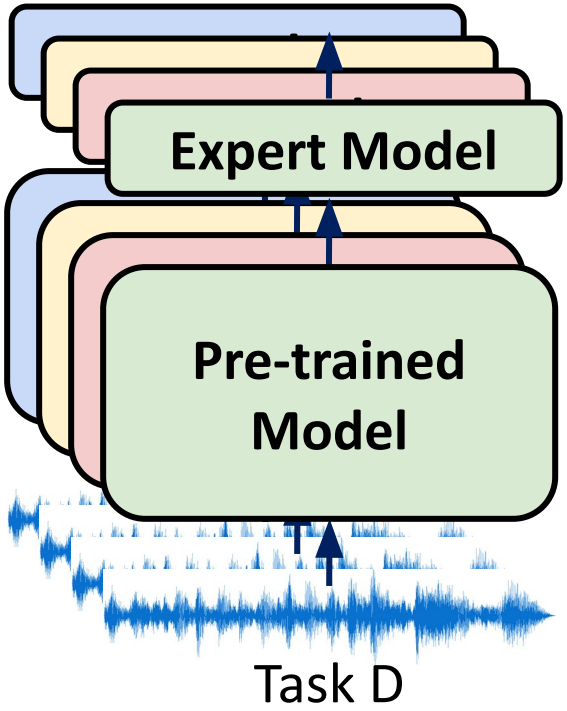
If you want to serve lots of users…

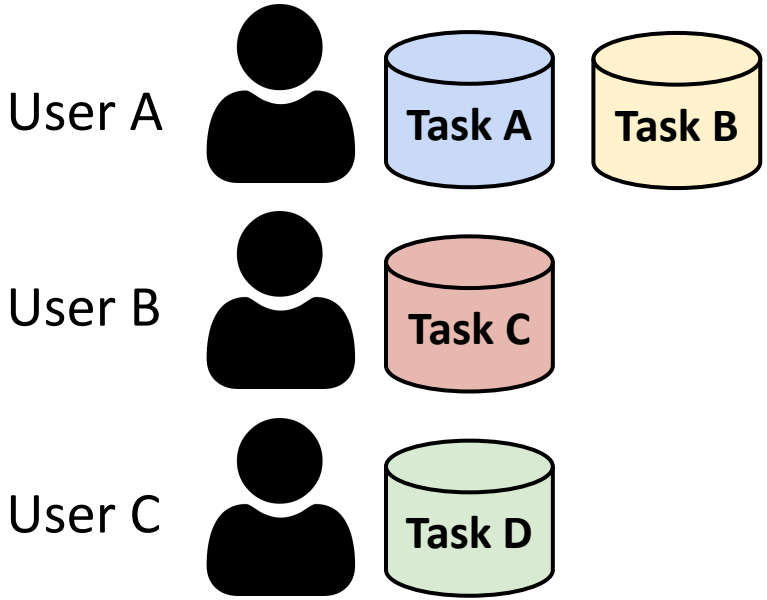




Pre-train, Fine-tune Paradigm

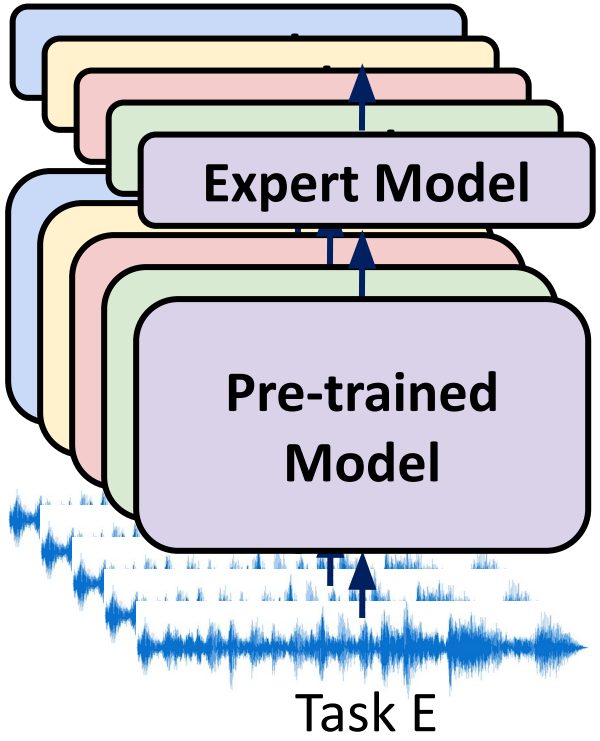
If you want to serve lots of users…

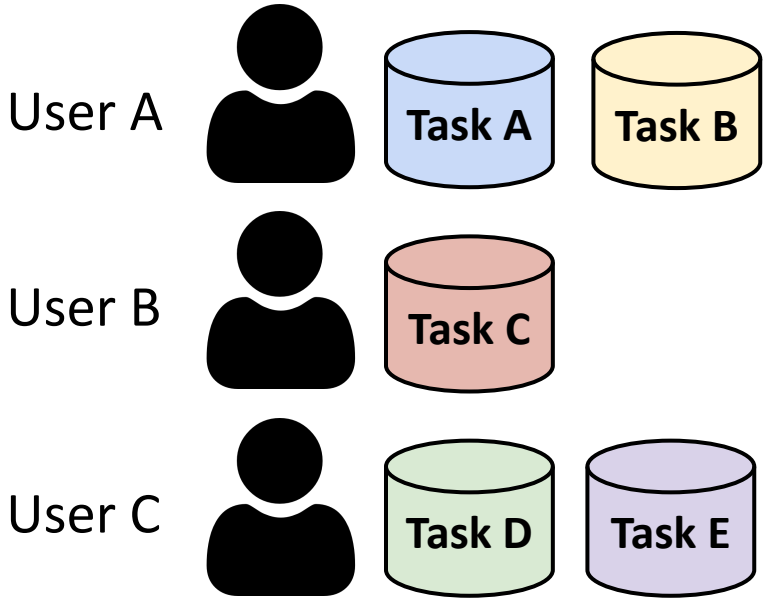




Pre-train, Fine-tune Paradigm

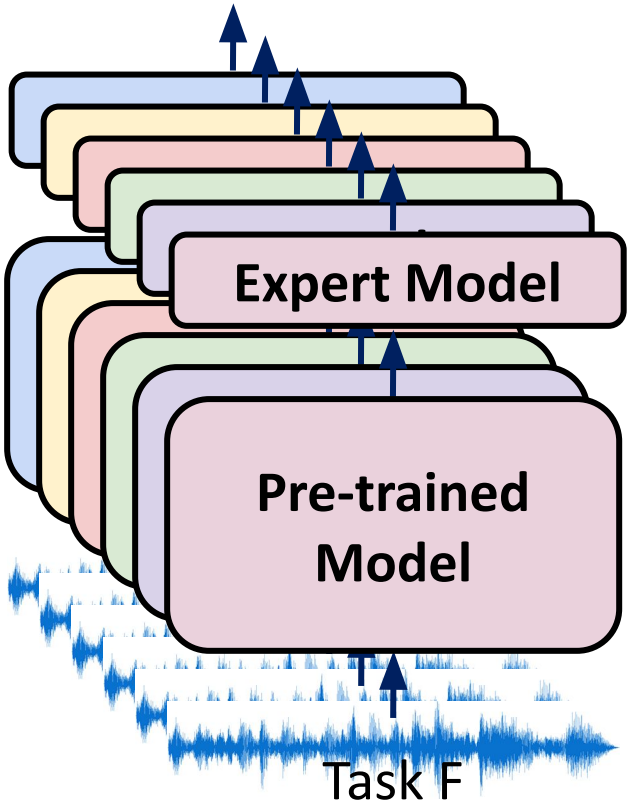
If you want to serve lots of users…

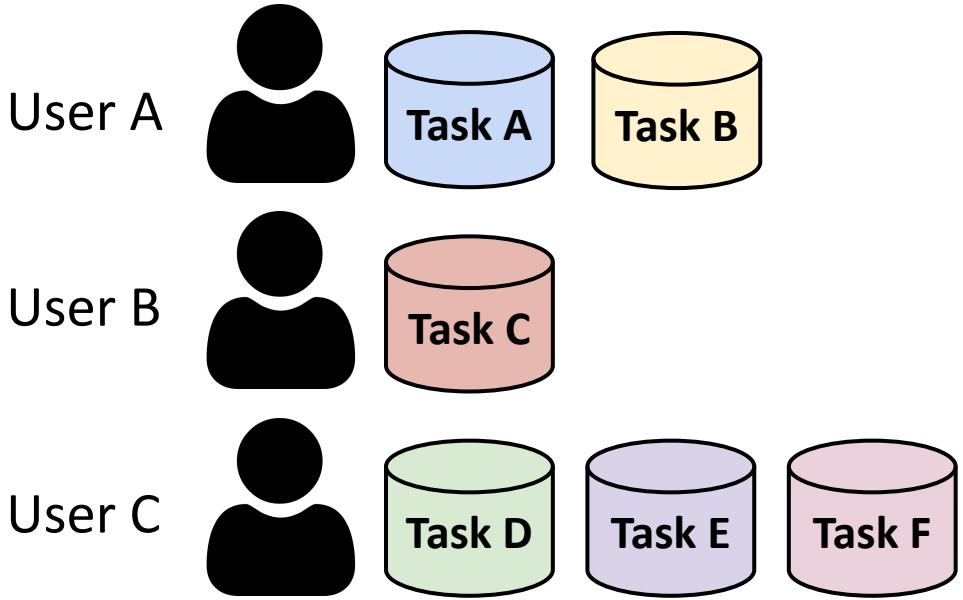




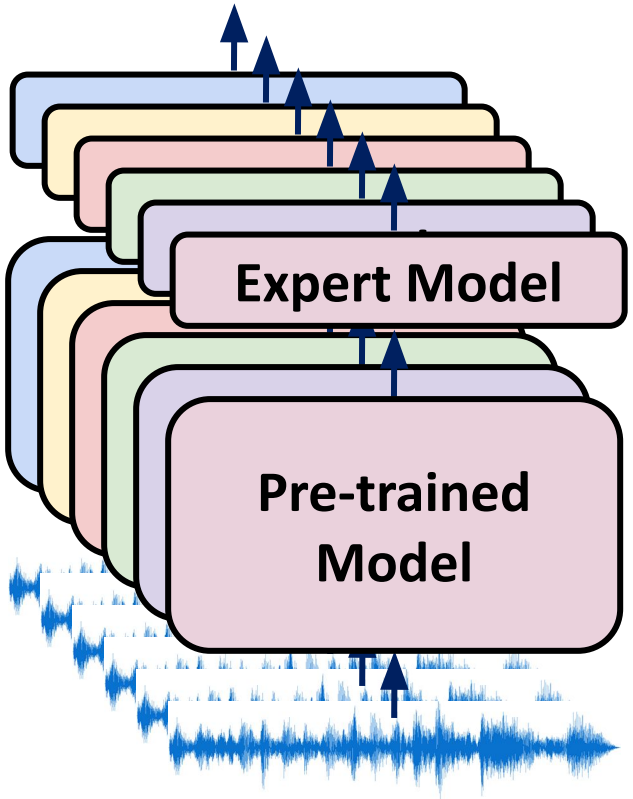
Pre-train, Fine-tune Paradigm

If you want to serve lots of users…





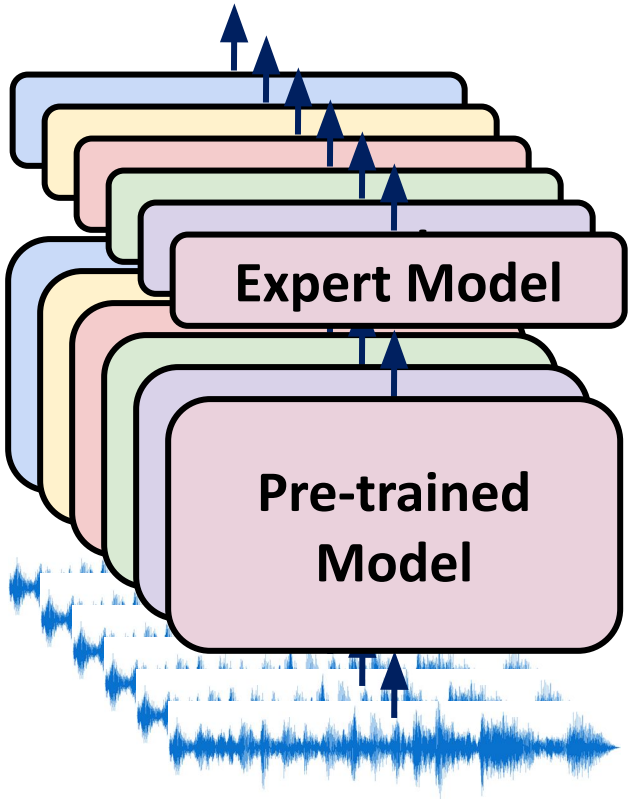
Pre-train, Fine-tune Paradigm



If there are lots of tasks to serve…

- design a head
  human labor
- fine-tune the model
  computational cost
- save the parameters
  storage cost

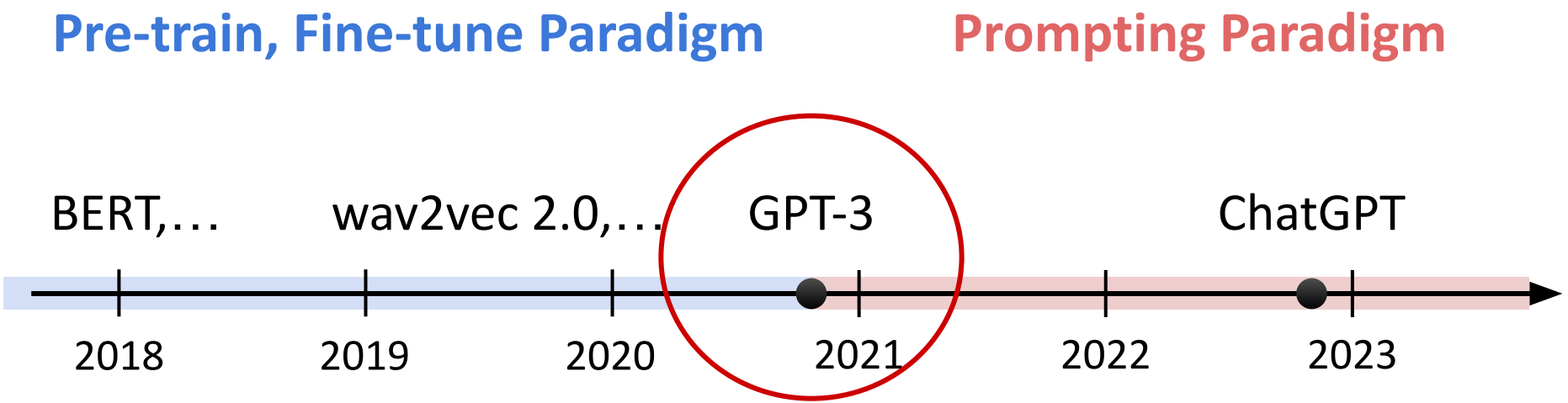
Pre-train, Fine-tune Paradigm



If there are lots of tasks to serve…

- design a head
  human labor
- fine-tune the model
  computational cost
- save the parameters
  storage cost

Difficult to scale!

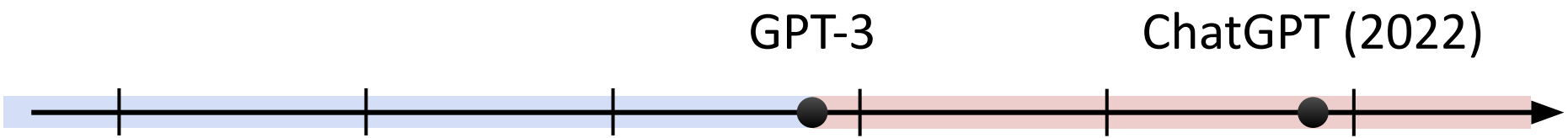


Prompting gained more and more attention

Prompting Paradigm

“prompt” a language model to perform various downstream tasks

Language Model



Prompting Paradigm

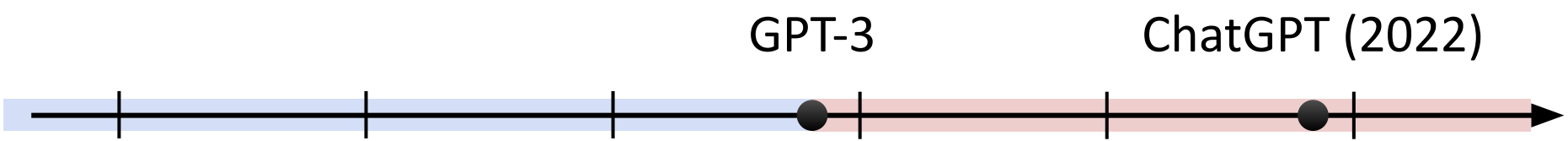
“prompt” a language model to perform various downstream tasks

Machine translation

Language Model

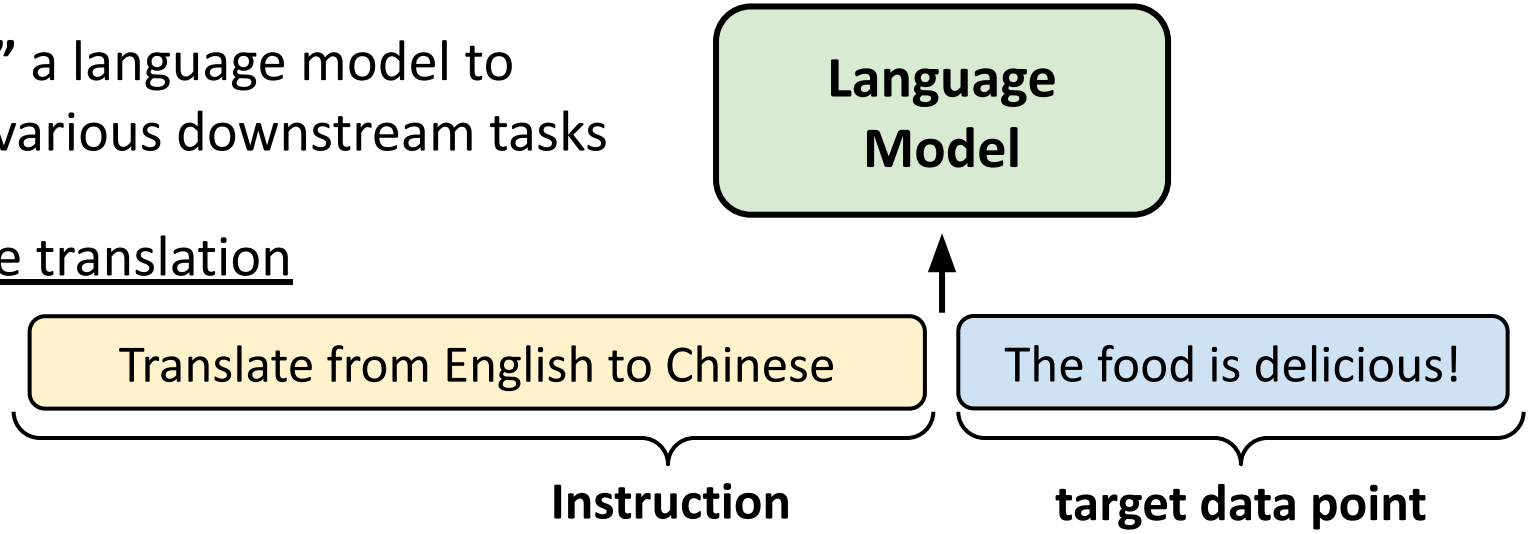
The food is delicious!

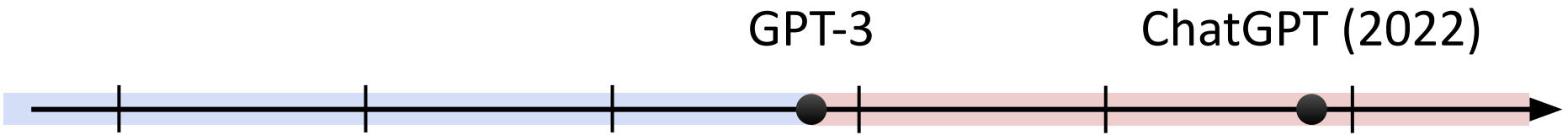
target data point



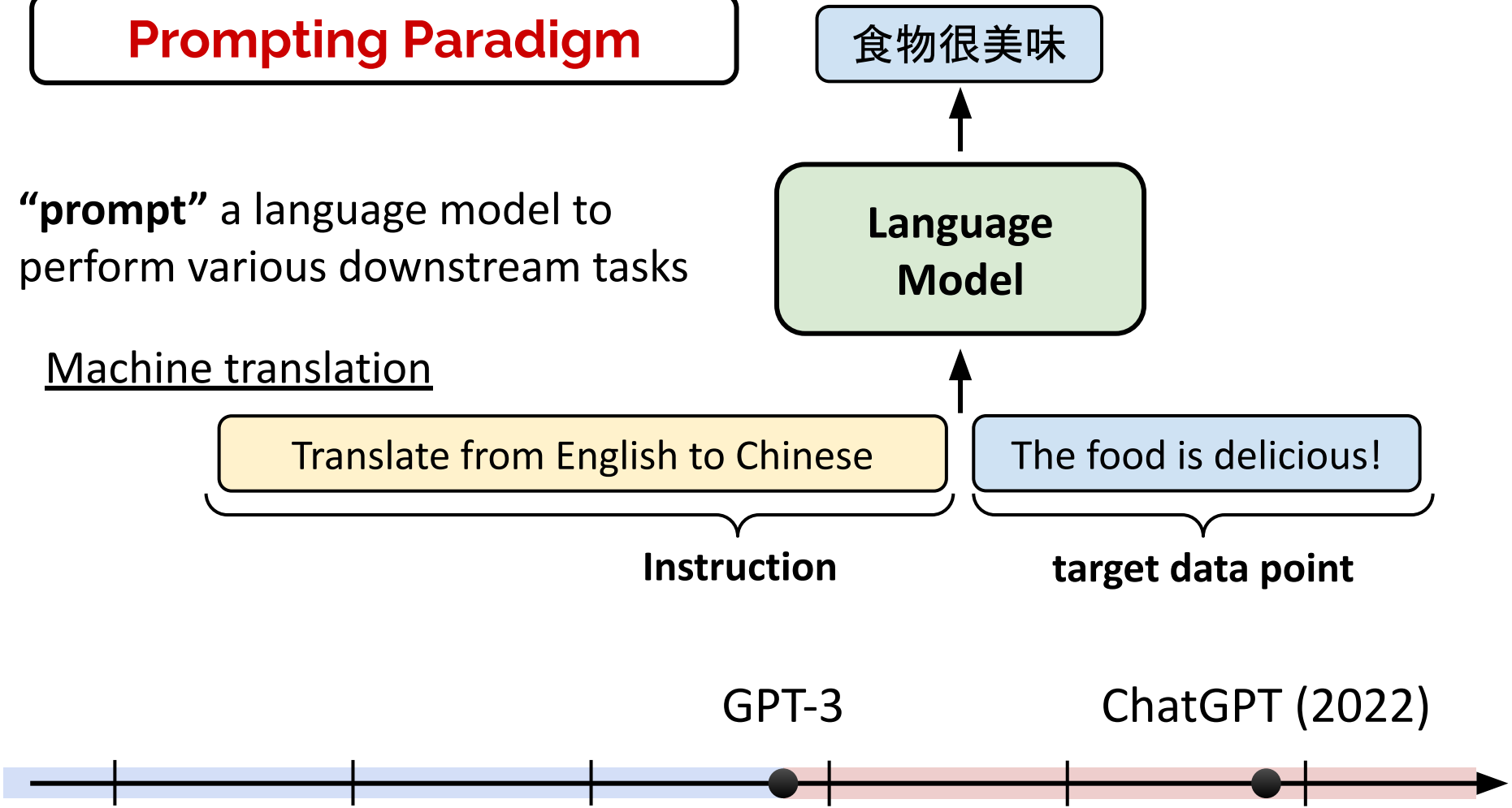
Prompting Paradigm

“prompt” a language model to perform various downstream tasks



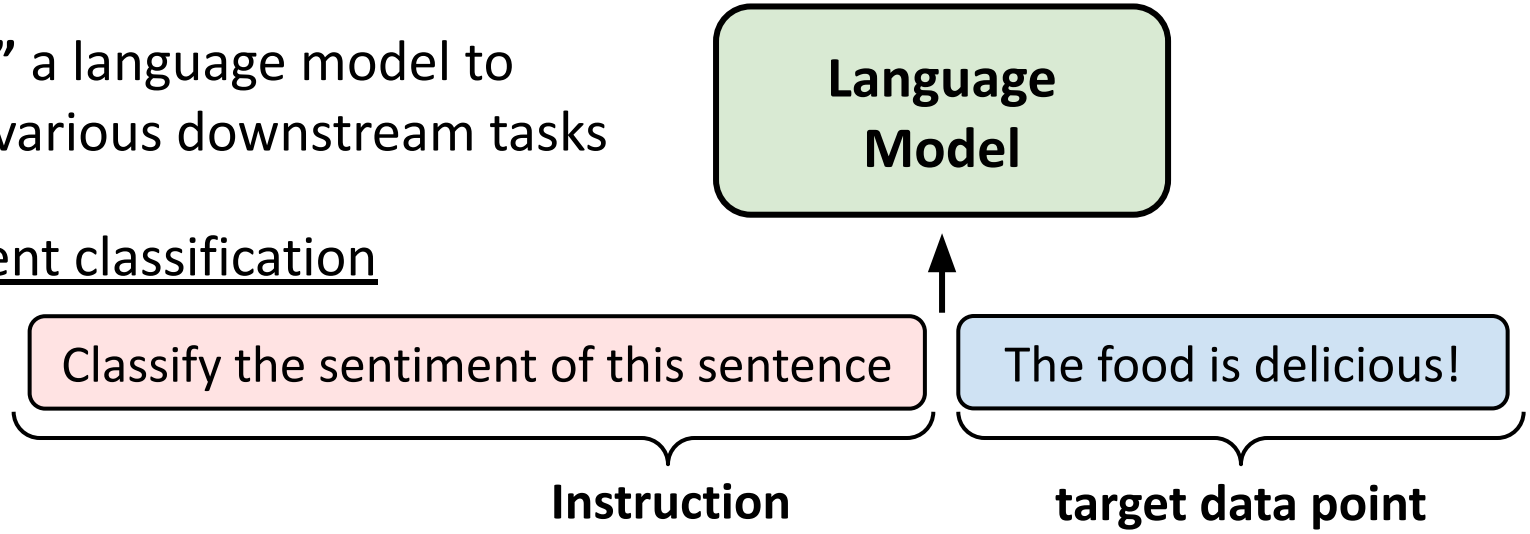


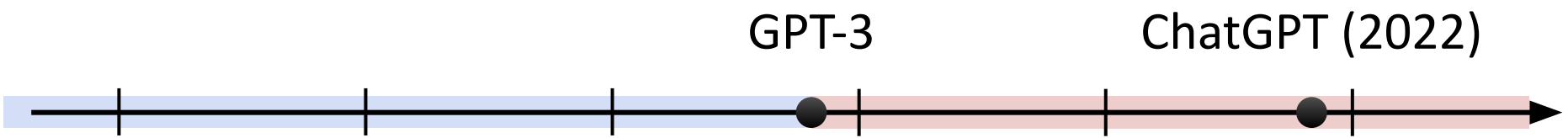
Prompting Paradigm



Prompting Paradigm

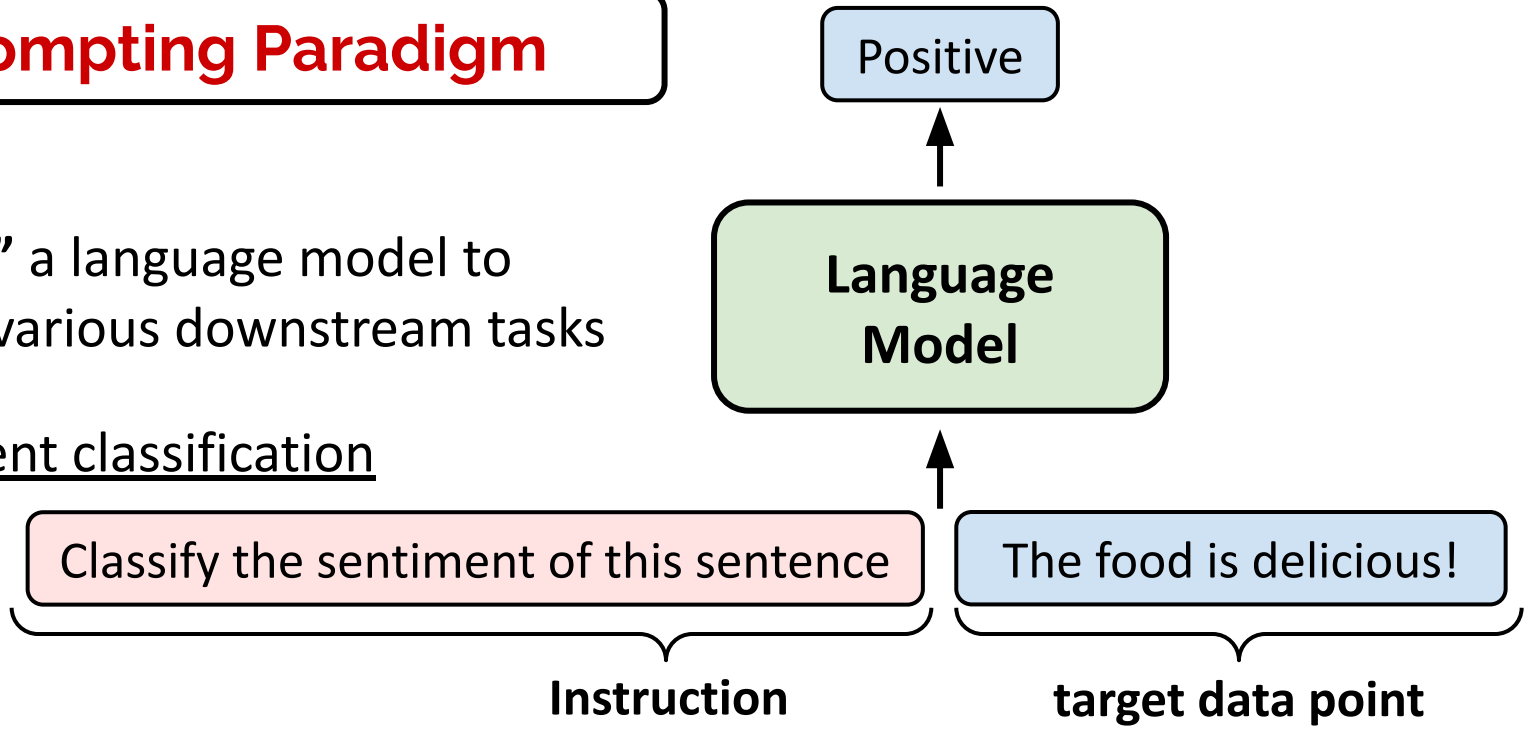
"prompt" a language model to perform various downstream tasks

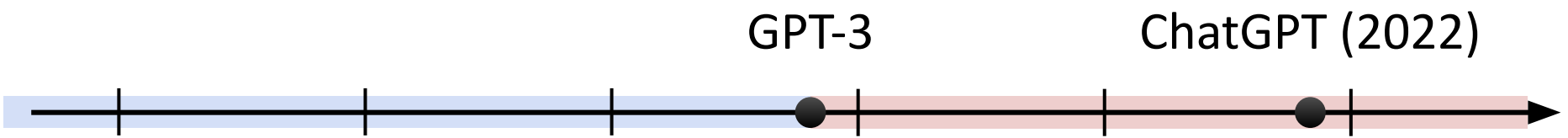




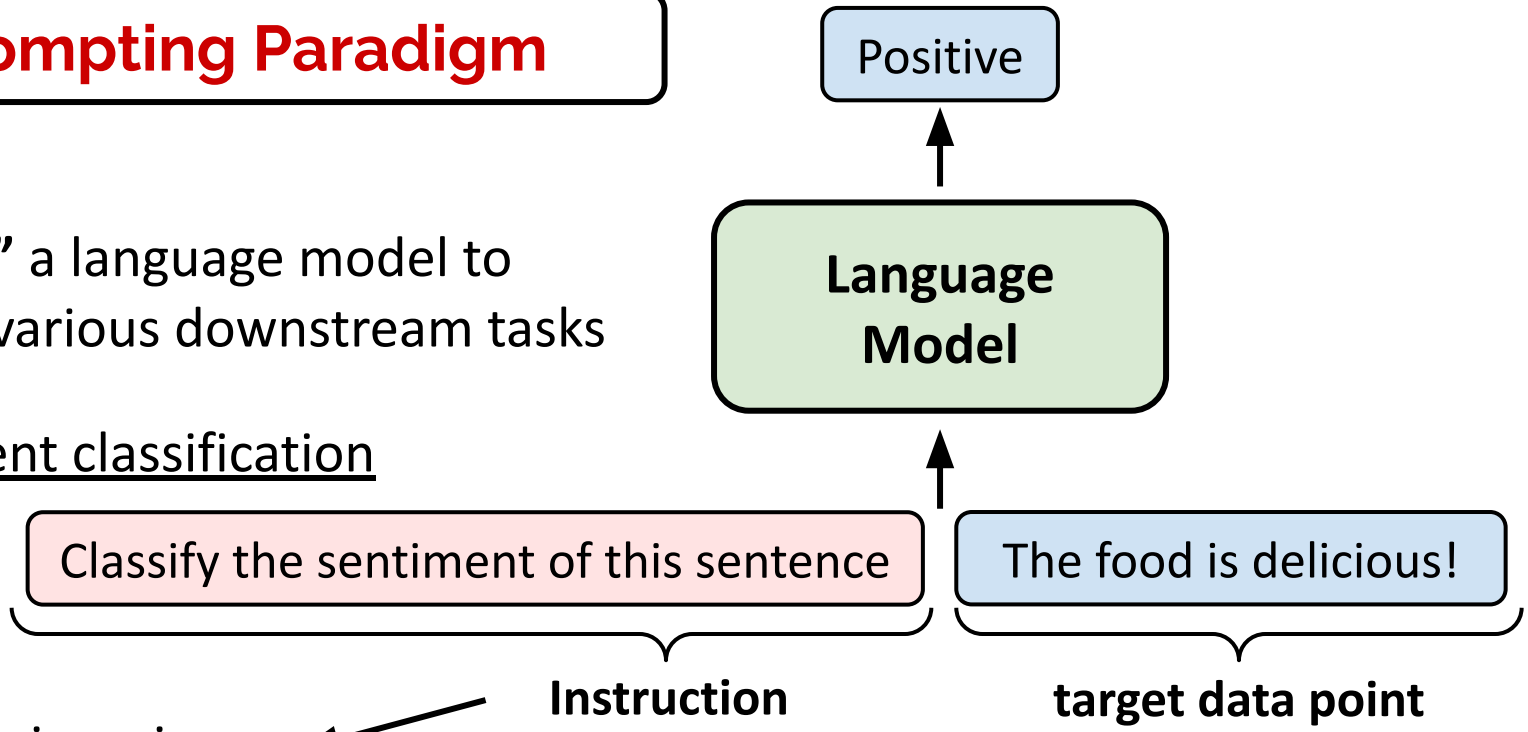
Prompting Paradigm

"prompt" a language model to perform various downstream tasks





Prompting Paradigm



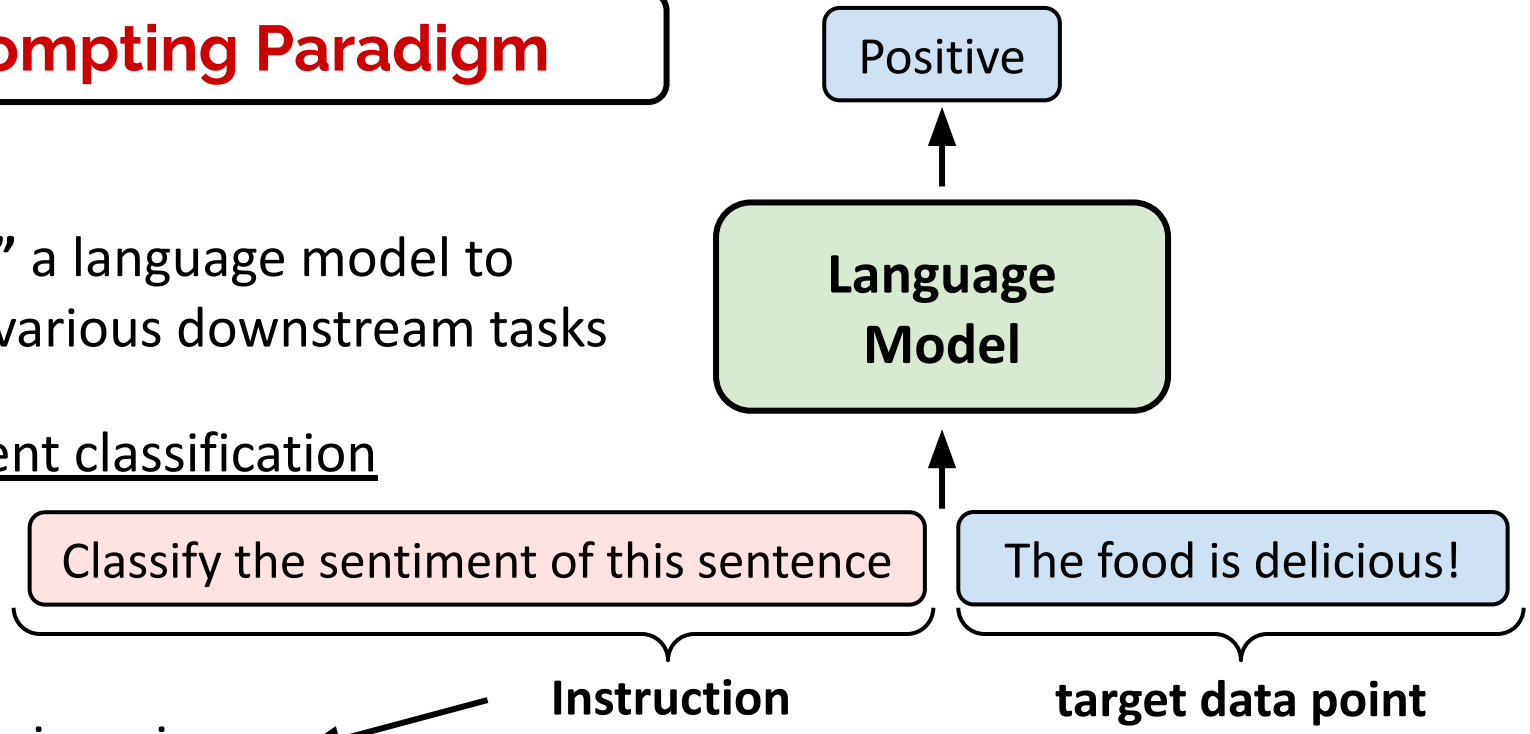
“prompt” a language model to perform various downstream tasks

Sentiment classification

Prompt Engineering

- Natural language (Discrete prompts) - Manually designed, interpretable.
- Continuous vectors (Soft prompts) - Trainable, more capable, difficult to interpret.

Prompting Paradigm



“prompt” a language model to perform various downstream tasks

Sentiment classification

Prompt Engineering

- Natural language (Discrete prompts) - Manually designed, interpretable.
- Continuous vectors (Soft prompts) - Trainable, more capable, difficult to interpret.

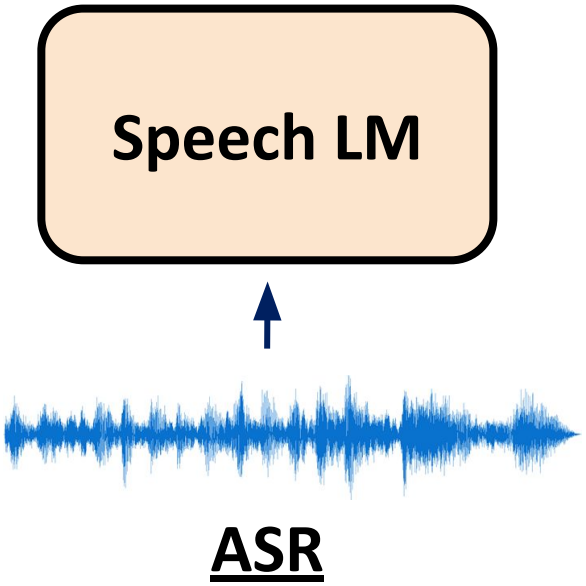
Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (ACM Computing Surveys 2023, https://arxiv.org/abs/2107.13586)

Prompting Paradigm

“prompt” a speech language model to perform various downstream tasks

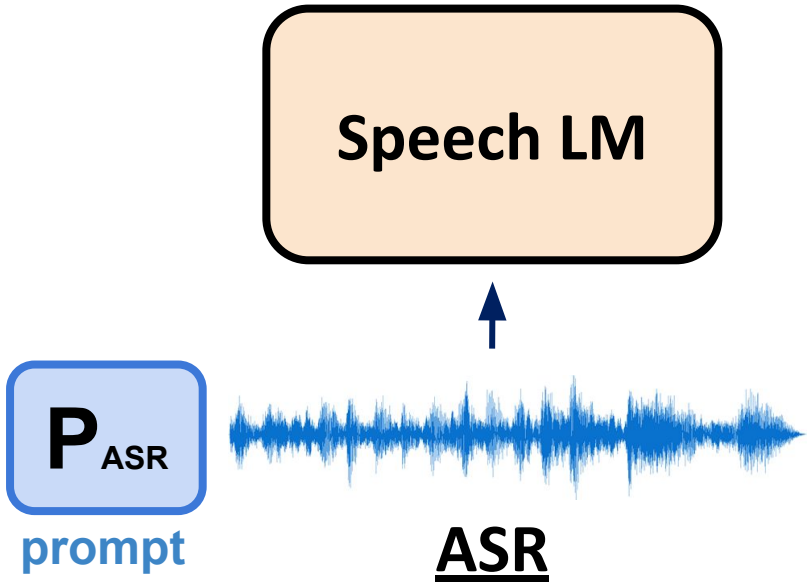
Prompting Paradigm

“prompt” a speech language model to perform various downstream tasks



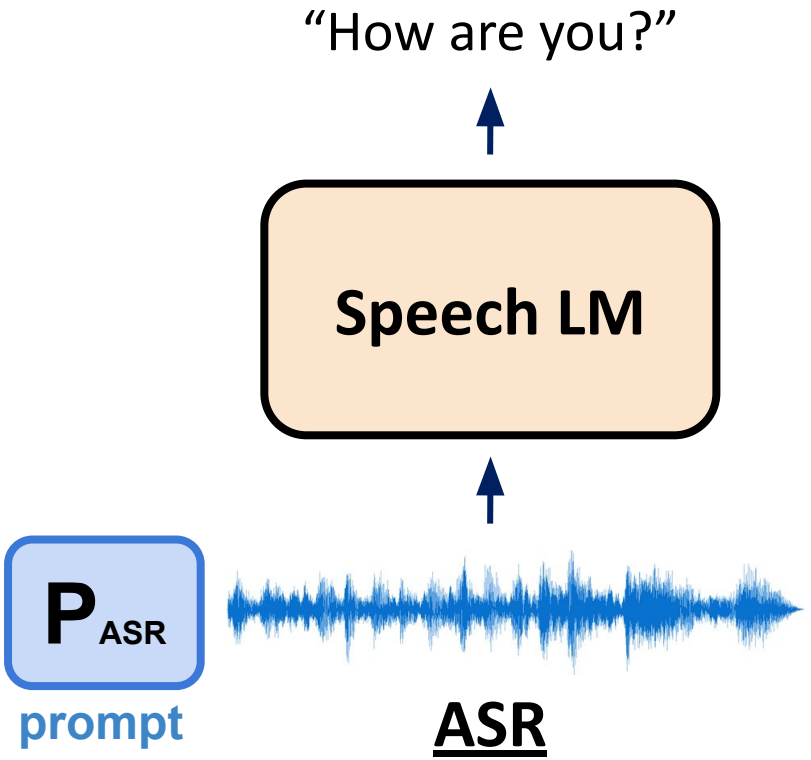
Prompting Paradigm

“prompt” a speech language model to perform various downstream tasks



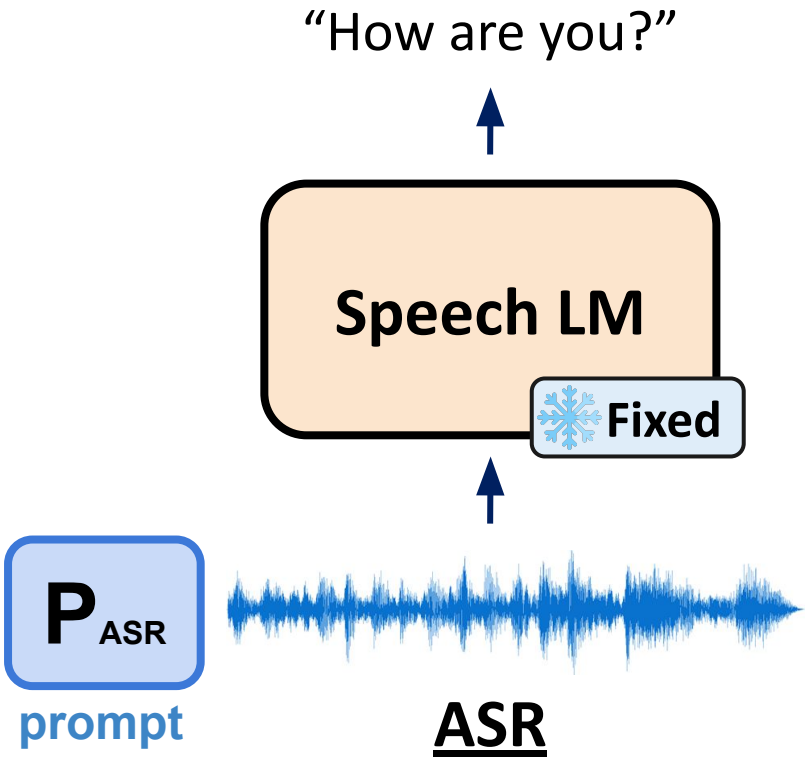
Prompting Paradigm

“prompt” a speech language model to perform various downstream tasks



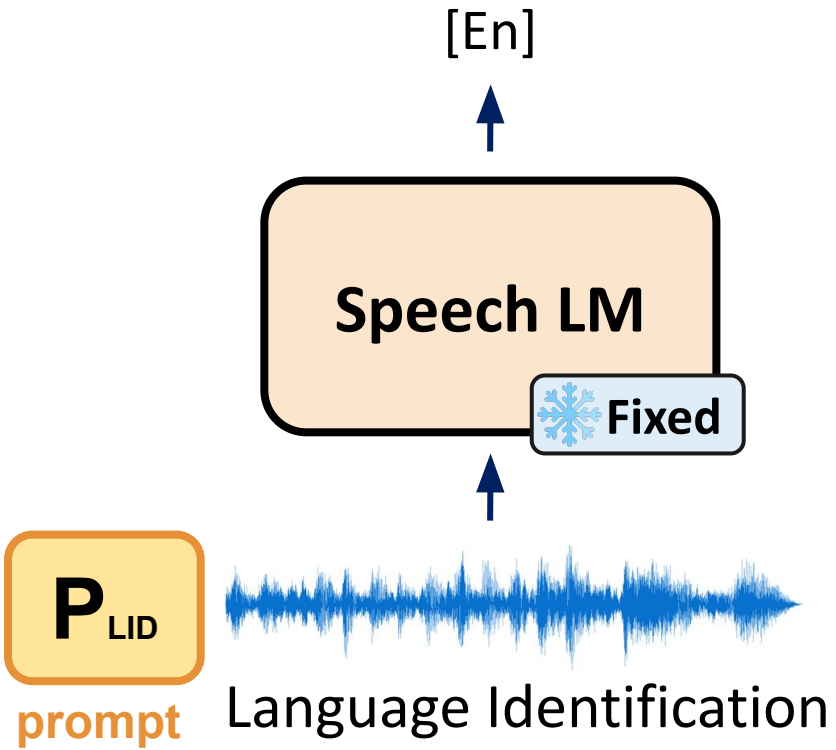
Prompting Paradigm

“prompt” a speech language model to perform various downstream tasks



Prompting Paradigm

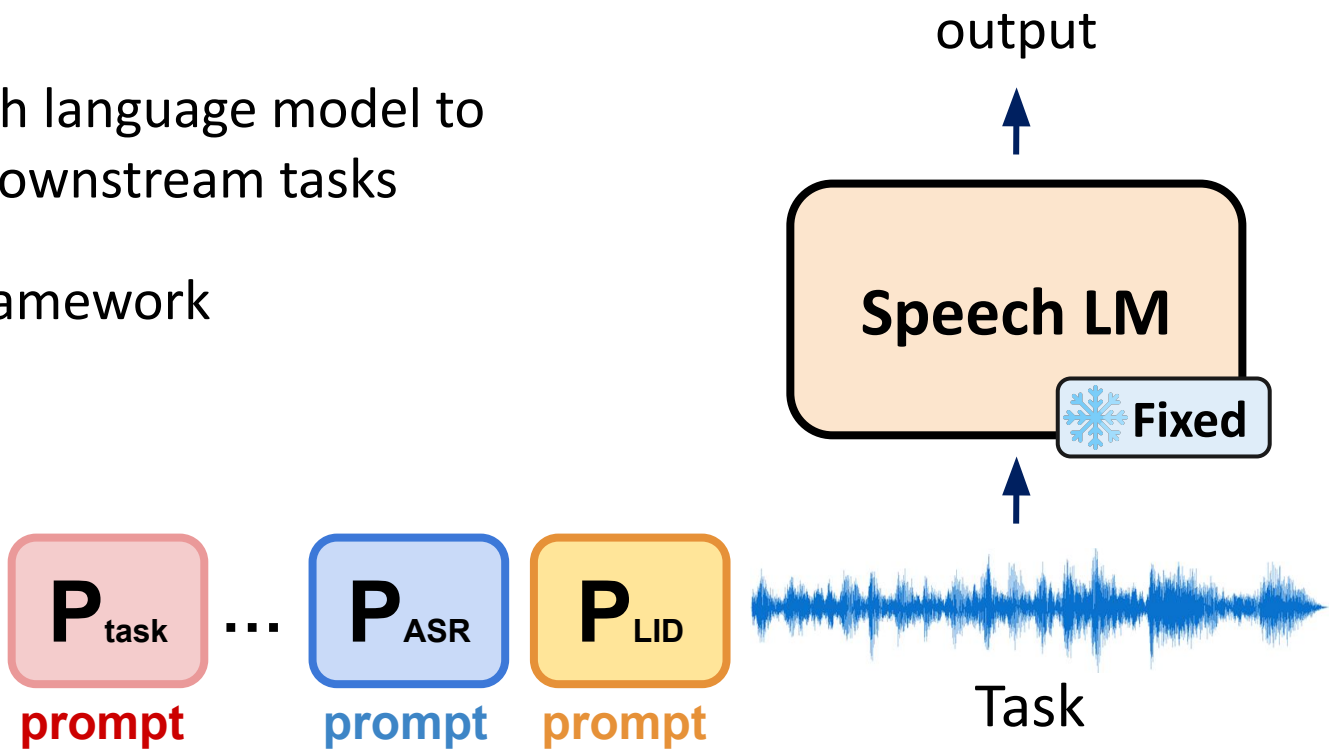
“prompt” a speech language model to perform various downstream tasks



Prompting Paradigm

“prompt” a speech language model to perform various downstream tasks

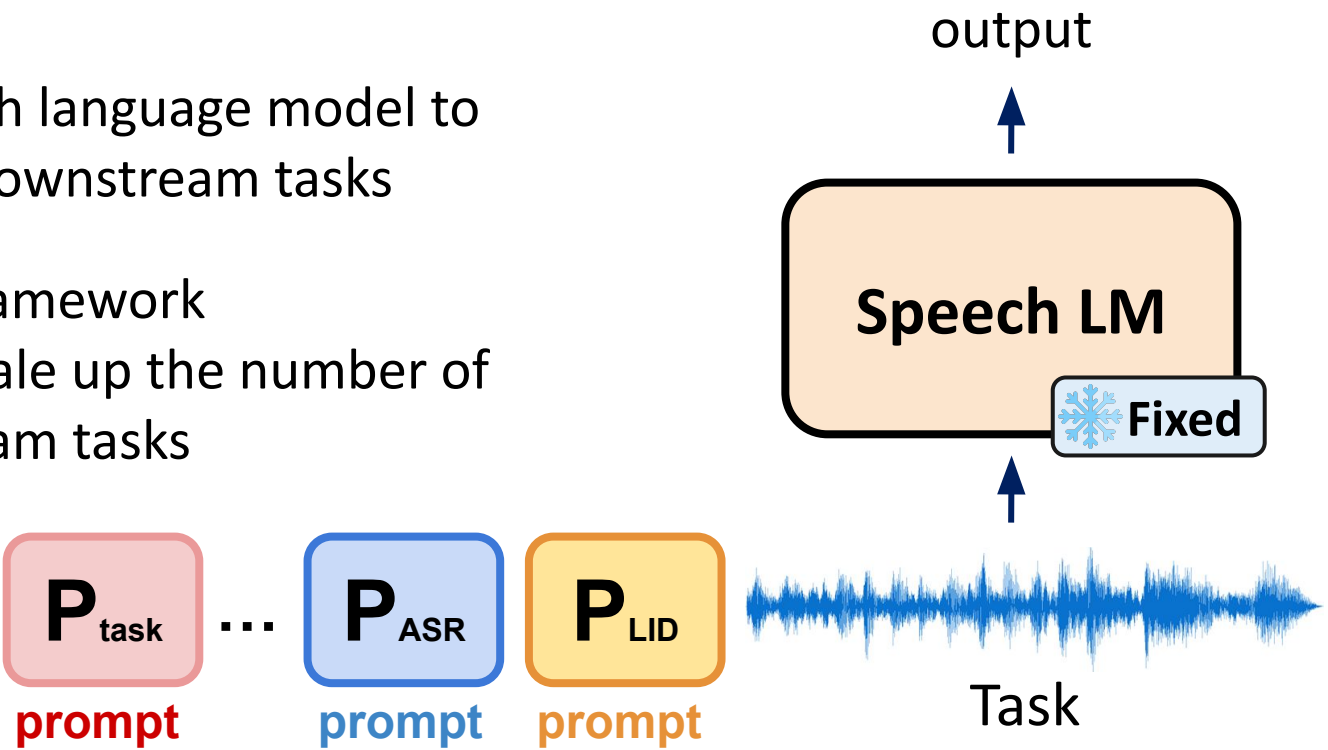
- Unified framework



Prompting Paradigm

“prompt” a speech language model to perform various downstream tasks

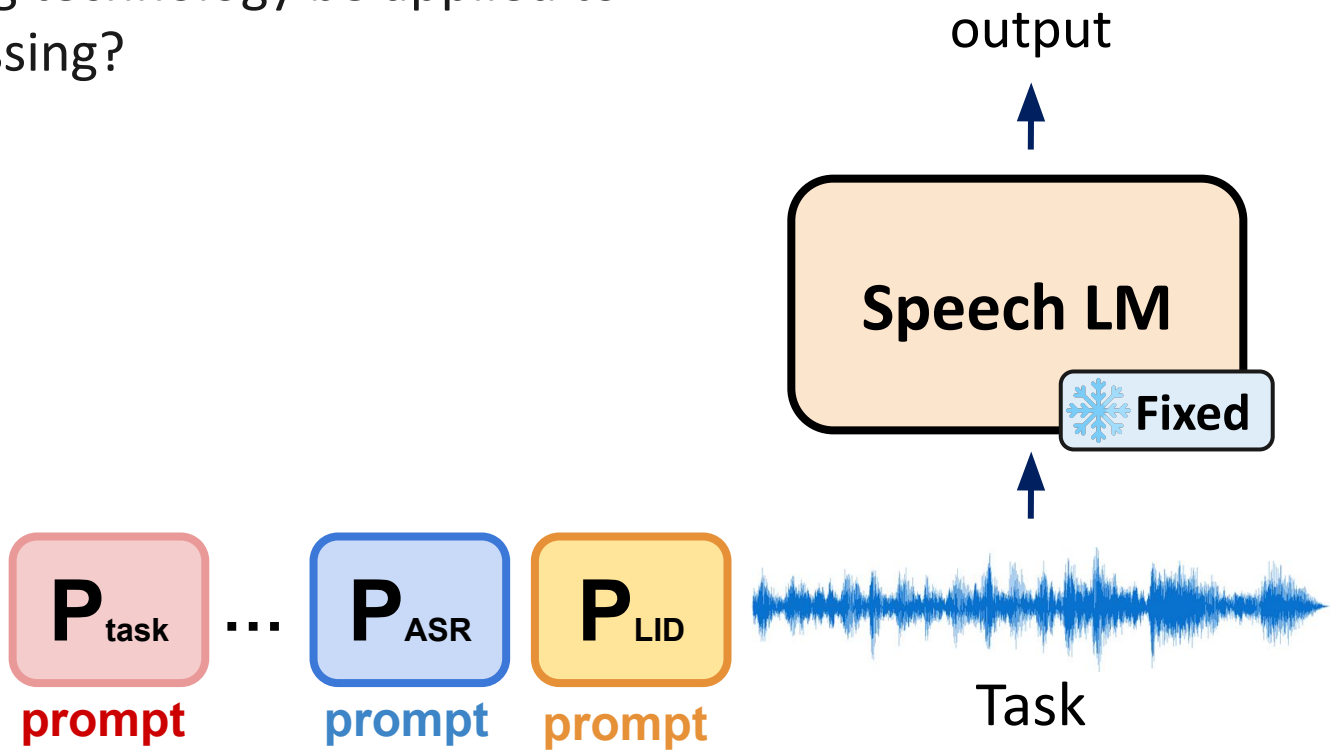
- Unified framework
- Easy to scale up the number of downstream tasks



Contain few parameters

Prompting Paradigm

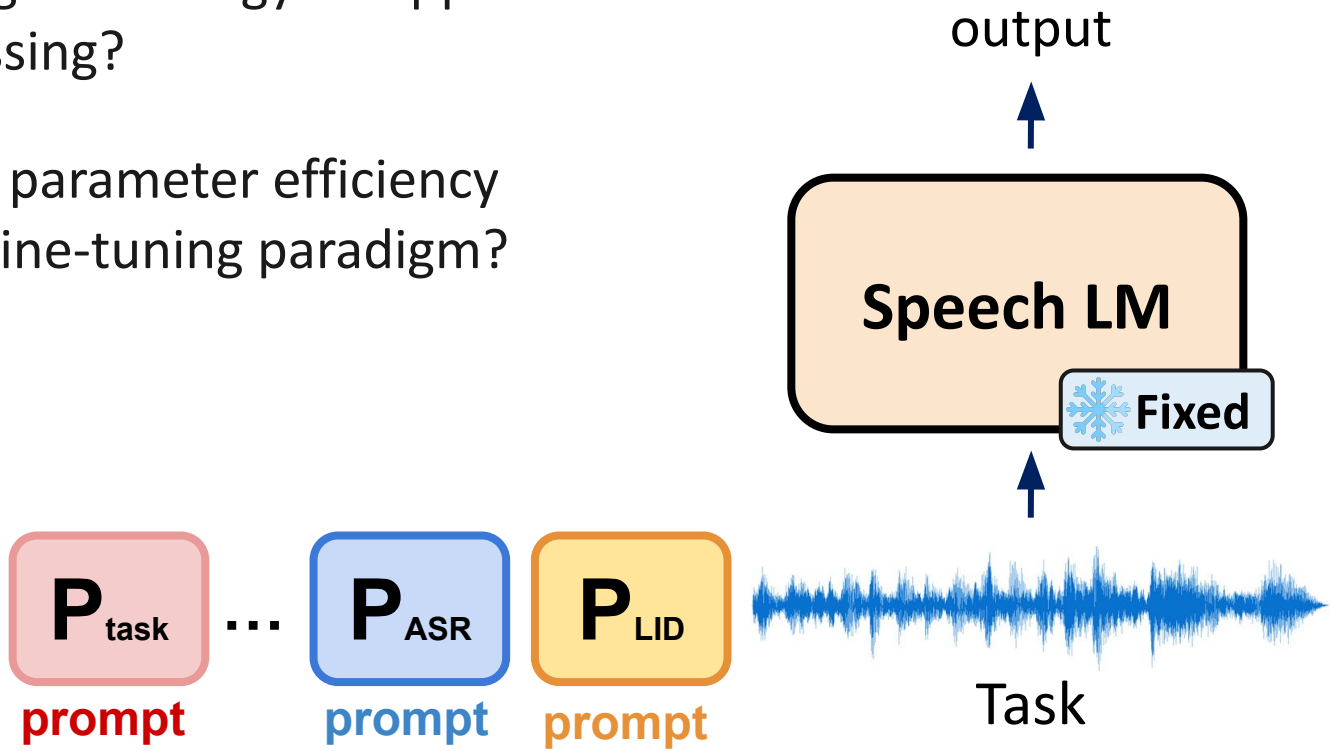
1. Can prompting technology be applied to speech processing?



Prompting Paradigm

1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency compared to fine-tuning paradigm?

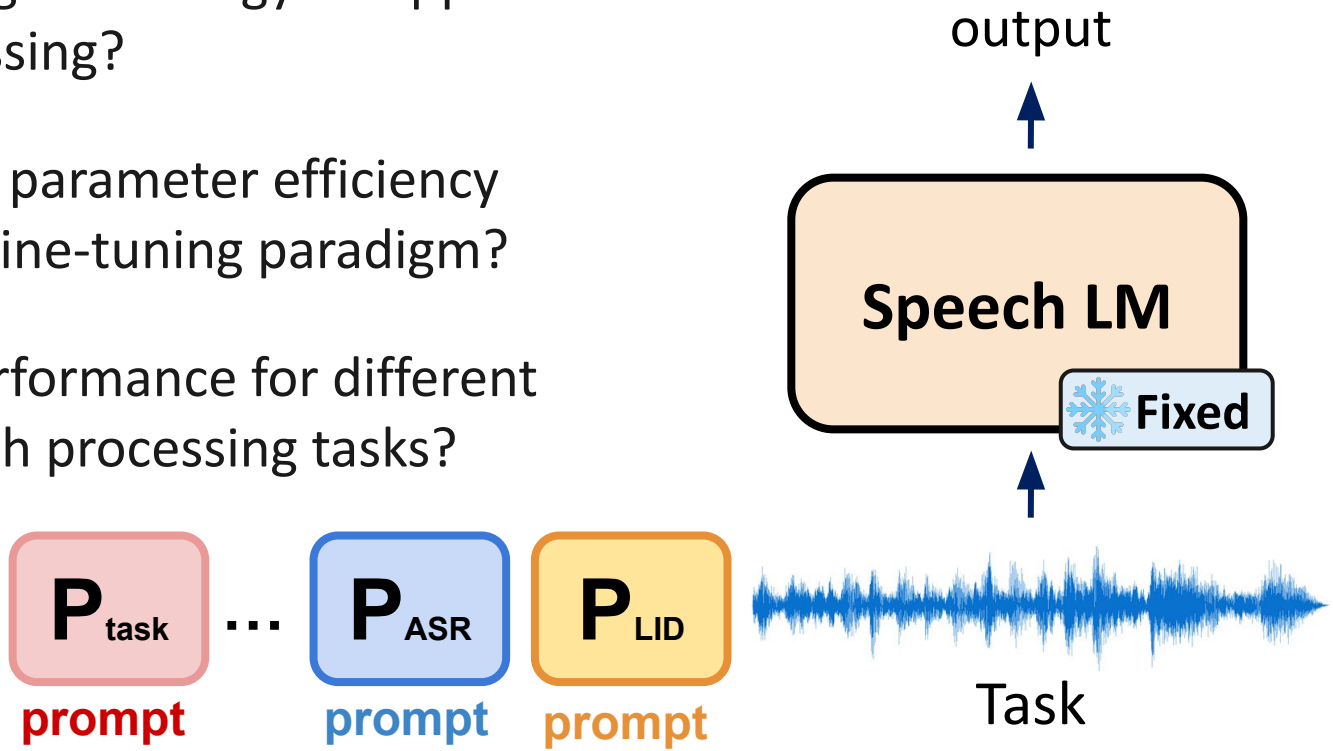


Prompting Paradigm

1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency compared to fine-tuning paradigm?

3. How is the performance for different kinds of speech processing tasks?



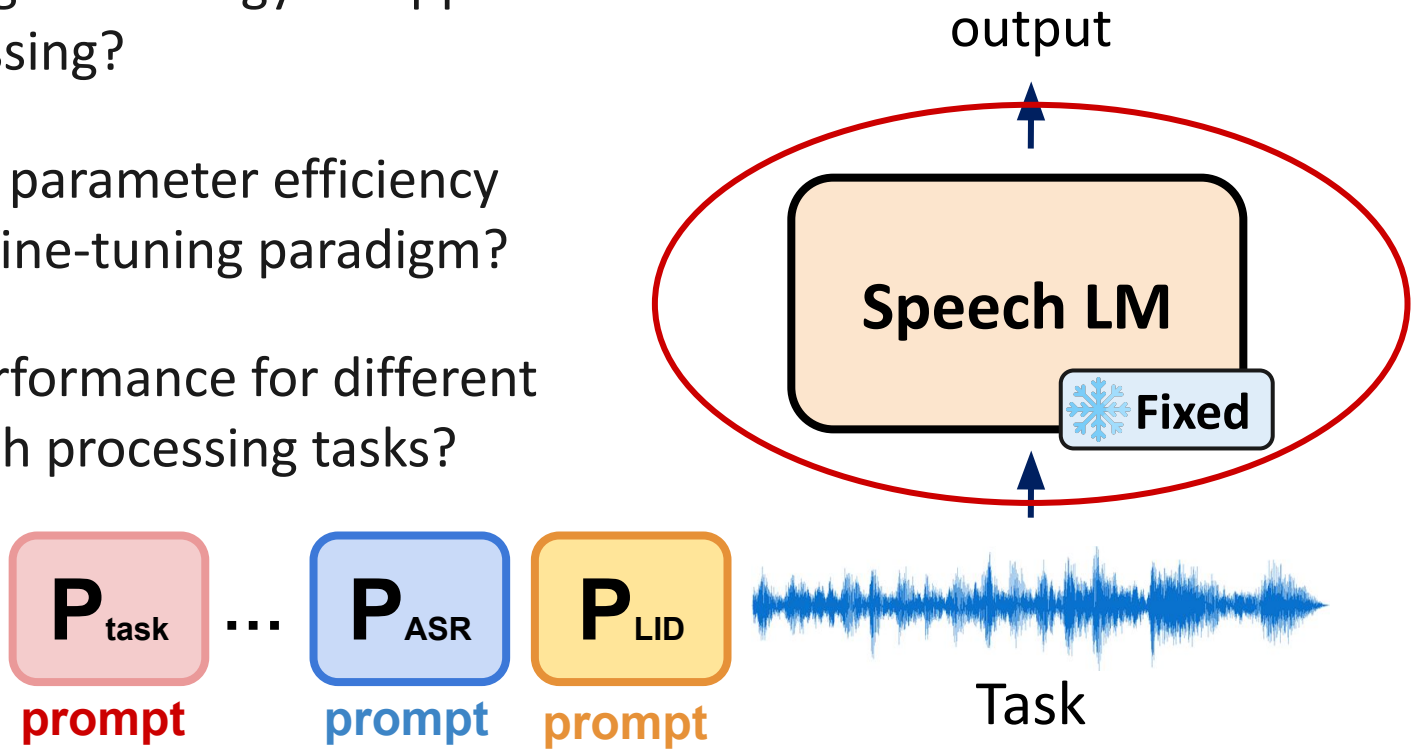
SpeechPrompt

Prompting Paradigm

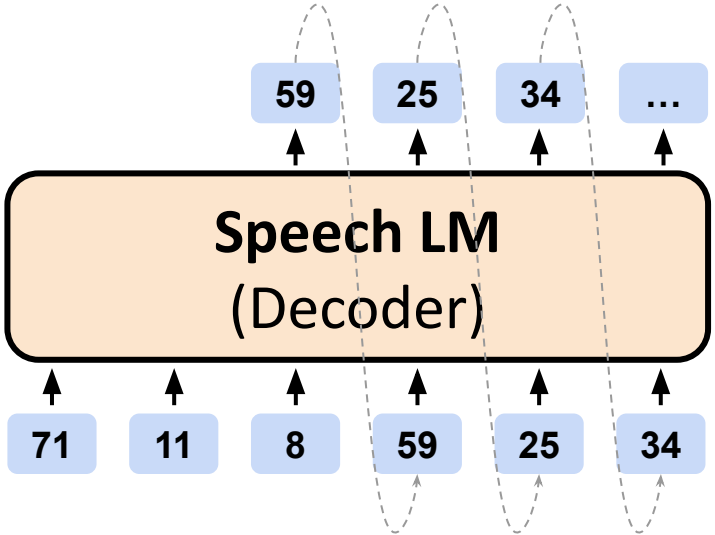
1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency compared to fine-tuning paradigm?

3. How is the performance for different kinds of speech processing tasks?



Textless Speech LM

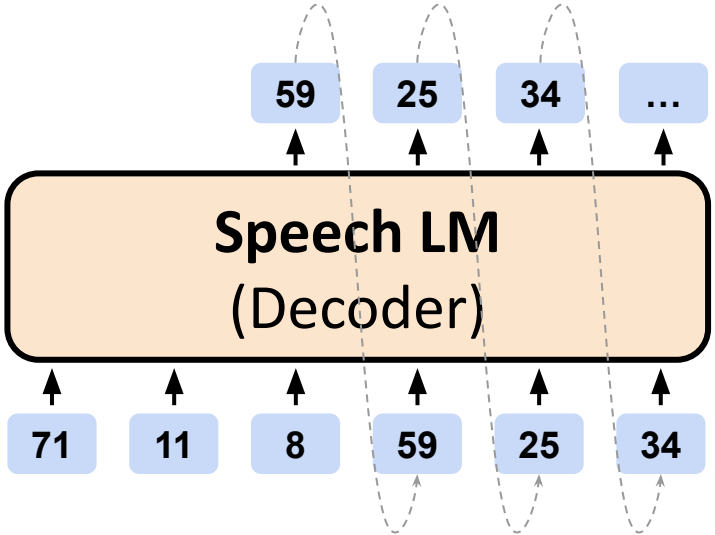


Speech tokens
(Discrete Units)

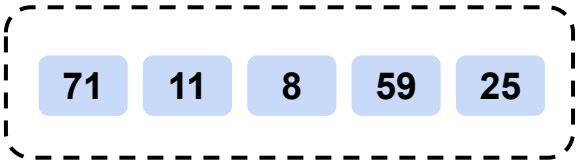
- Task: Next-token prediction
- Example: GSLM

Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

Textless Speech LM



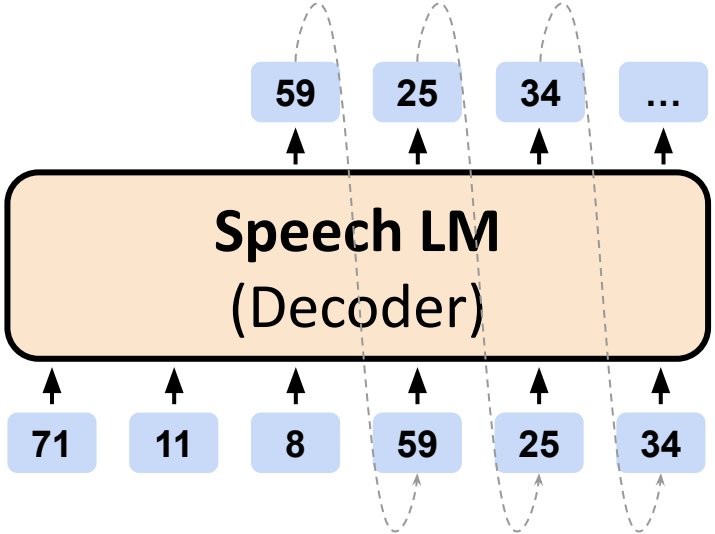
Speech tokens
(Discrete Units)



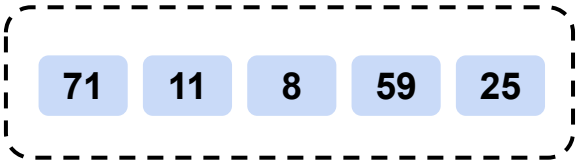
- Task: Next-token prediction
- Example: GSLM

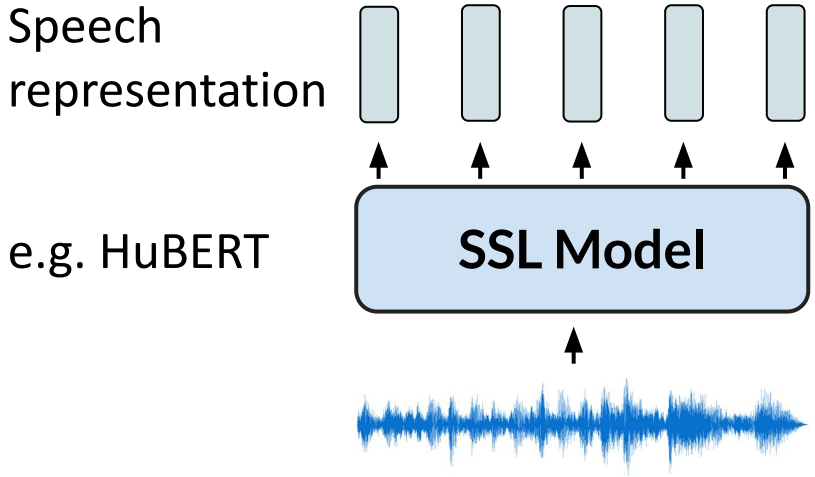
Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

Textless Speech LM



Speech tokens
(Discrete Units)

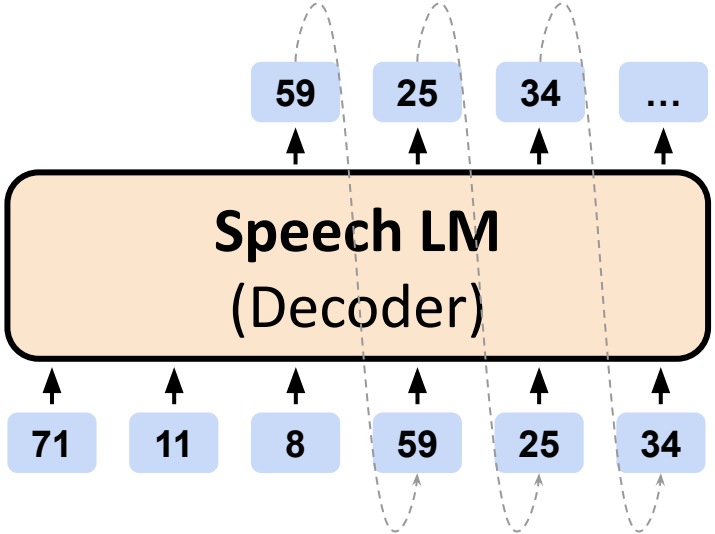


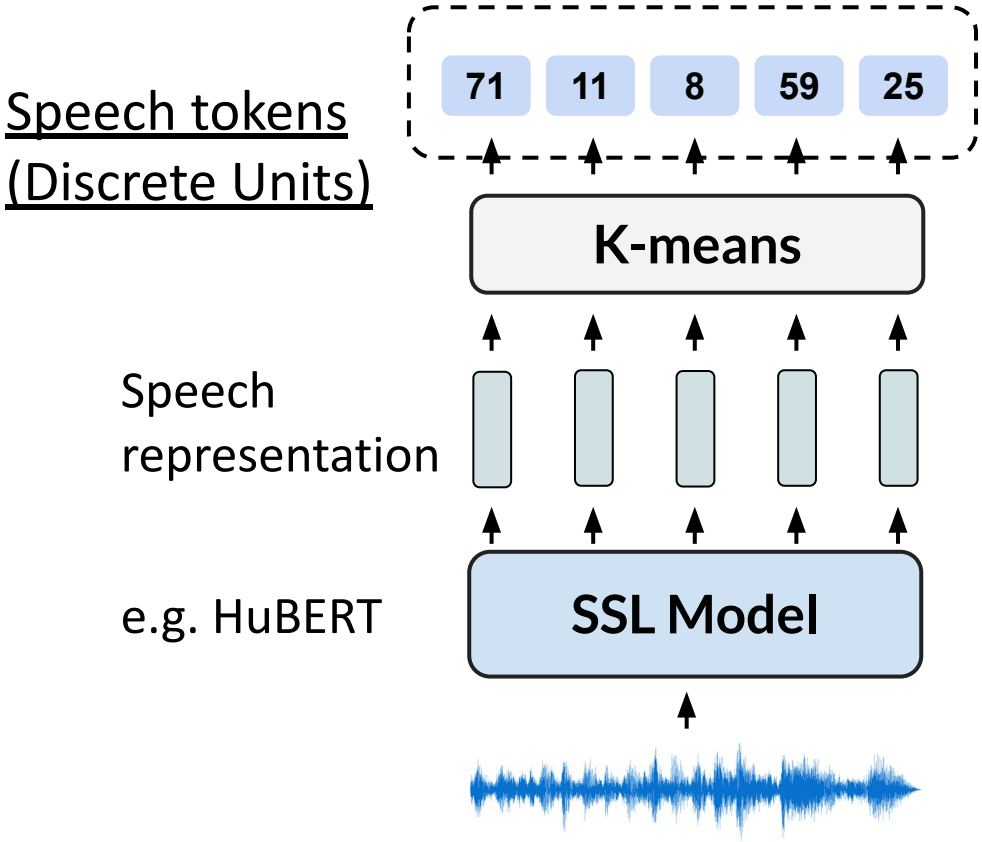


- Task: Next-token prediction
- Example: GSLM

Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

Textless Speech LM

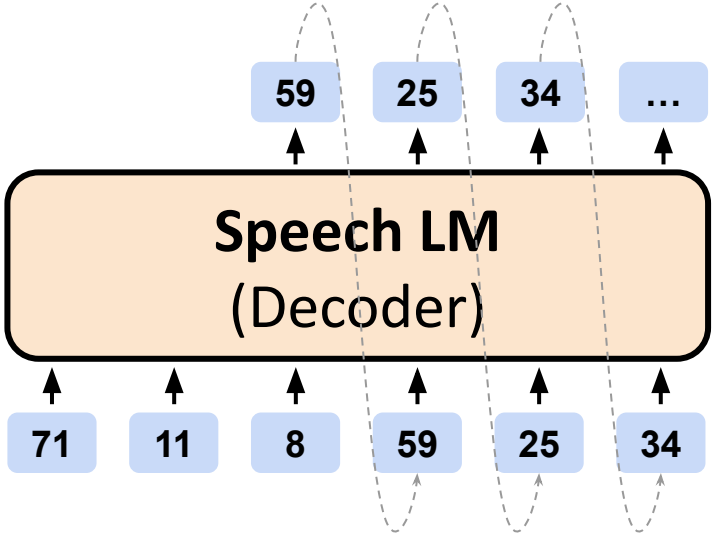




- Task: Next-token prediction
- Example: GSLM

Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

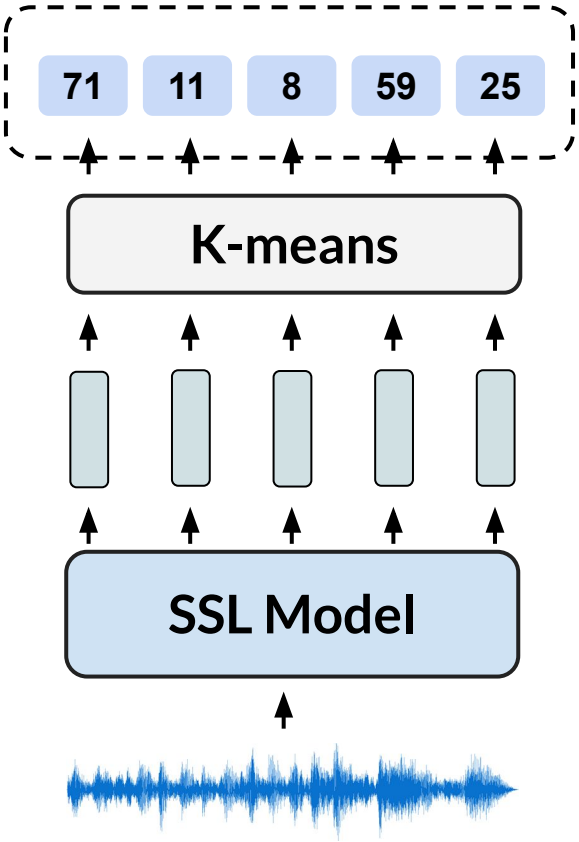
Textless Speech LM



- Task: Next-token prediction
- Example: GSLM

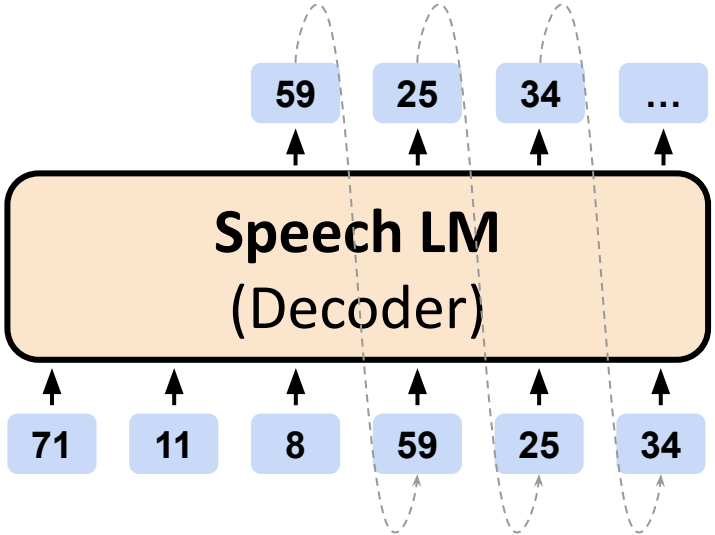
Speech tokens
(Discrete Units)

- Phonetic
- Semantic



Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

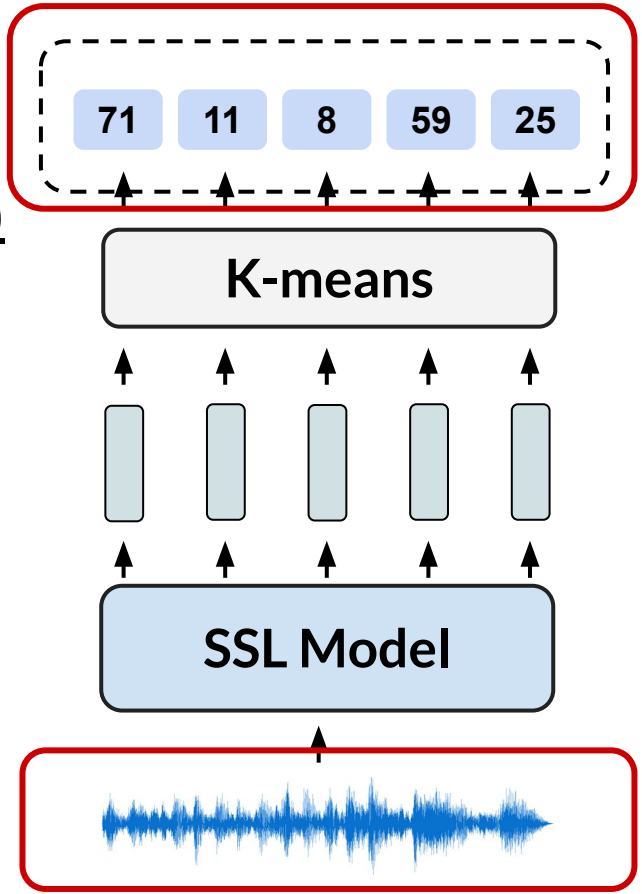
Textless Speech LM



- Task: Next-token prediction
- Example: GSLM

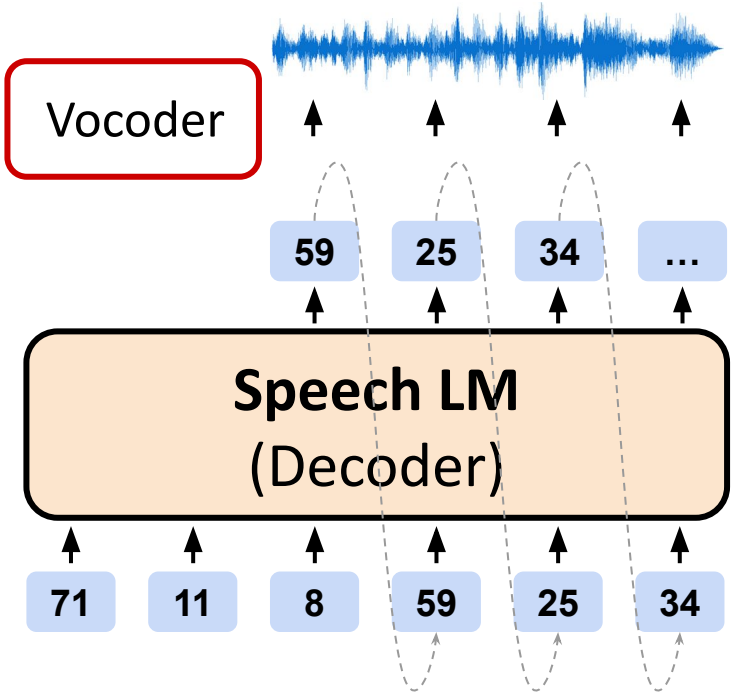
Speech tokens
(Discrete Units)

- Phonetic
- Semantic



Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

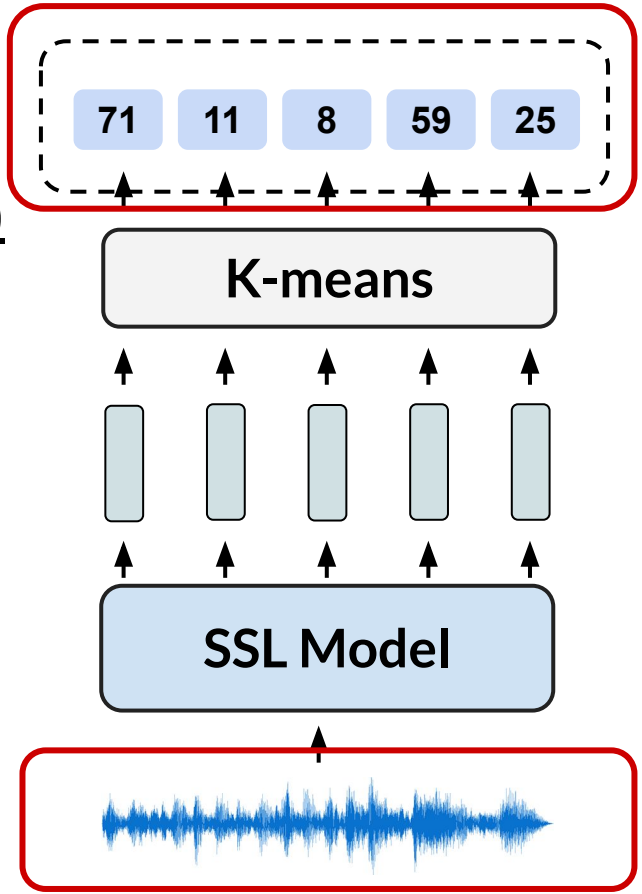
Textless Speech LM



- Task: Next-token prediction
- Example: GSLM

Speech tokens
(Discrete Units)

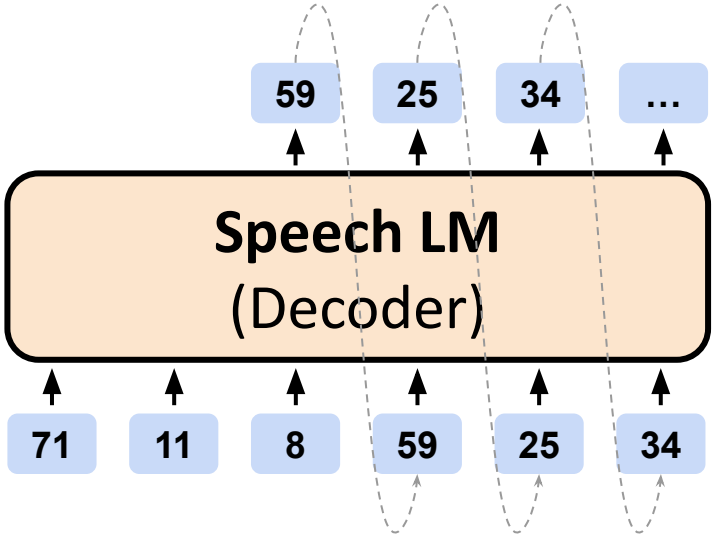
- Phonetic
- Semantic



Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

Textless Speech LM

Decoder-only Speech LM

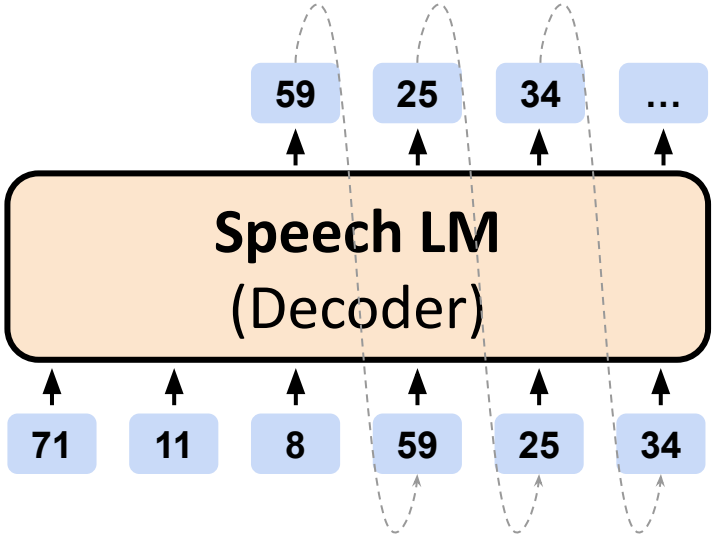


- Task: Next-token prediction
- Example: GSLM

Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

Textless Speech LM

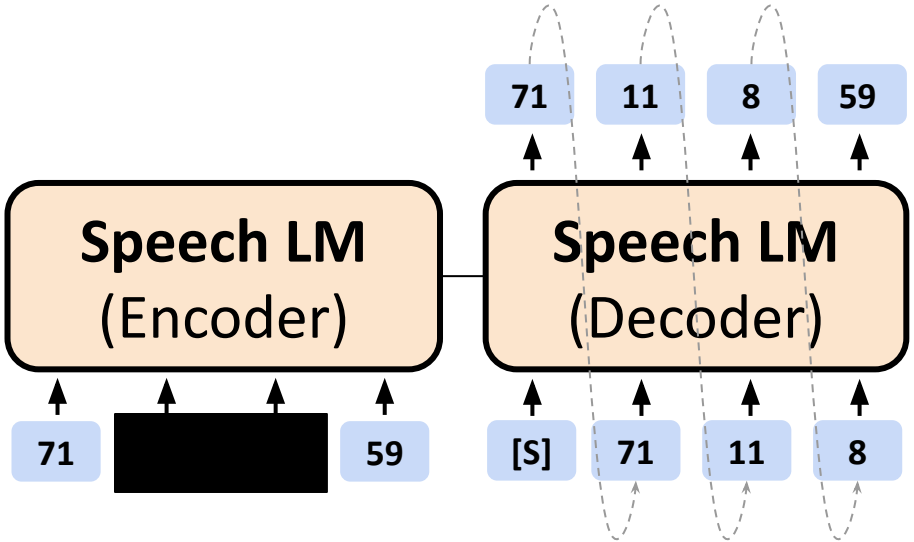
Decoder-only Speech LM



- Task: Next-token prediction
- Example: GSLM

Generative Spoken Language Modeling from Raw Audio (https://arxiv.org/abs/2102.01192)

Encoder-Decoder Speech LM



- Task: Reconstruction
- Example: Unit mBART

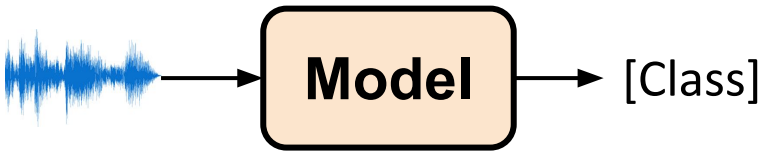
Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation (https://arxiv.org/abs/2204.02967)

3 kinds of speech processing tasks that take speech as input

3 kinds of speech processing tasks that take speech as input

1. Speech Classification

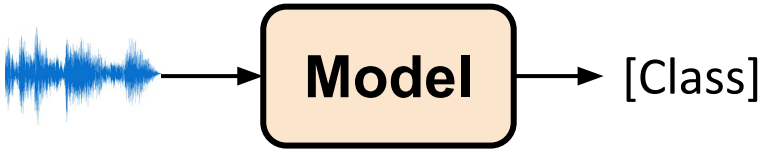
- ➔ Speech to class
- ➔ e.g. Langauge Identification



3 kinds of speech processing tasks that take speech as input

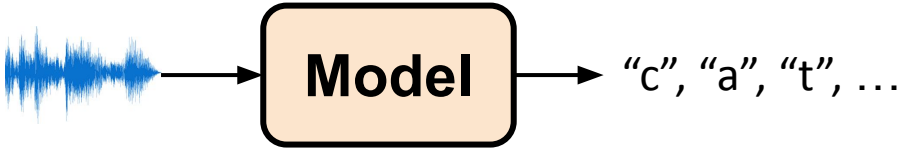
1. Speech Classification

- ➔ Speech to class
- ➔ e.g. Langauge Identification



2. Sequence Generation

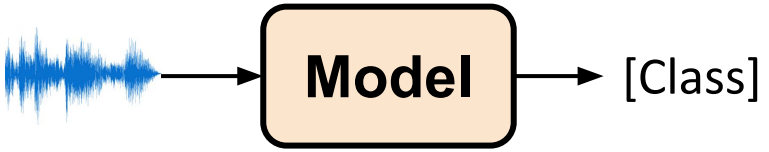
- ➔ Speech to label sequence
- ➔ e.g. ASR



3 kinds of speech processing tasks that take speech as input

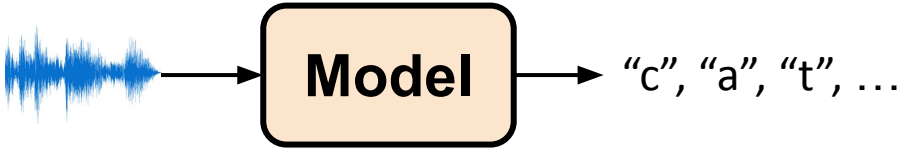
1. Speech Classification

- ➔ Speech to class
- ➔ e.g. Langauge Identification



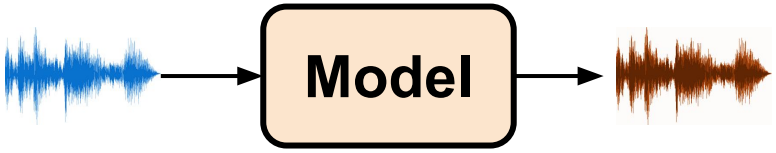
2. Sequence Generation

- ➔ Speech to label sequence
- ➔ e.g. ASR



3. Speech Generation

- ➔ Speech to speech
- ➔ e.g. Speech translation



Prompting Speech LM

Prompting Speech LM

Sequence Generation Tasks

(e.g. ASR)

Speech LM

(cat)

Prompting Speech LM

downstream task's label

Sequence Generation Tasks

(e.g. ASR)

c a t

Speech LM

(cat)

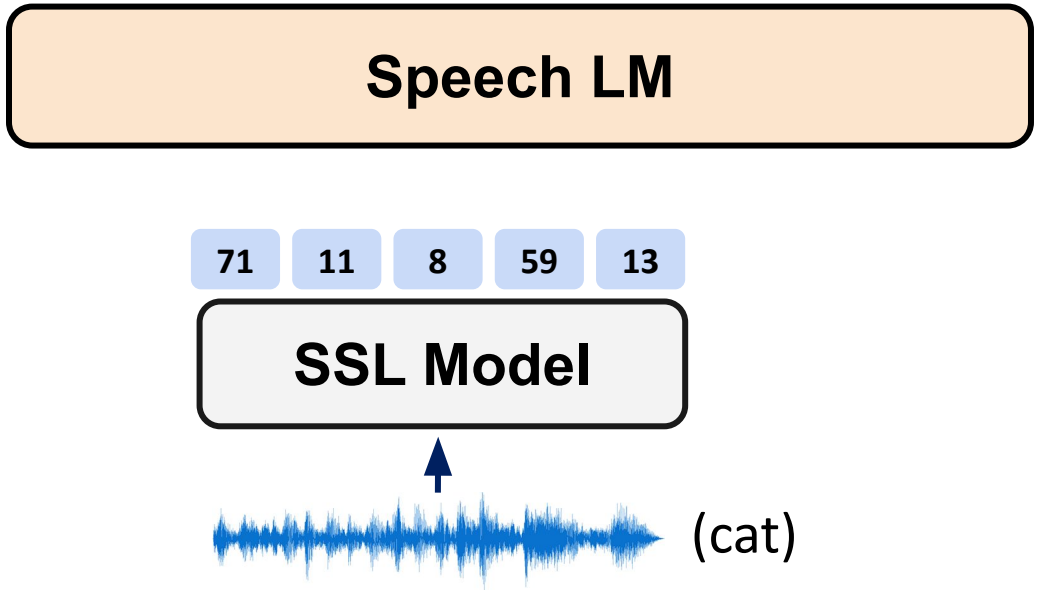
Prompting Speech LM

downstream task's label

Sequence Generation Tasks

(e.g. ASR)

c a t



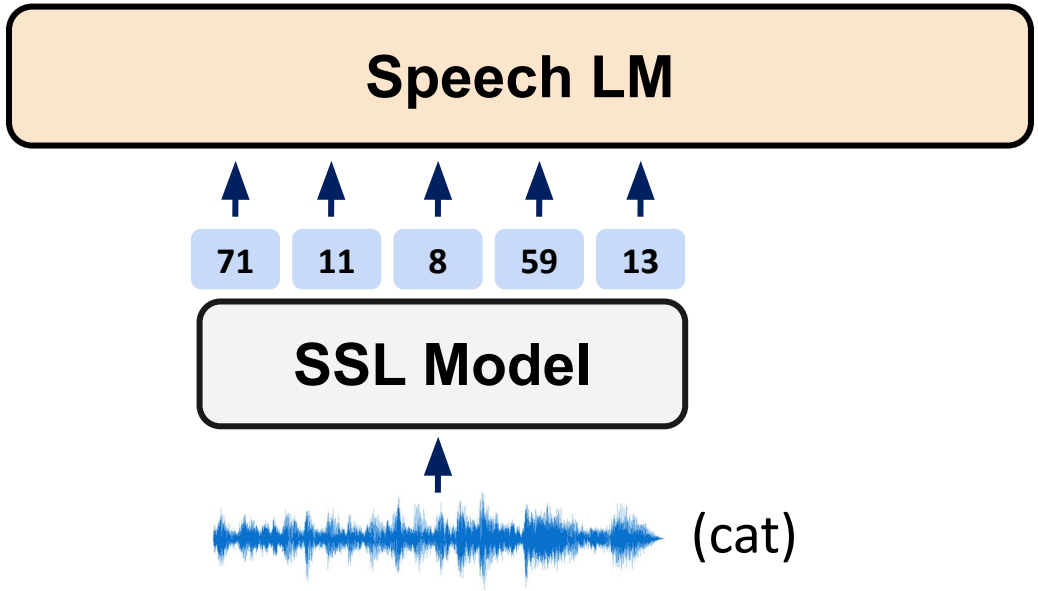
Prompting Speech LM

downstream task's label

Sequence Generation Tasks

(e.g. ASR)

c a t



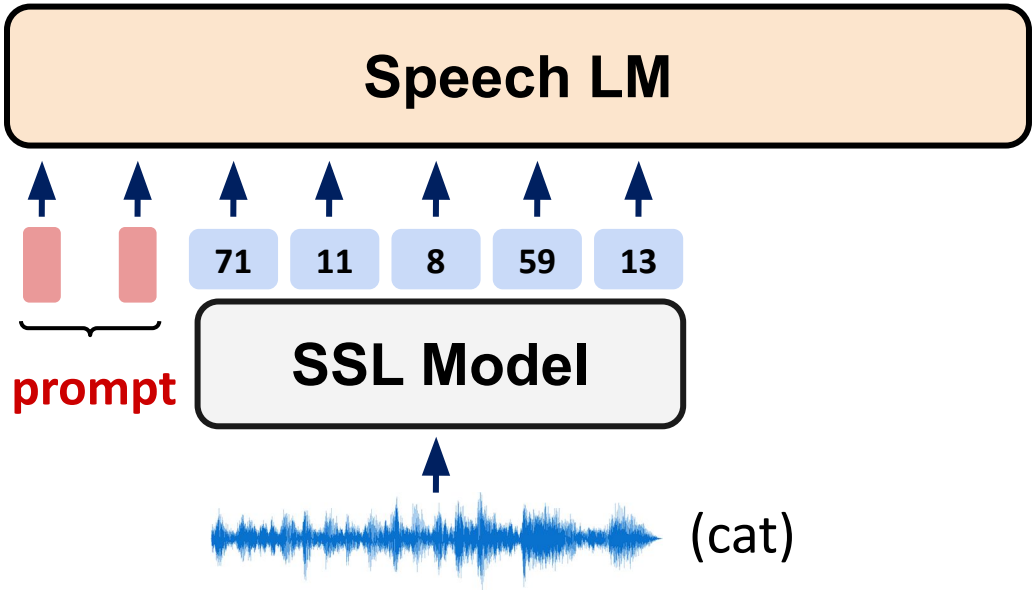
Prompting Speech LM

downstream task's label

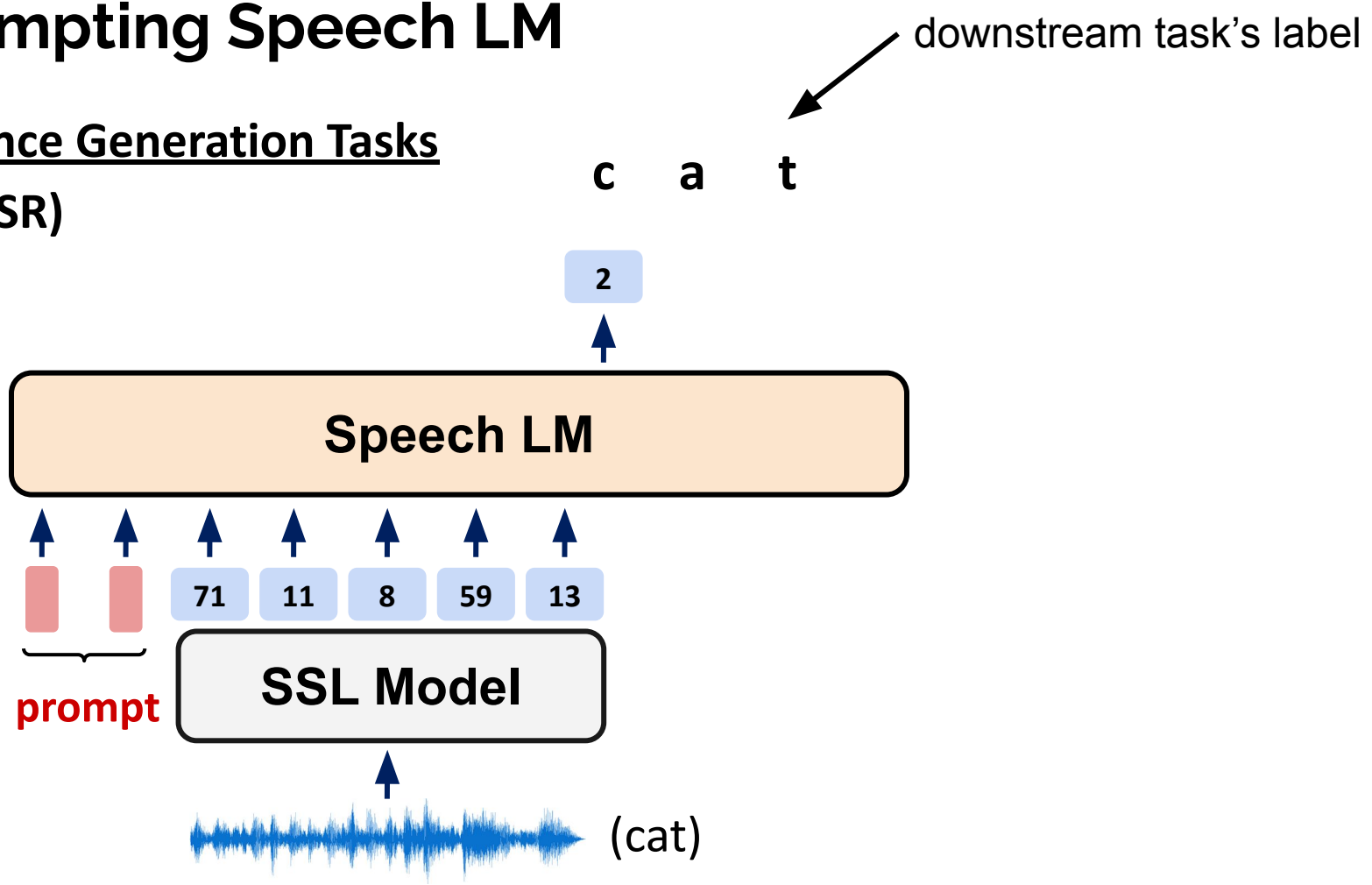
Sequence Generation Tasks

(e.g. ASR)

c a t



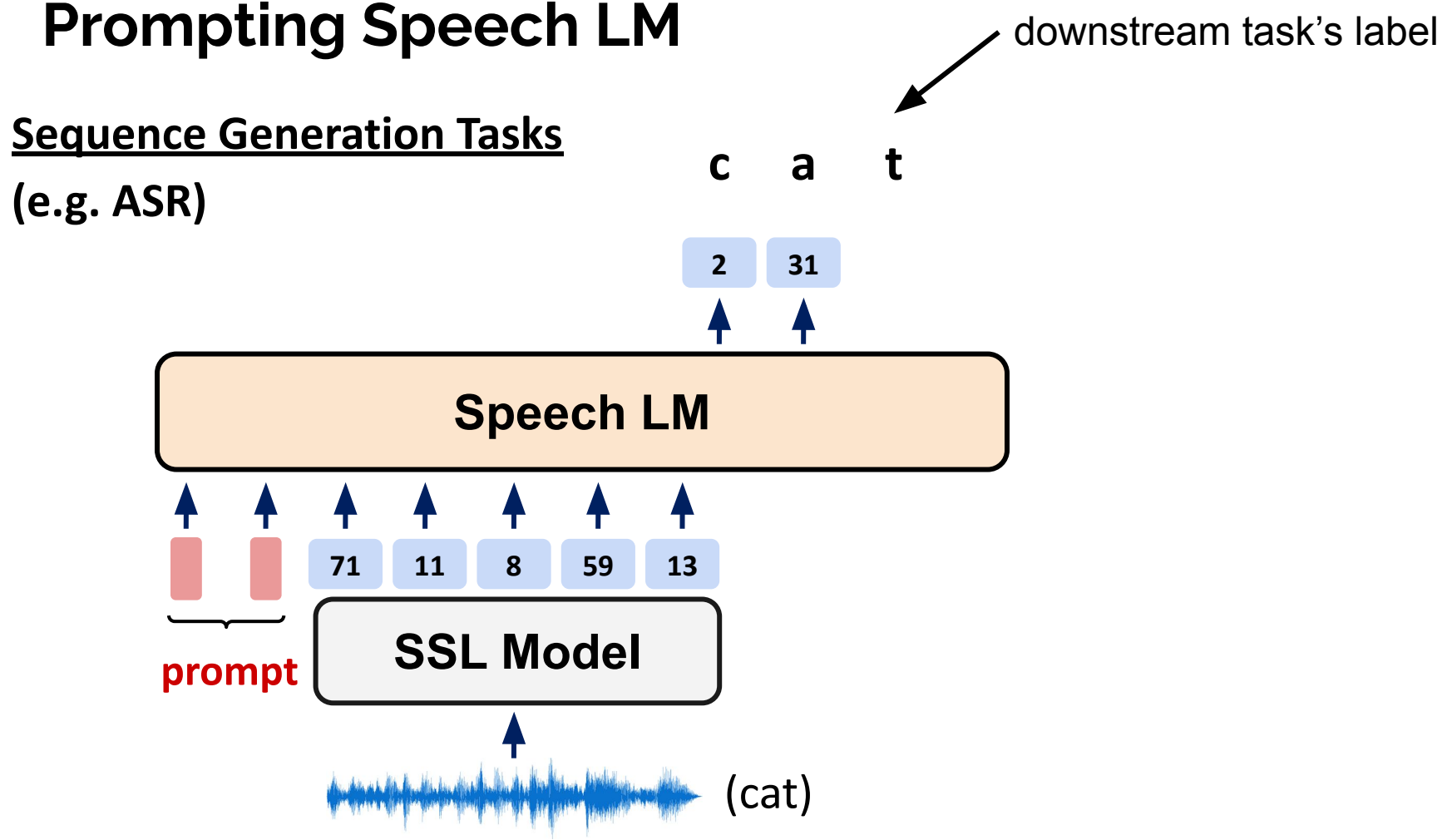
Prompting Speech LM



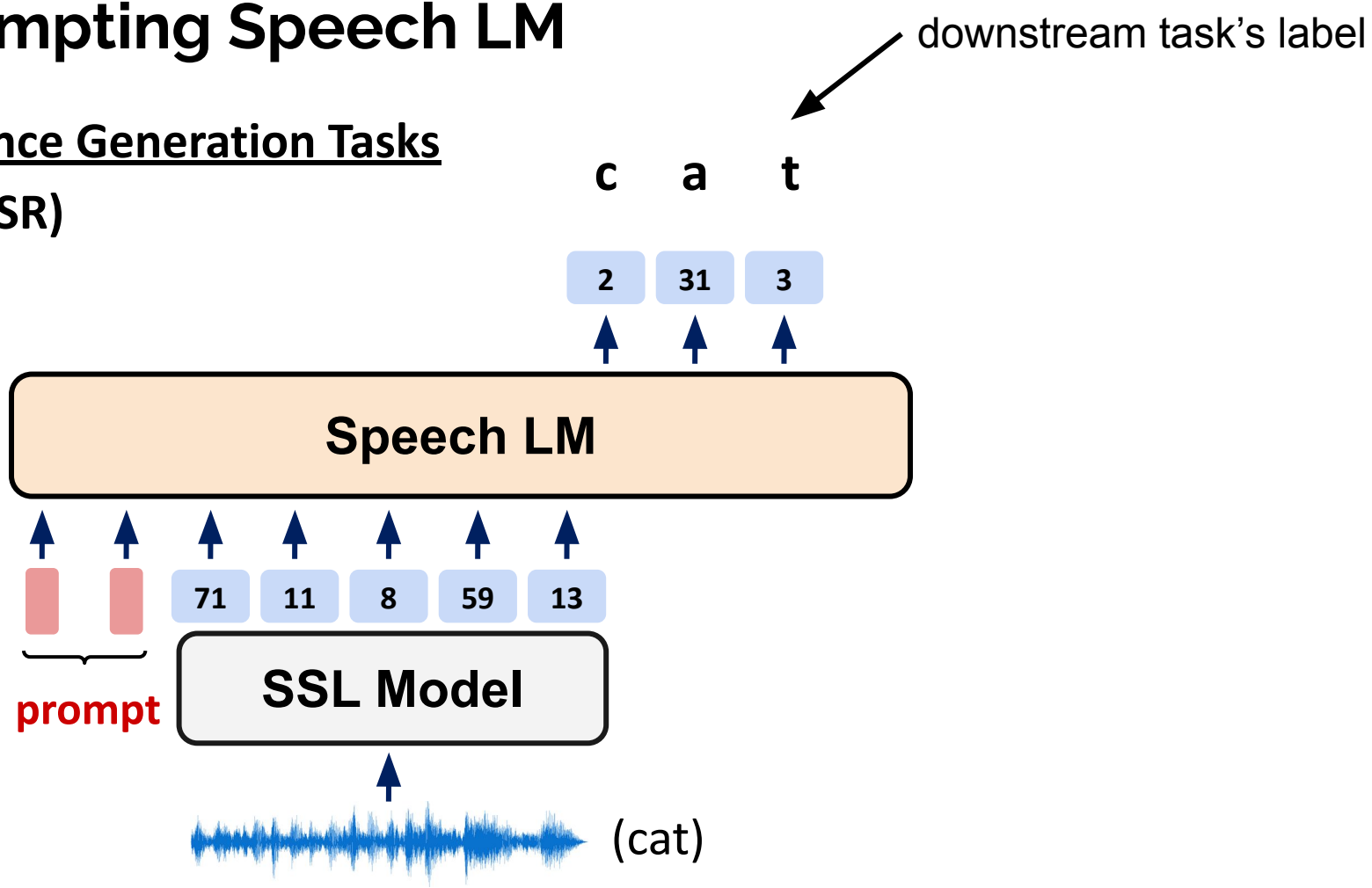
Sequence Generation Tasks

(e.g. ASR)

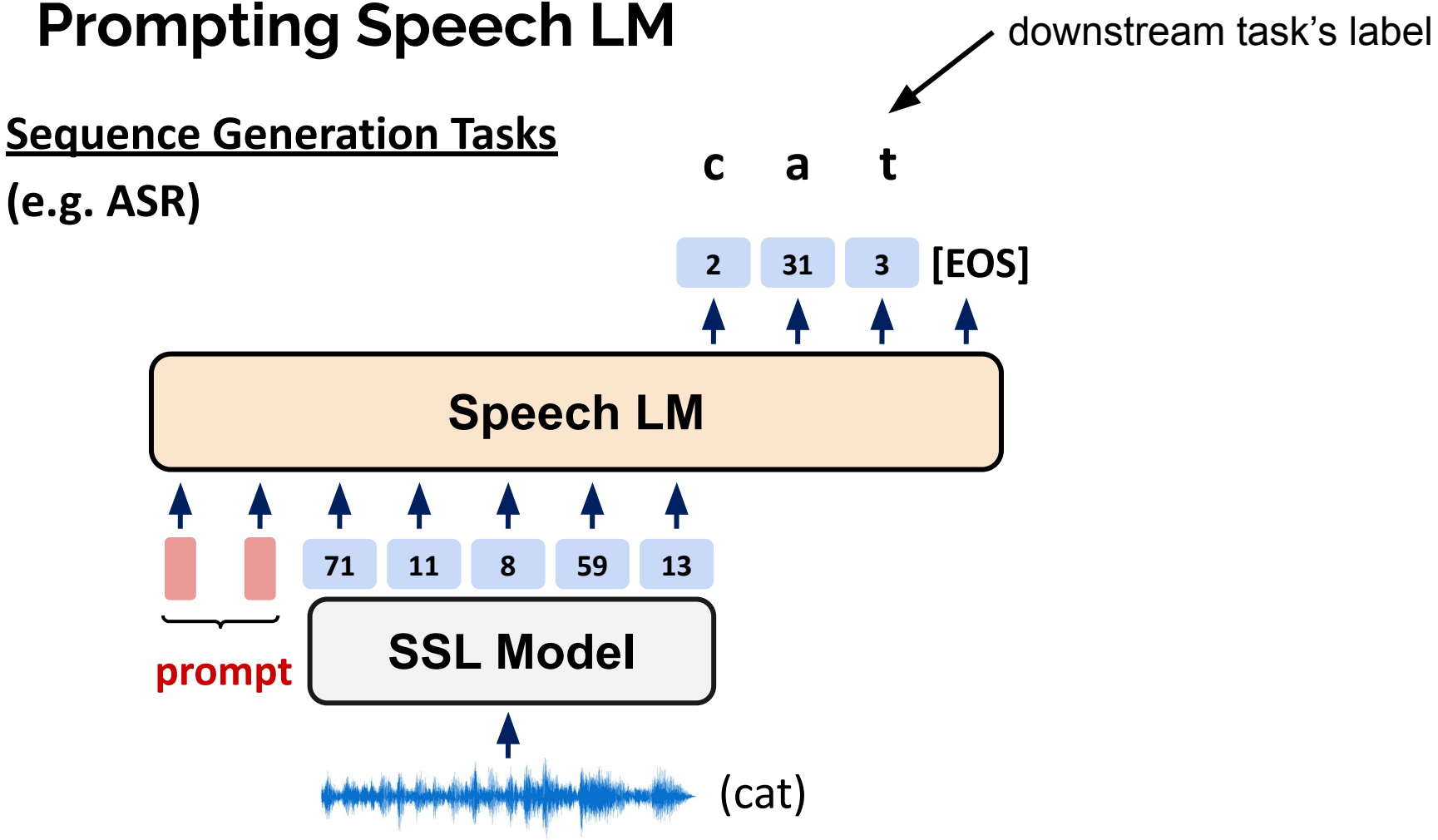
Prompting Speech LM



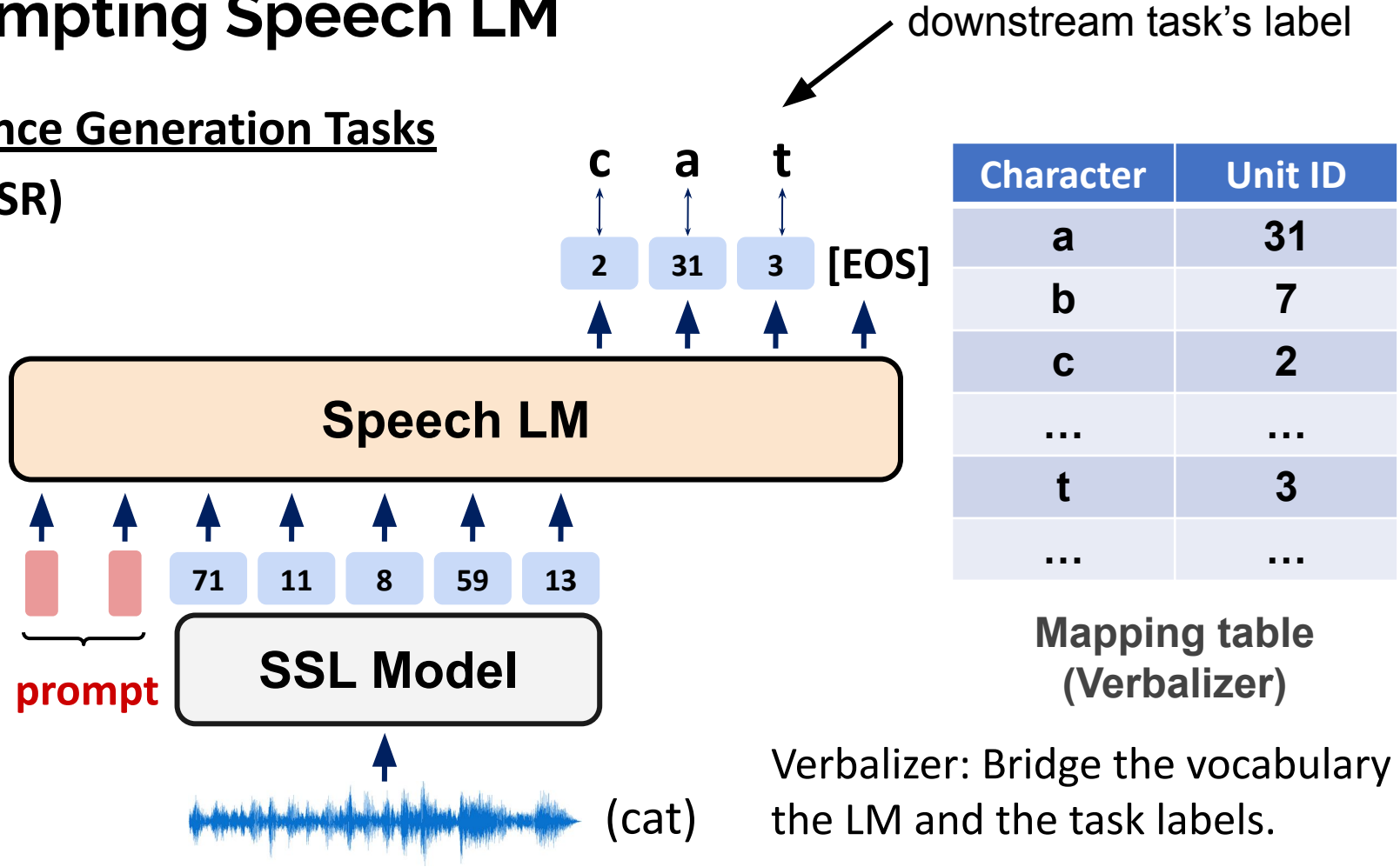
Prompting Speech LM



Prompting Speech LM



Prompting Speech LM

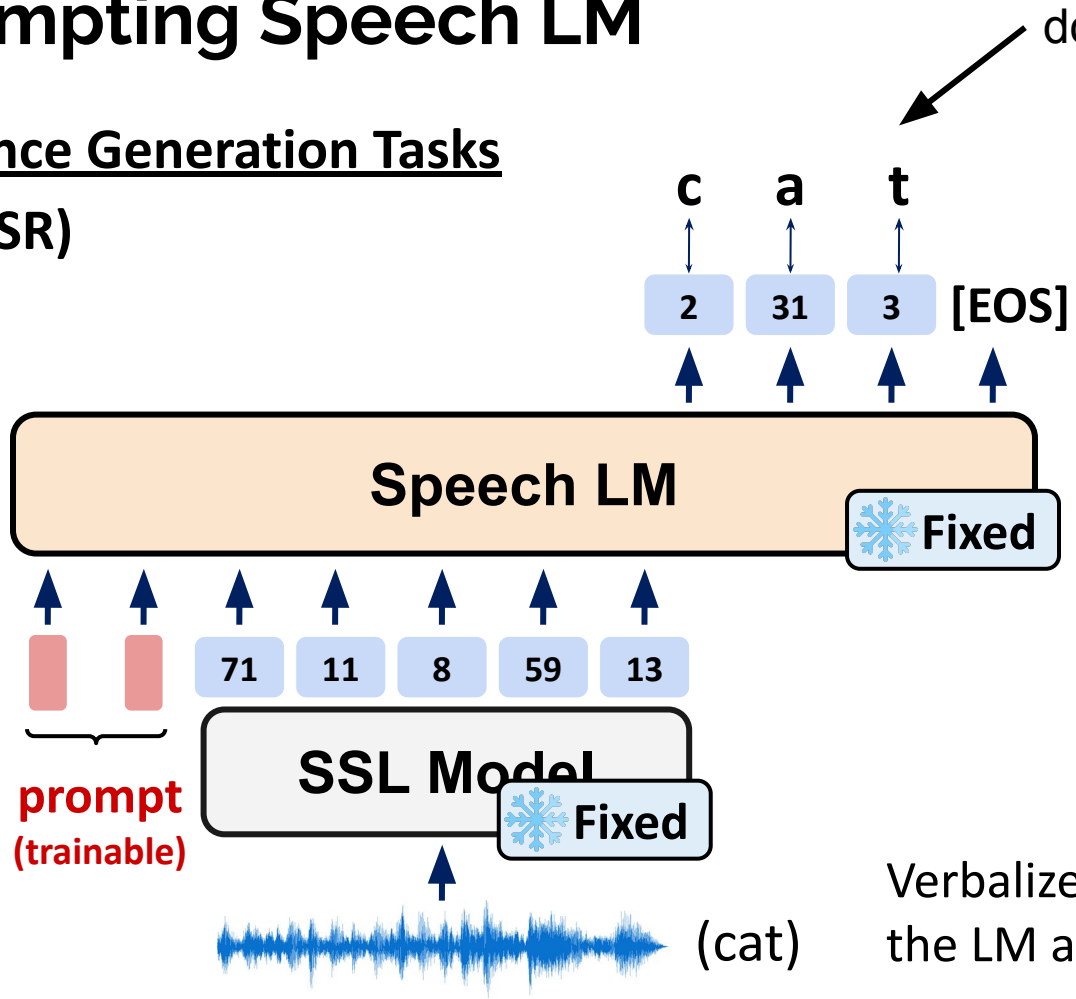


Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)

Verbalizer: Bridge the vocabulary of the LM and the task labels.

Prompting Speech LM

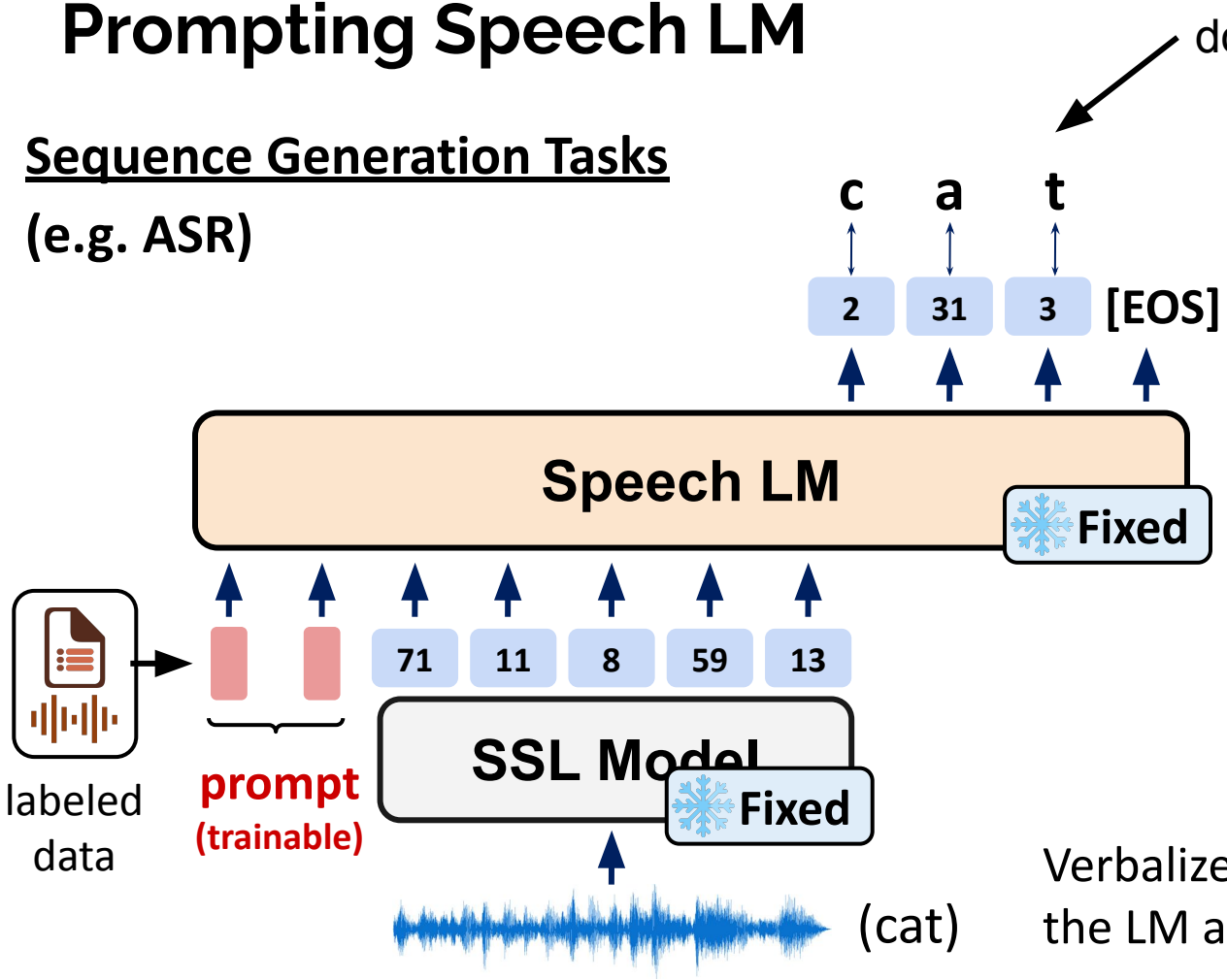


Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)

Verbalizer: Bridge the vocabulary of the LM and the task labels.

Prompting Speech LM



Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)

Verbalizer: Bridge the vocabulary of the LM and the task labels.

Prompting Speech LM

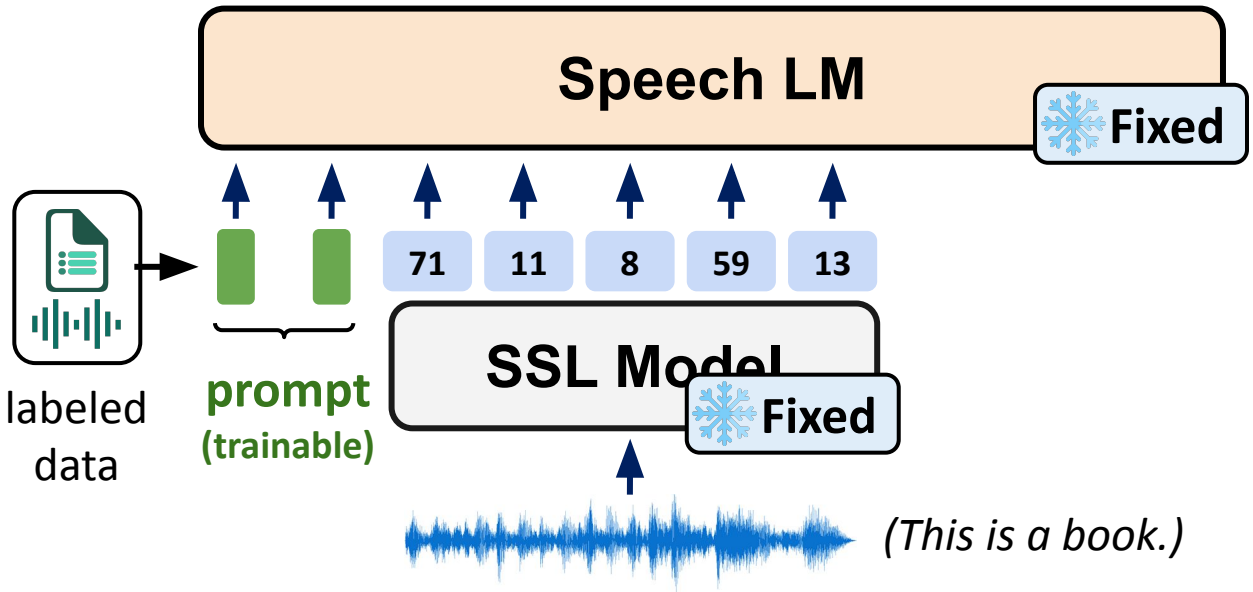
downstream task's label

Speech Classification

(e.g. Language Identification)

Labels	Unit ID
[EN]	22
[ES]	29
[CN]	17
…	…
[LT]	3
…	…

Mapping table
(Verbalizer)

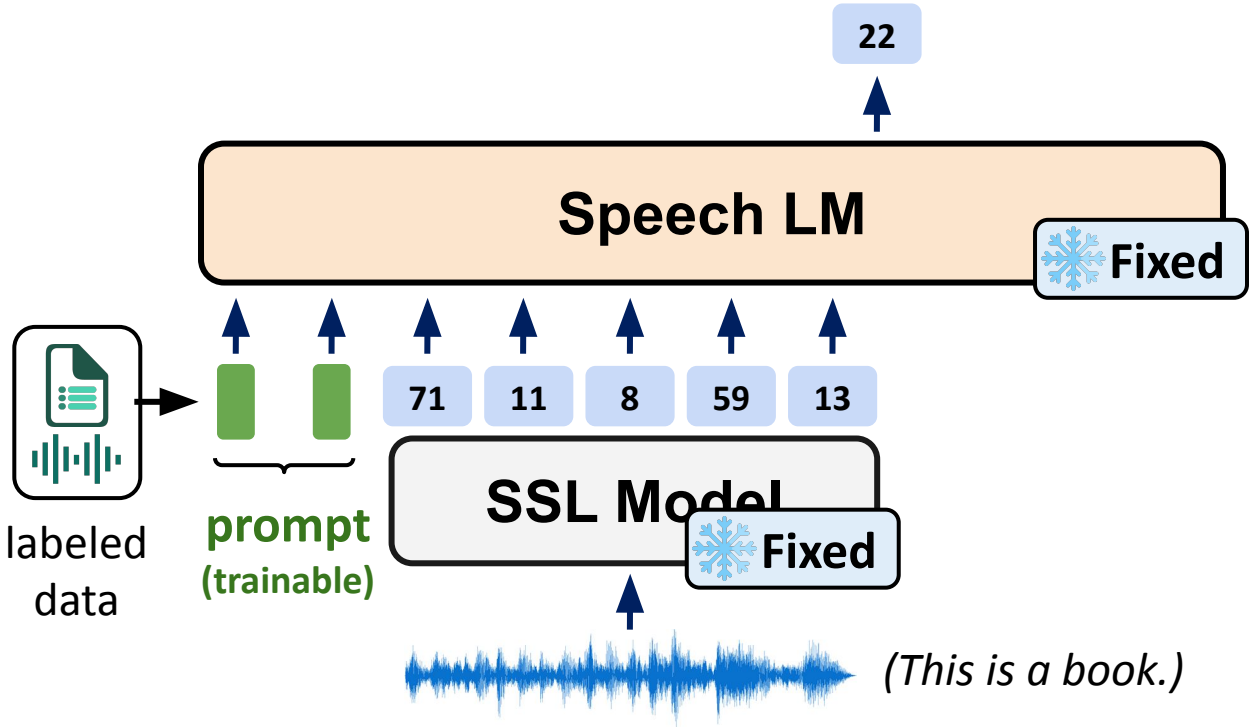


Prompting Speech LM

downstream task's label

Speech Classification

(e.g. Language Identification)



Labels	Unit ID
[EN]	22
[ES]	29
[CN]	17
…	…
[LT]	3
…	…

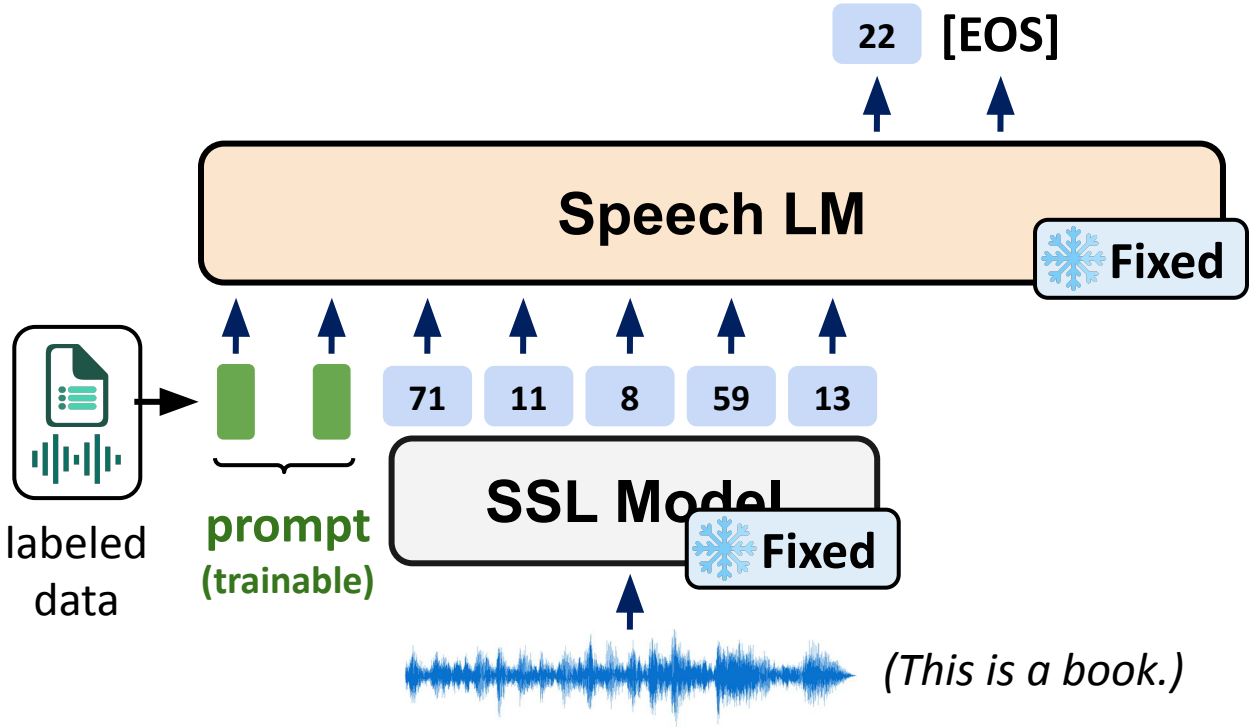
Mapping table
(Verbalizer)

Prompting Speech LM

downstream task's label

Speech Classification

(e.g. Language Identification)



Labels	Unit ID
[EN]	22
[ES]	29
[CN]	17
…	…
[LT]	3
…	…

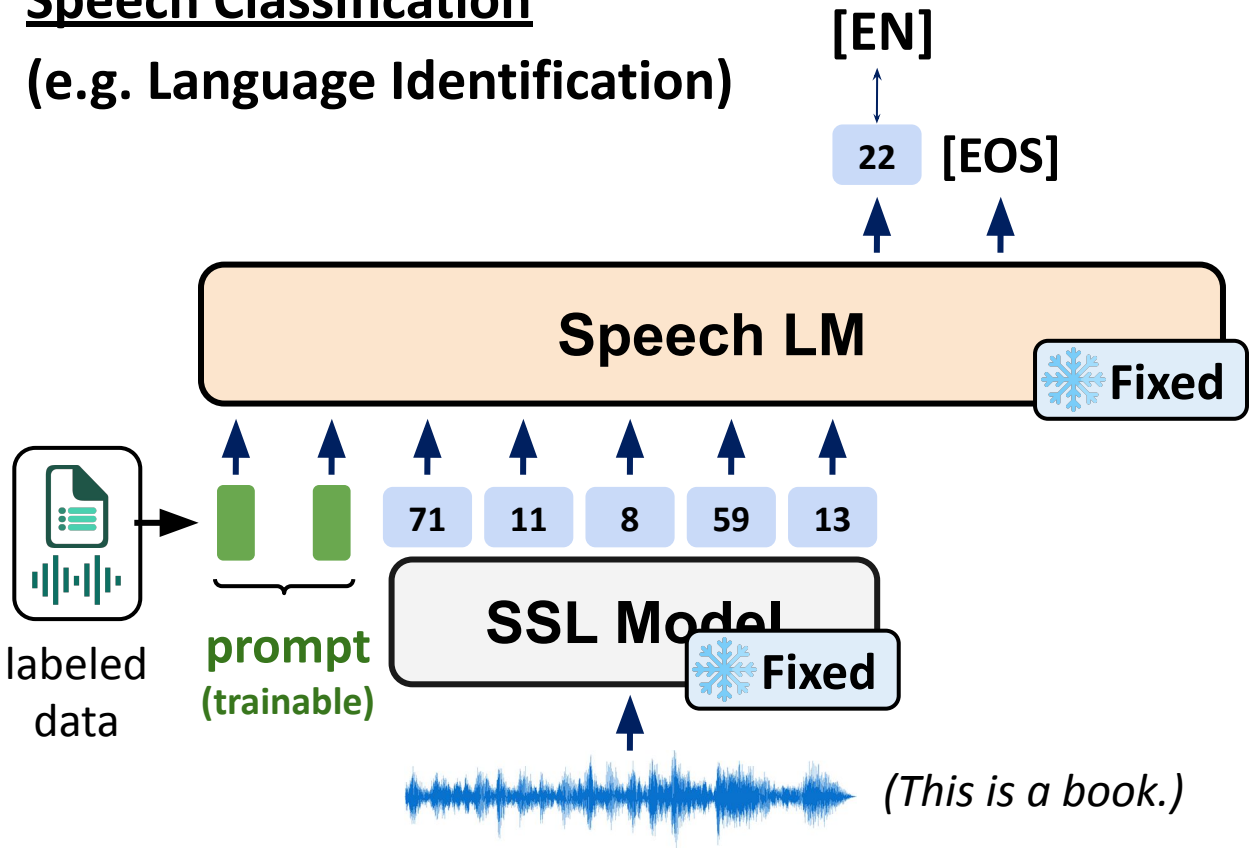
Mapping table (Verbalizer)

Prompting Speech LM

downstream task's label

Speech Classification

(e.g. Language Identification)



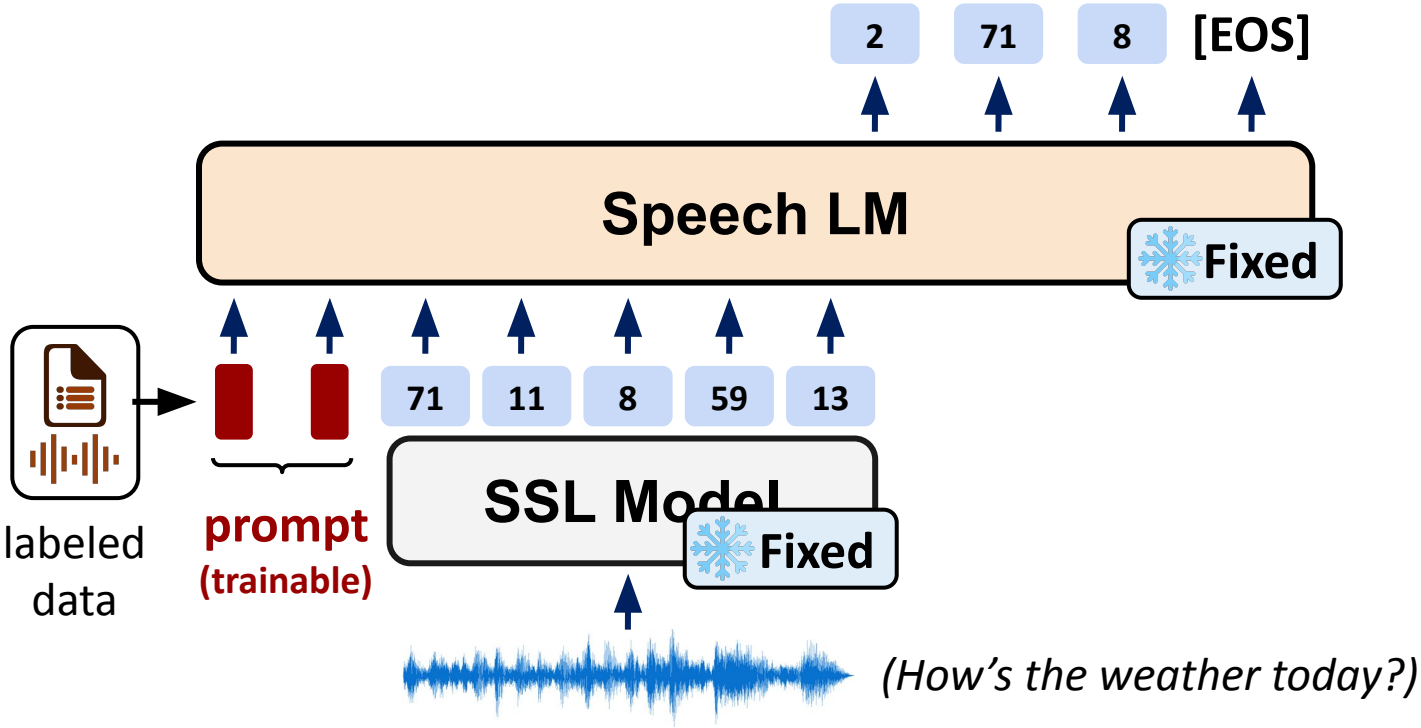
Labels	Unit ID
[EN]	22
[ES]	29
[CN]	17
…	…
[LT]	3
…	…

Mapping table
(Verbalizer)

Prompting Speech LM

Speech Generation

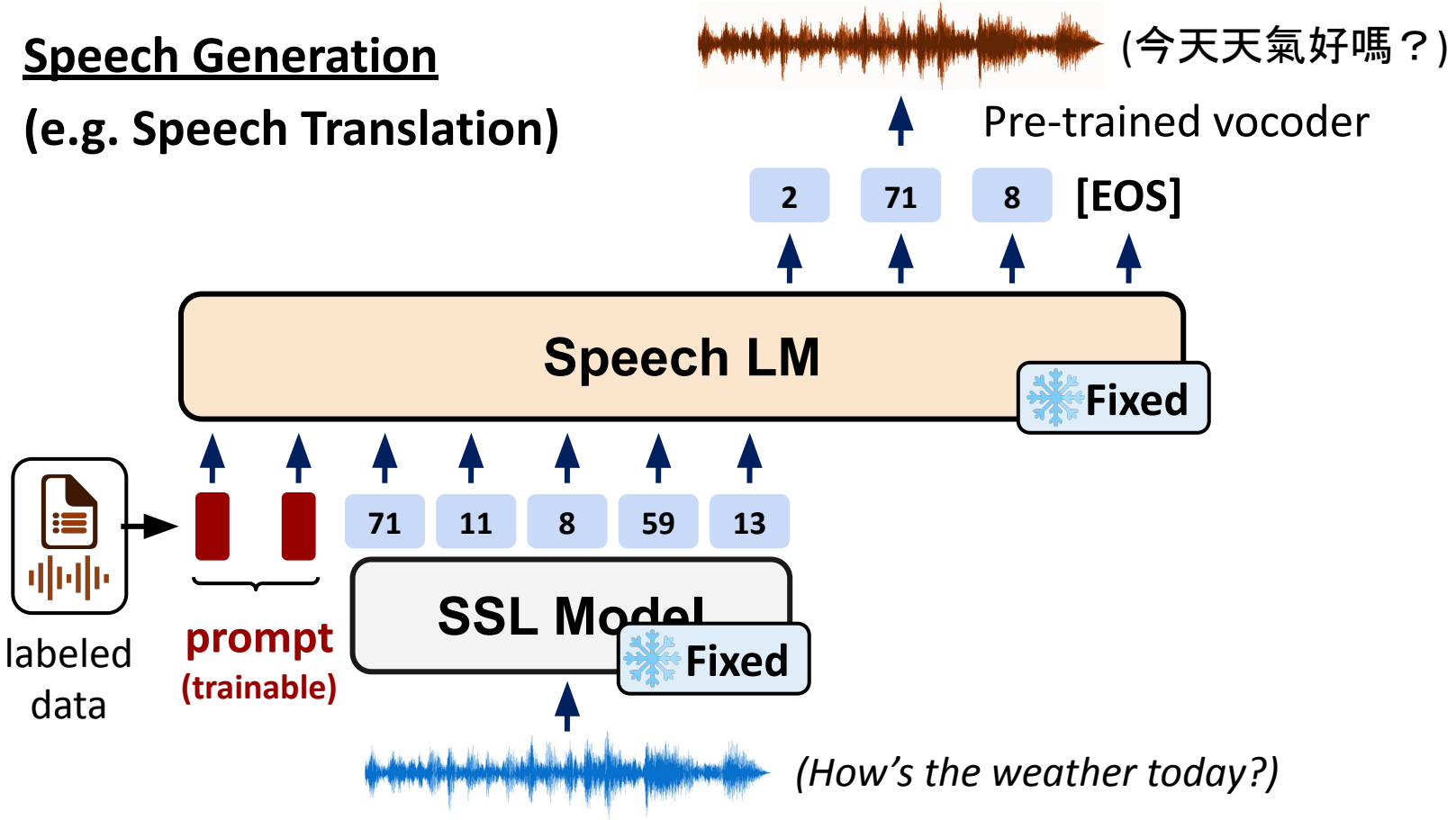
(e.g. Speech Translation)



Prompting Speech LM

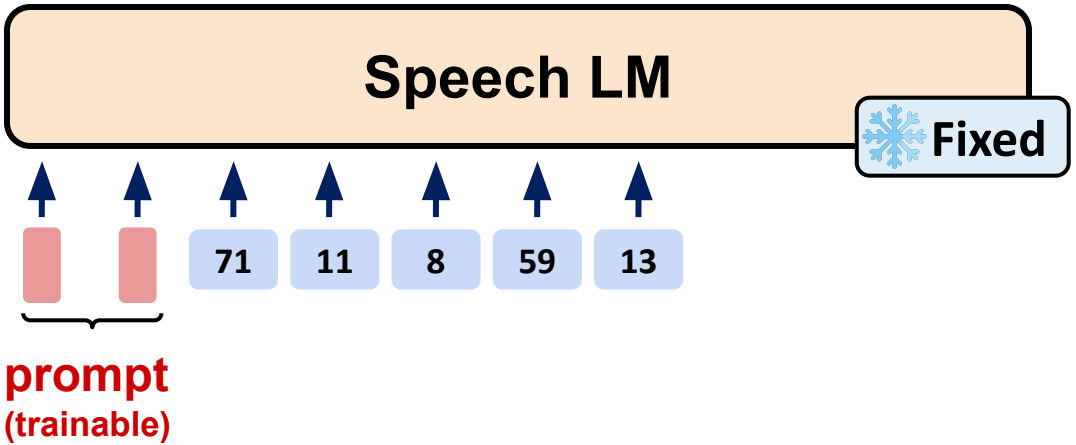
Speech Generation

(e.g. Speech Translation)



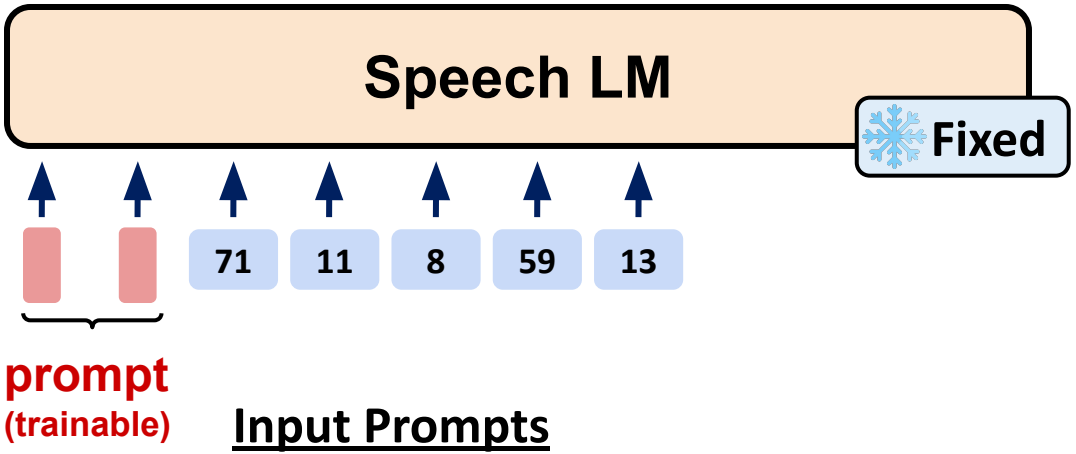
Prompting Speech LM

Prompting: Find the prompts and put them at the **input** without modifying the LM's architecture



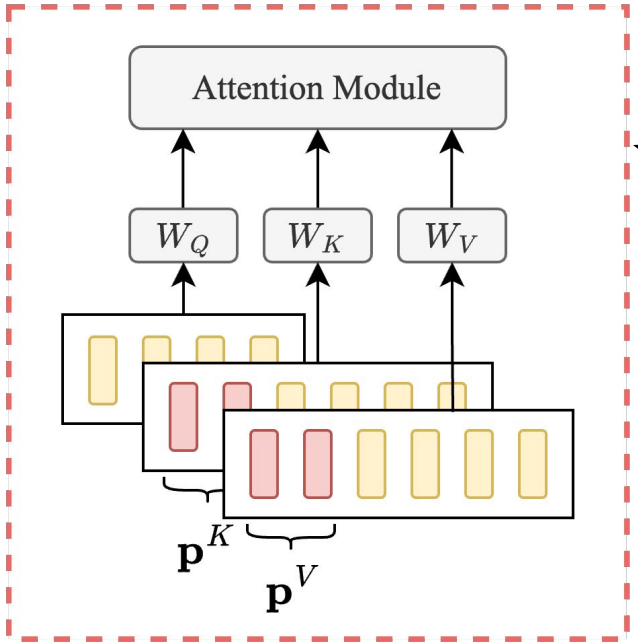
Prompting Speech LM

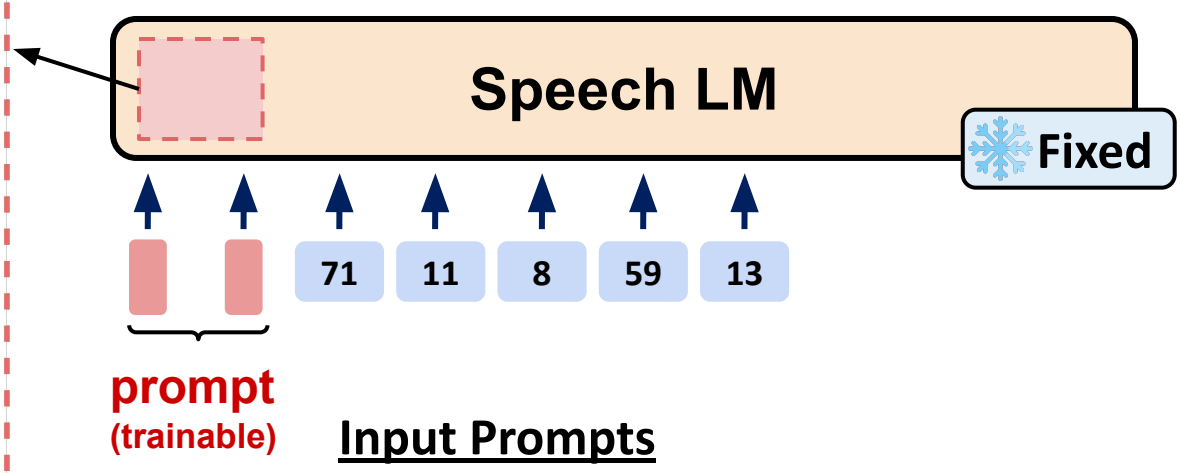
Prompting: Find the prompts and put them at the **input** without modifying the LM's architecture



Prompting Speech LM

The prompts are prepended at the **input** of each transformer layer.

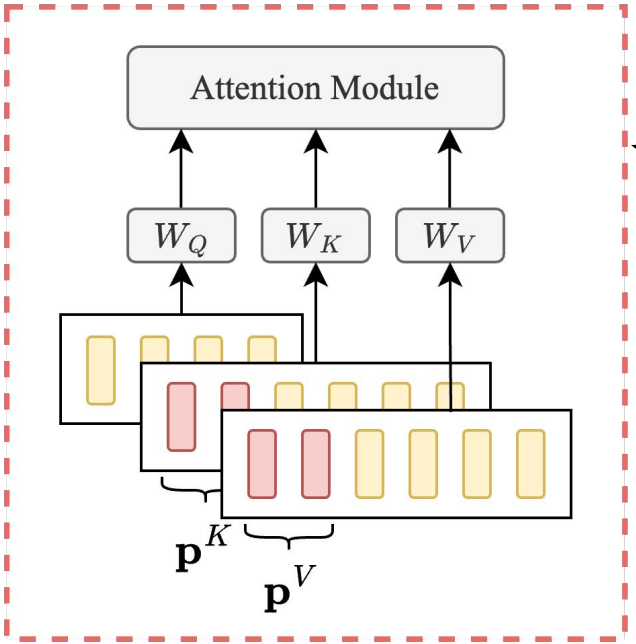




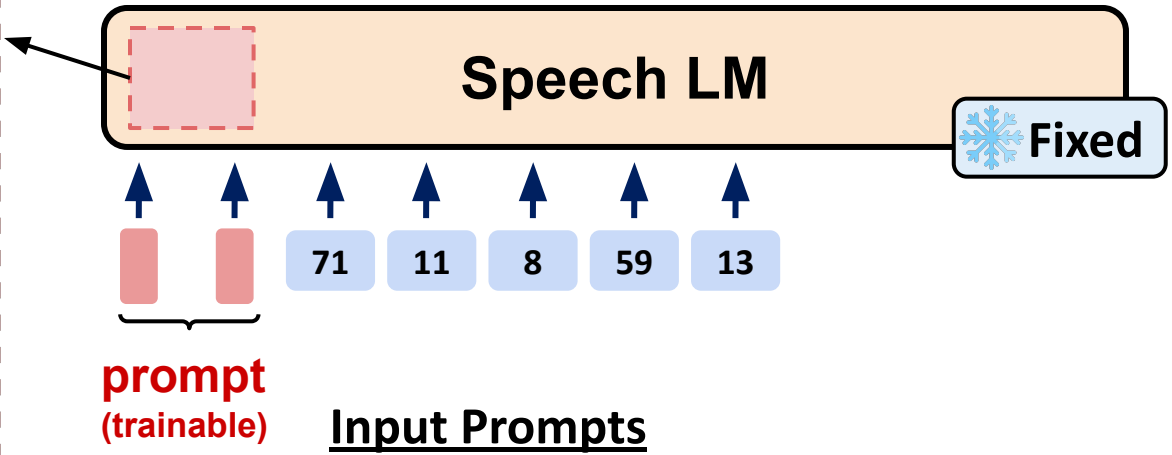
Deep Prompts guiding the attention mechnism

Prompting Speech LM

The prompts are prepended at the **input** of each transformer layer.



Deep Prompts guiding the attention mechnism

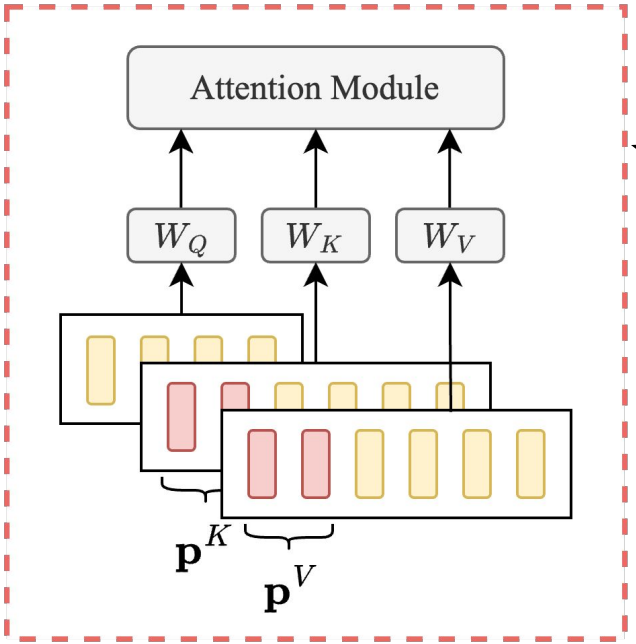


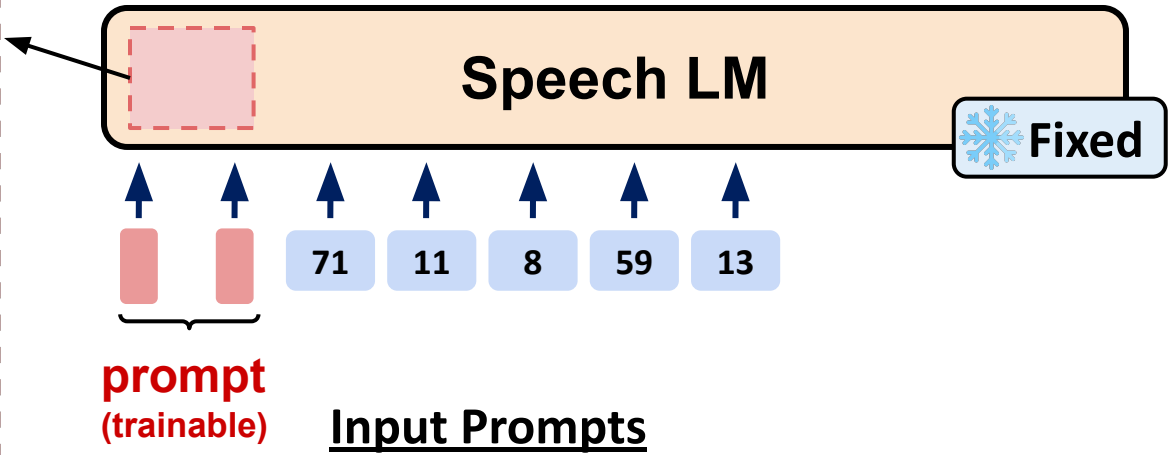
Input Prompts

The following exps. show the performance of input prompts + deep prompts

Prompting Speech LM

The prompts are prepended at the **input** of each transformer layer.



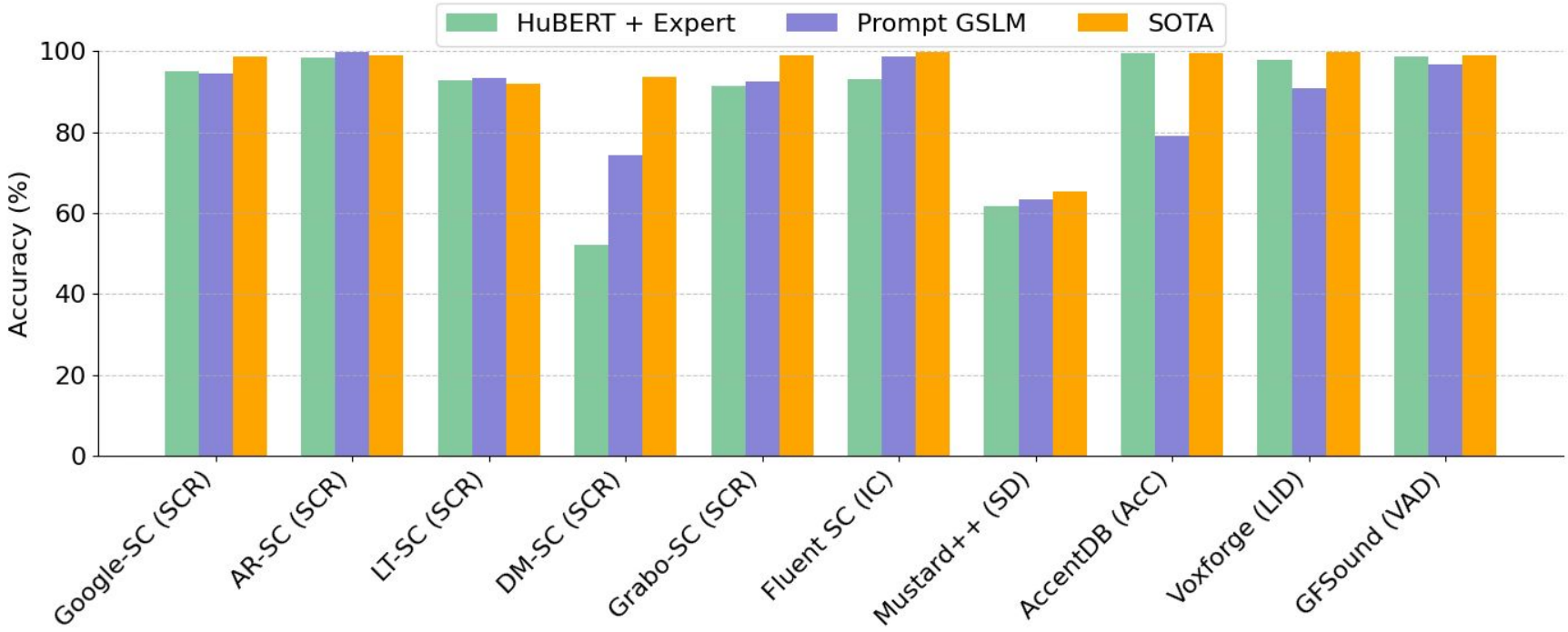


Deep Prompts guiding the attention mechnism

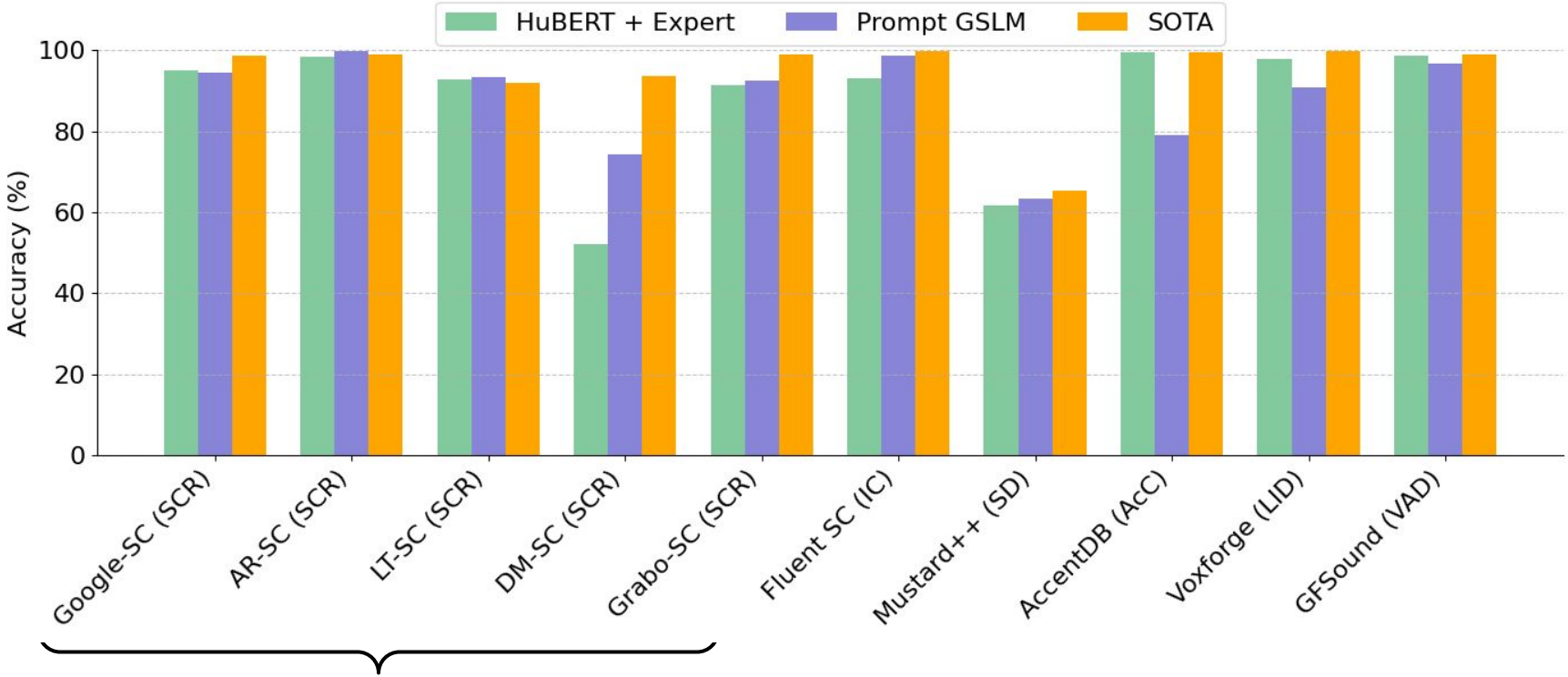
The following exps. show the performance of input prompts + deep prompts

SpeechPrompt: An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks (Interspeech 2022, https://arxiv.org/abs/2203.16773)

Speech Classification - Prompt GSLM



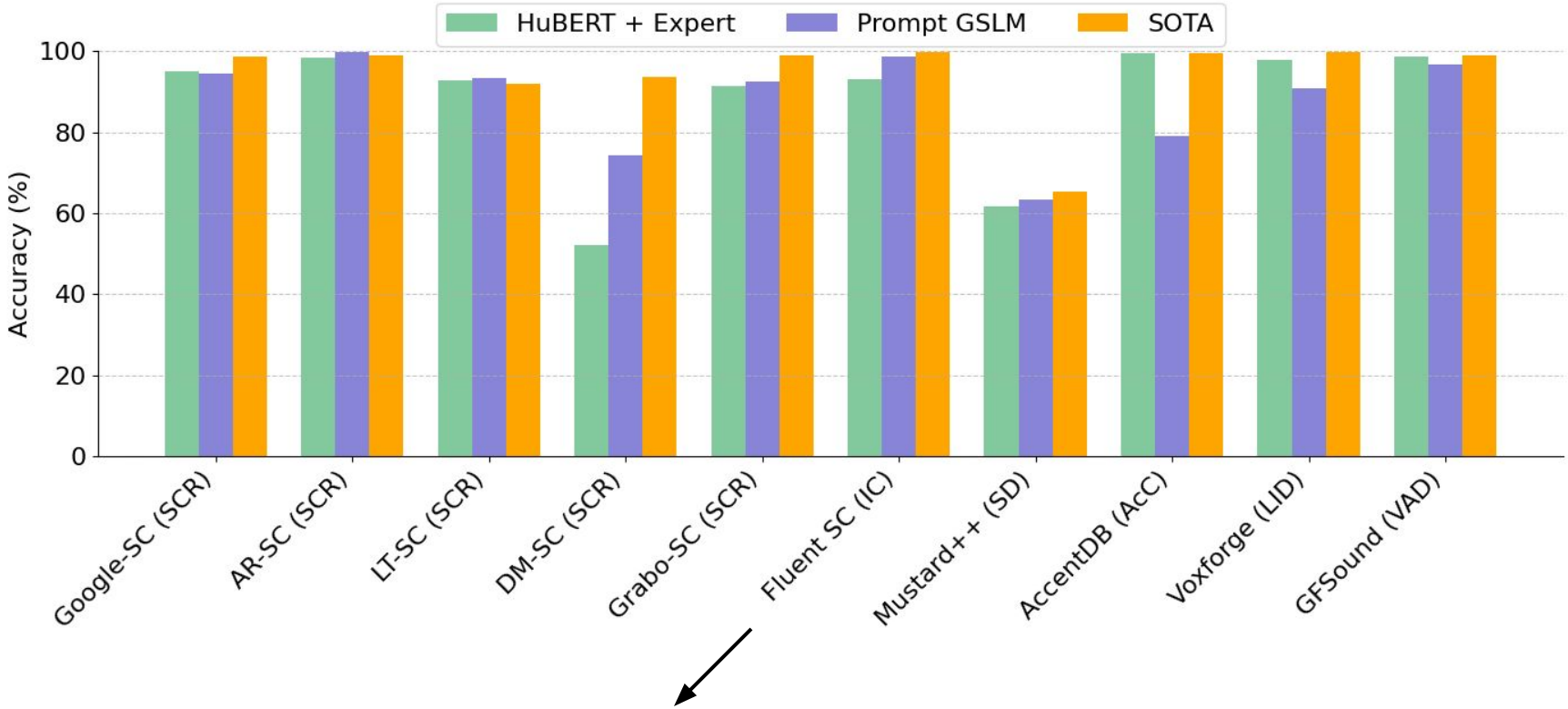
Speech Classification - Prompt GSLM



Spoken Command Recognition (SCR)
(English, Arabic, Lithuanian, Mandarin, German)

Classify an utterance into a predefined keyword set.

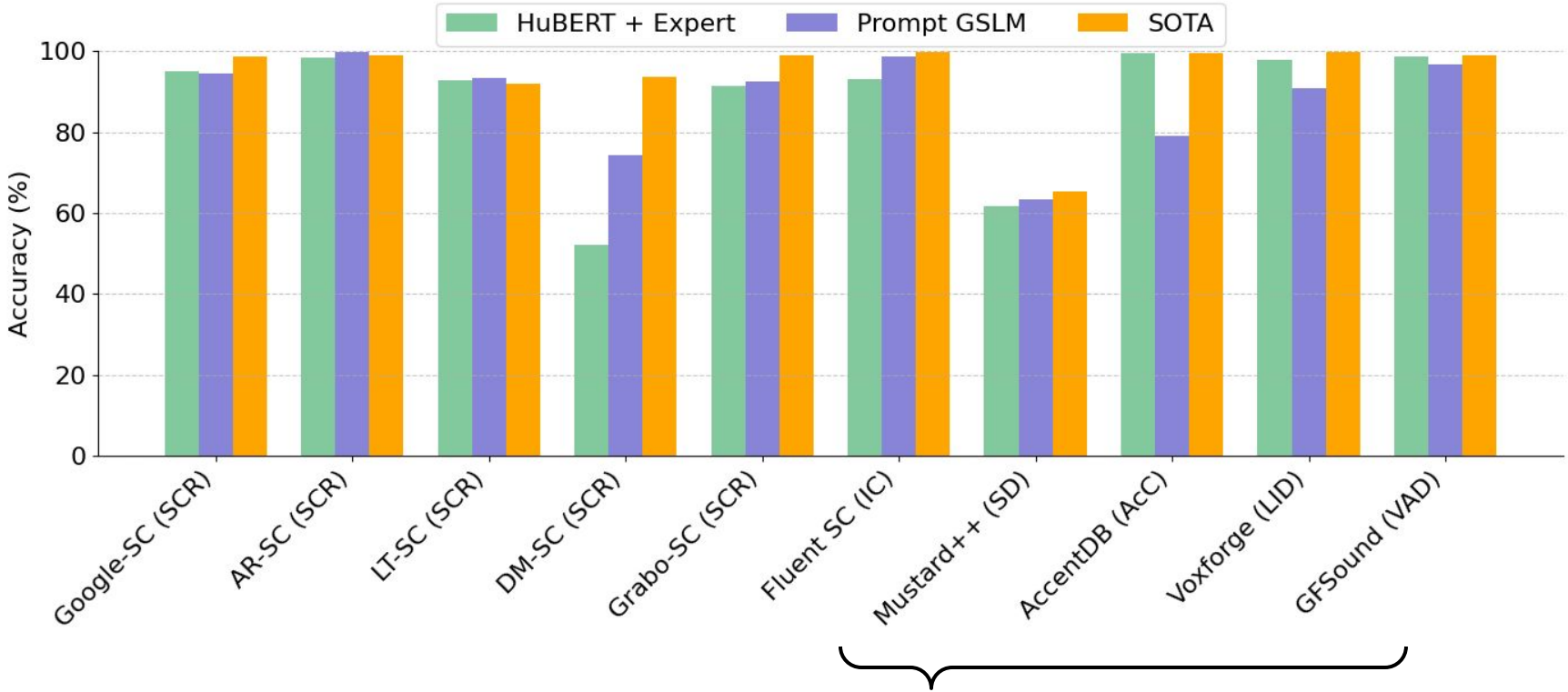
Speech Classification - Prompt GSLM



Intent Classification (IC)

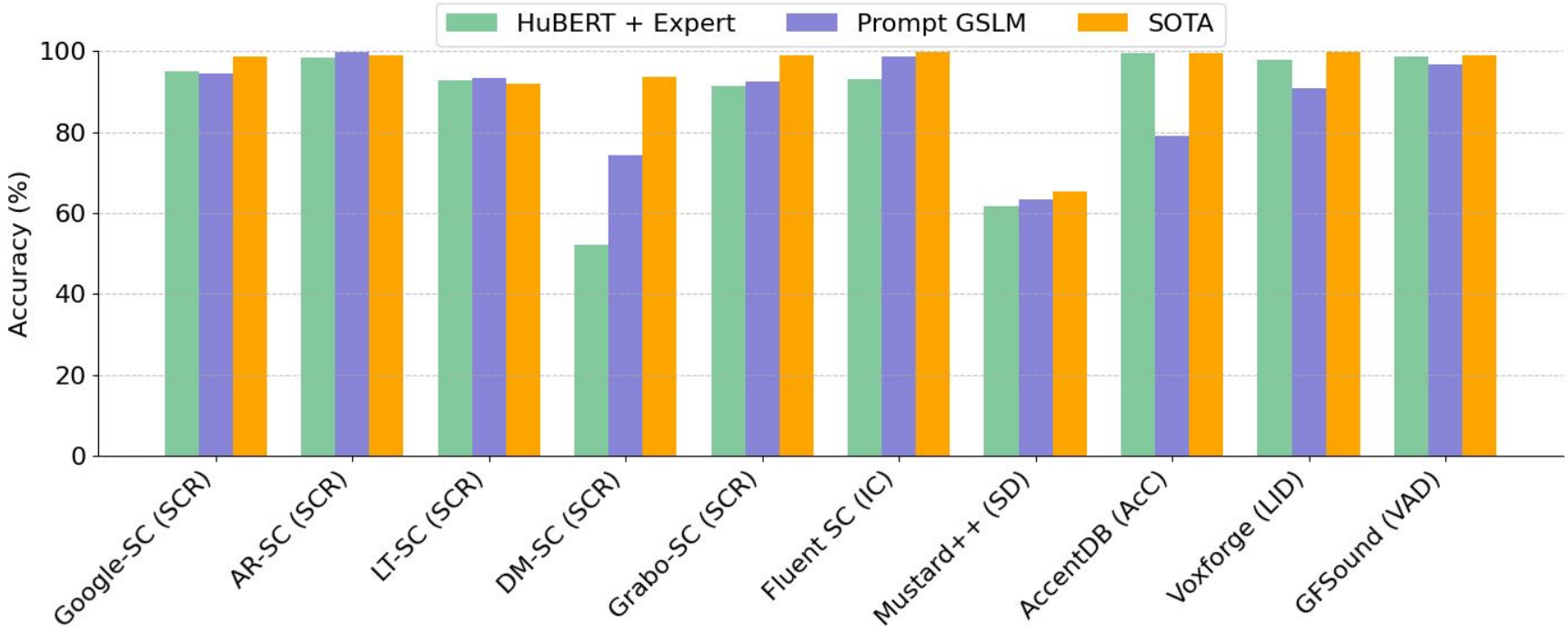
Multi-label classification: e.g. [ACTIVATE], [LIGHTS], [BEDROOM]

Speech Classification - Prompt GSLM



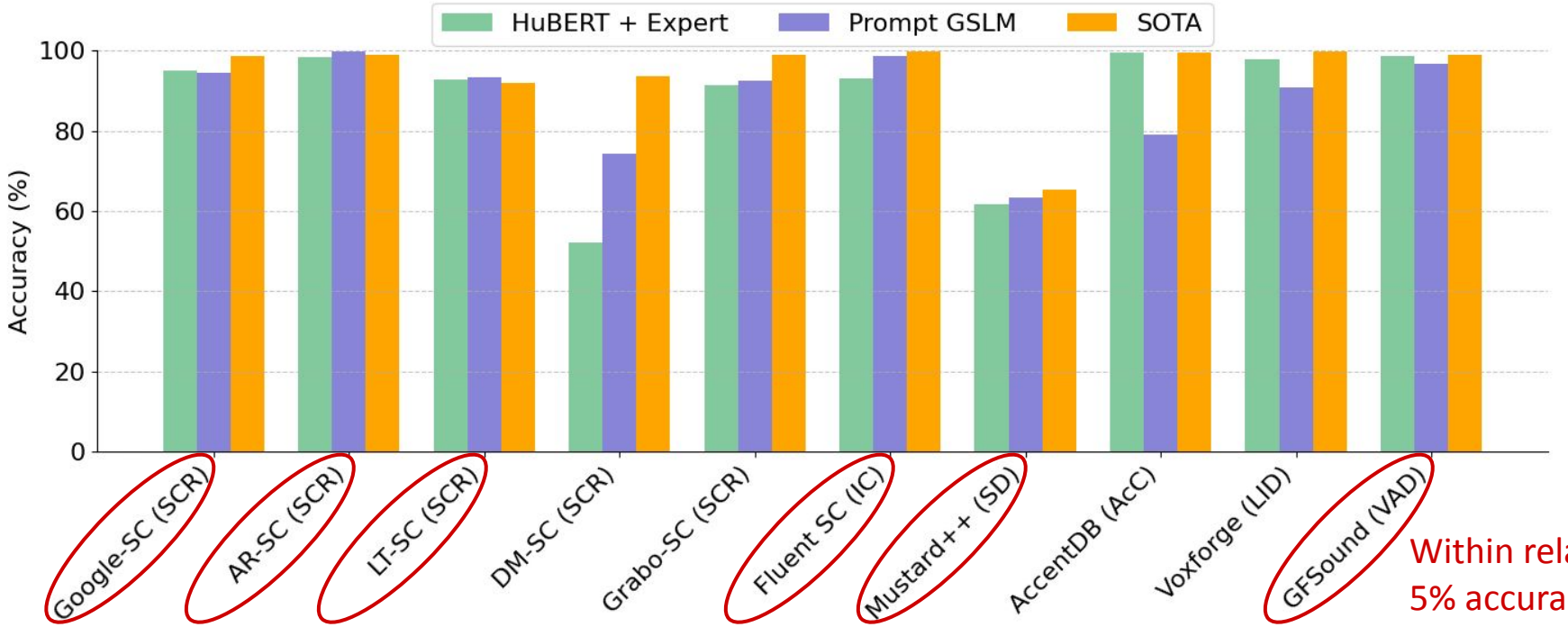
Sarcasm Detection (SD), Accent Classification(AcC)
Language Identification (LID), Voice Activity Detection (VAD)

Speech Classification - Prompt GSLM



- **HuBERT + Expert**: Pre-train, Fine-tune paradigm (**0.2M** trainable parameters)
- **Prompt GSLM**: Prompting paradigm (**0.15M** trainable parameters)
- **SOTA**: Best model from the literature (dedicated trained model)

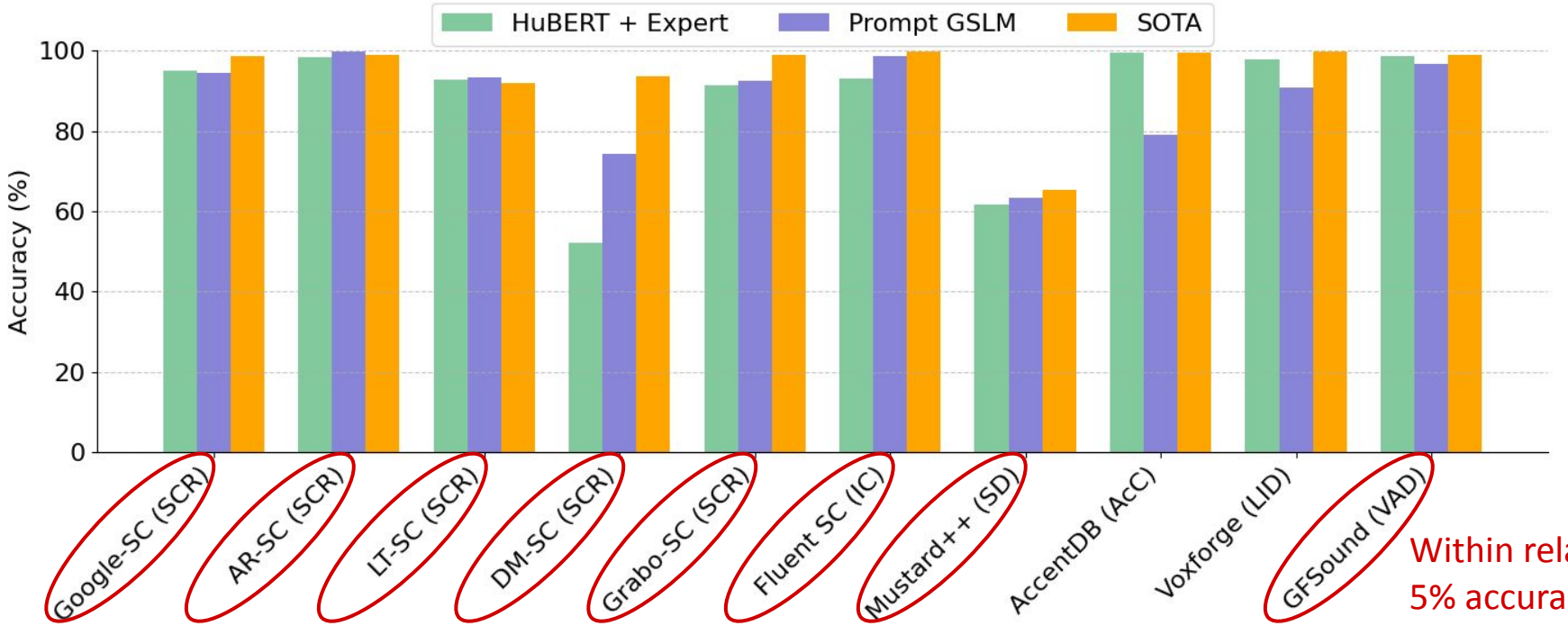
Speech Classification - Prompt GSLM



Within relative 5% accuracy difference

- Comparing **Prompt GSLM** and **SOTA:** For some tasks, prompting can achieve comparible performance to SOTA.
- Note that prompting adopts a unified framework.

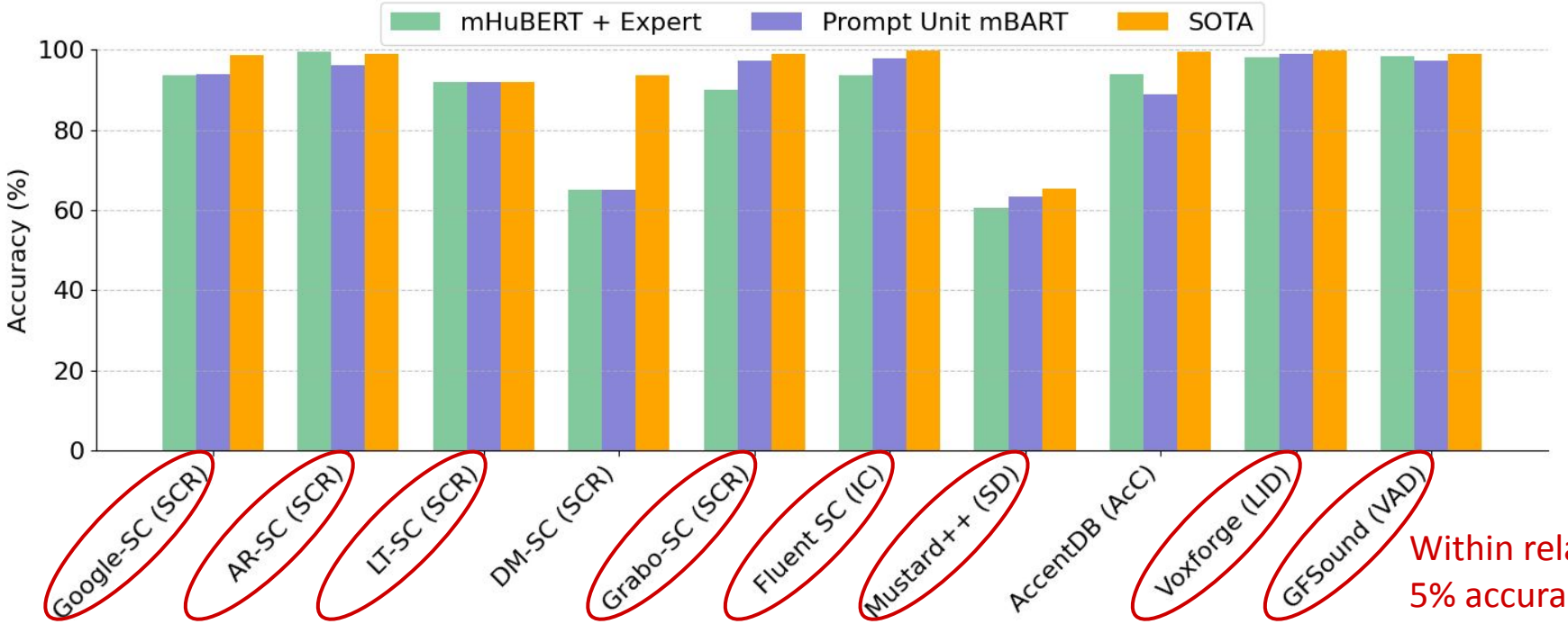
Speech Classification - Prompt GSLM



Within relative 5% accuracy difference

- Comparing **Prompt GSLM** and **HuBERT + Expert:** Prompting is competitive to pre-train, fine-tune paradigm in 8 out of 10 tasks.

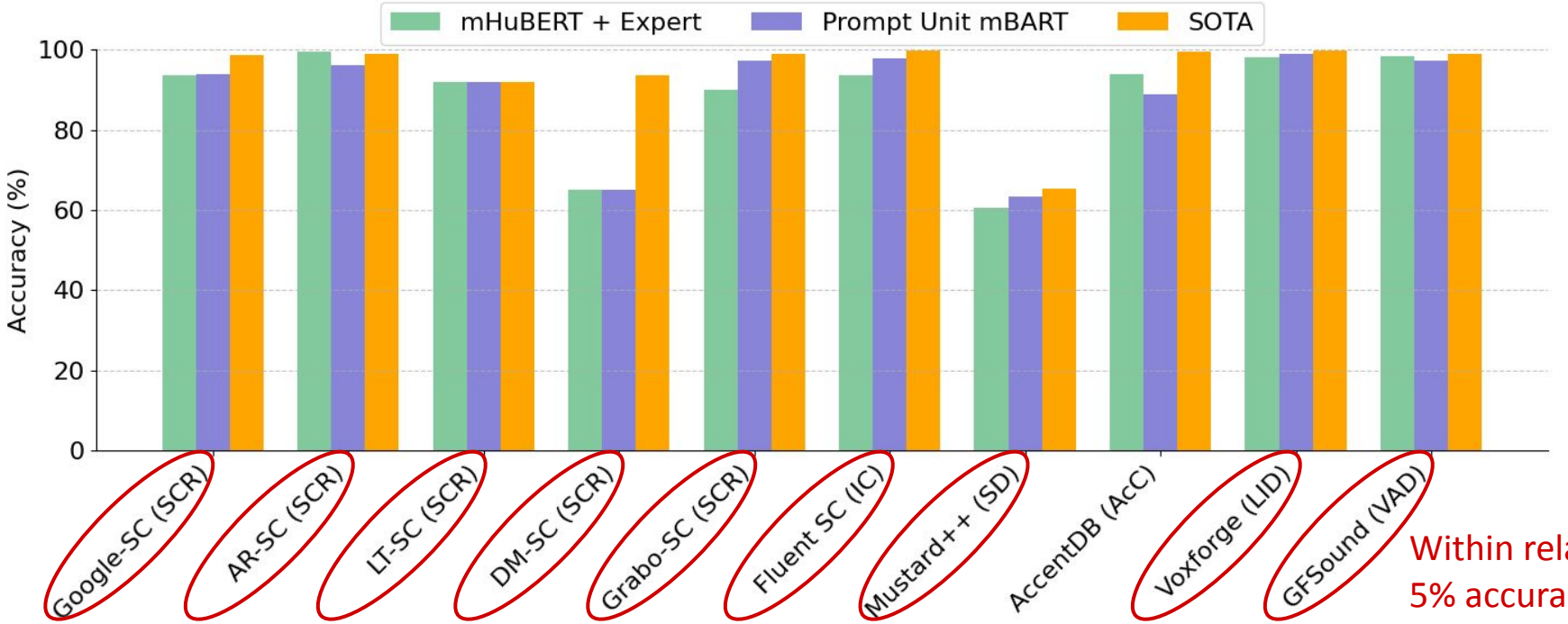
Speech Classification - Prompt Unit mBART



Within relative 5% accuracy difference

- Comparing **Prompt Unit mBART** and **SOTA:** Prompting can achieve competitive performance to SOTA in 8 out of 10 tasks.

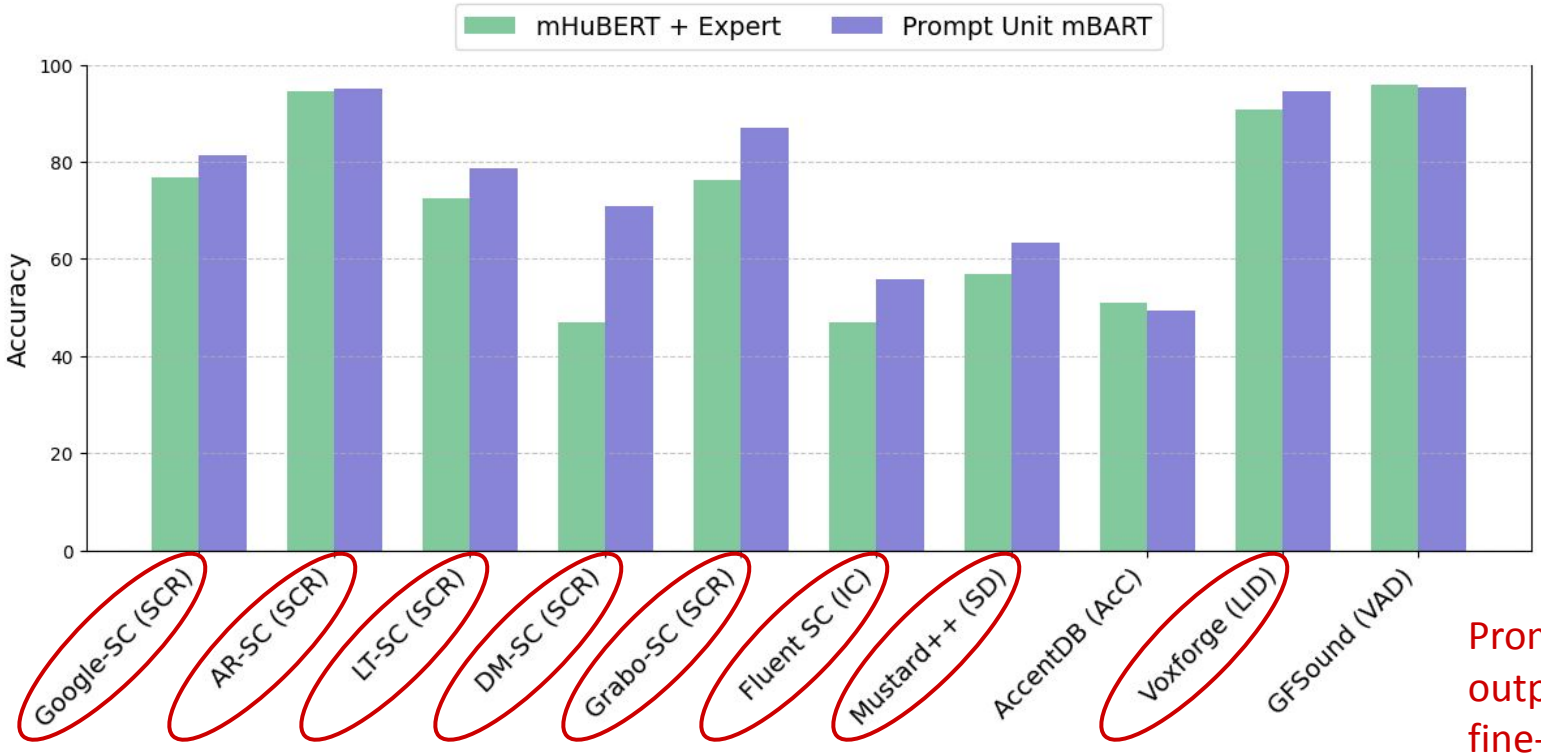
Speech Classification - Prompt Unit mBART



Within relative 5% accuracy difference

- Comparing **Prompt Unit mBART** and **mHuBERT + Expert:** Prompting is competitive to pre-train, fine-tune paradigm

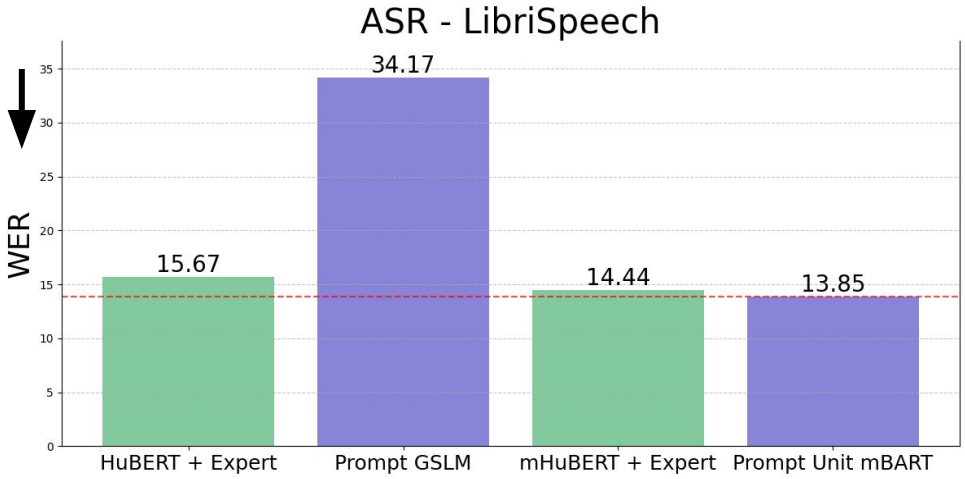
Speech Classification - Prompt Unit mBART



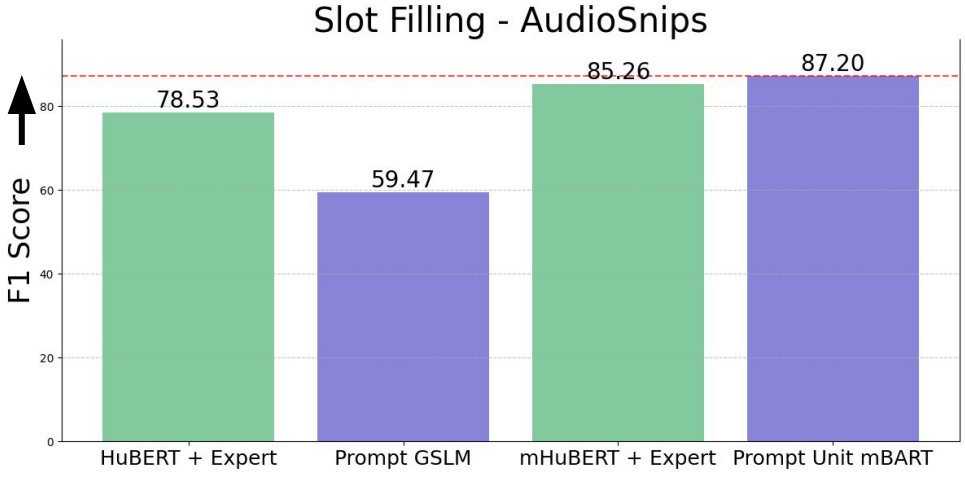
Prompting outperforms fine-tuning

- **10-shot Learning**. Each class contains only 10 training data.
- Comparing **Prompt Unit mBART** and **mHuBERT + Expert:** Prompting outperforms in 8 out of 10 tasks.

Sequence Generation Tasks



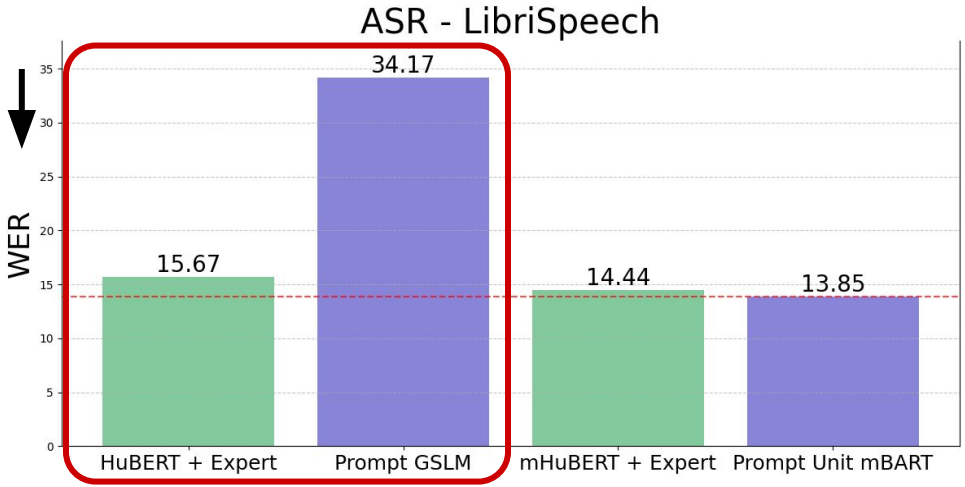
- ASR: transcribe an utterance into characters

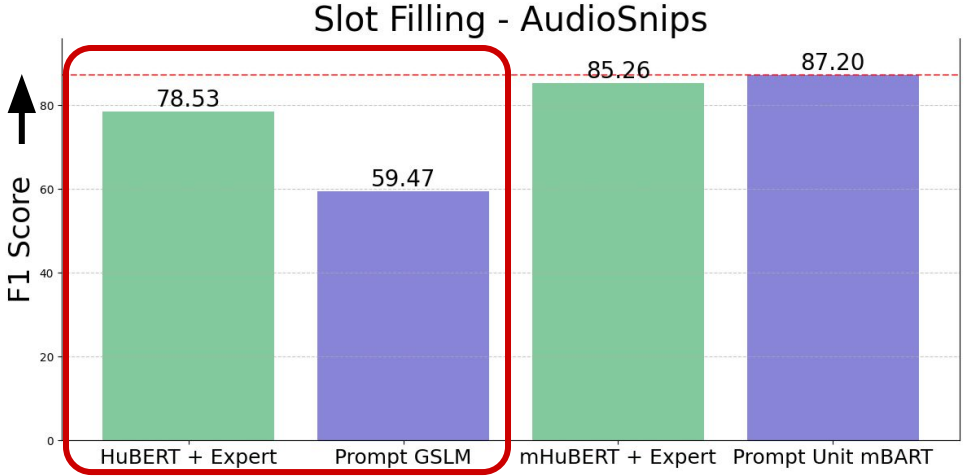


- Slot Filling: conduct ASR and identify the slot types at the same time.

What's the weather like in <L>NewYork<L/> <T>tomorrow</T>?

Sequence Generation Tasks

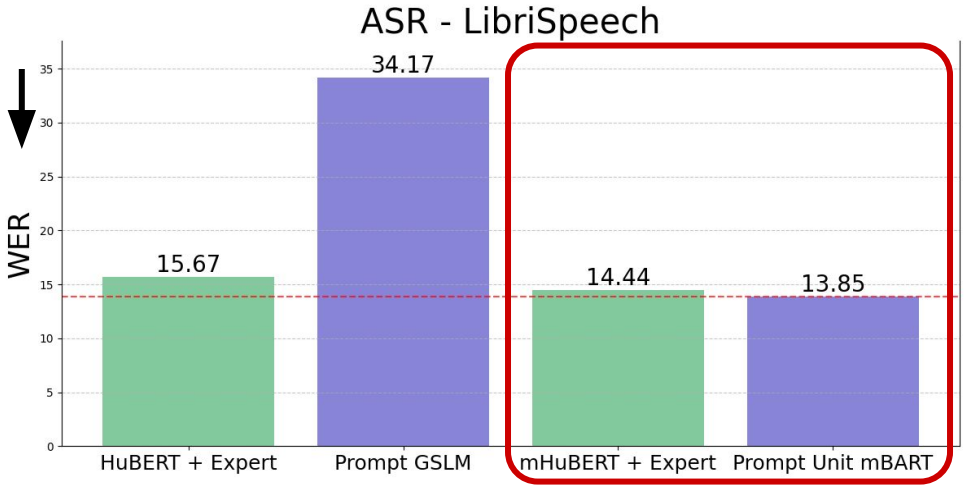


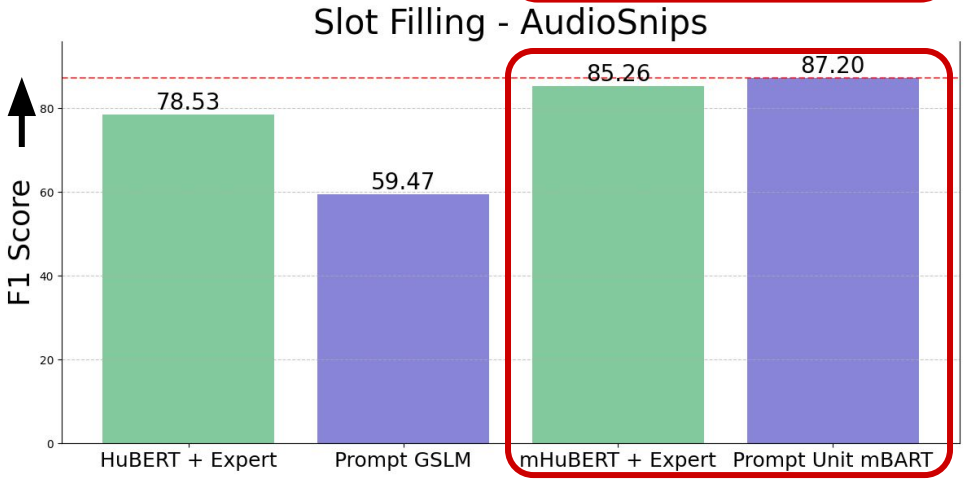


Scenario	Traniable Params.
HuBERT + Expert	2.9M
Prompt **GSLM**	4.5M

- GSLM underperforms compared to the pre-train, fine-tune paradigm.

Sequence Generation Tasks

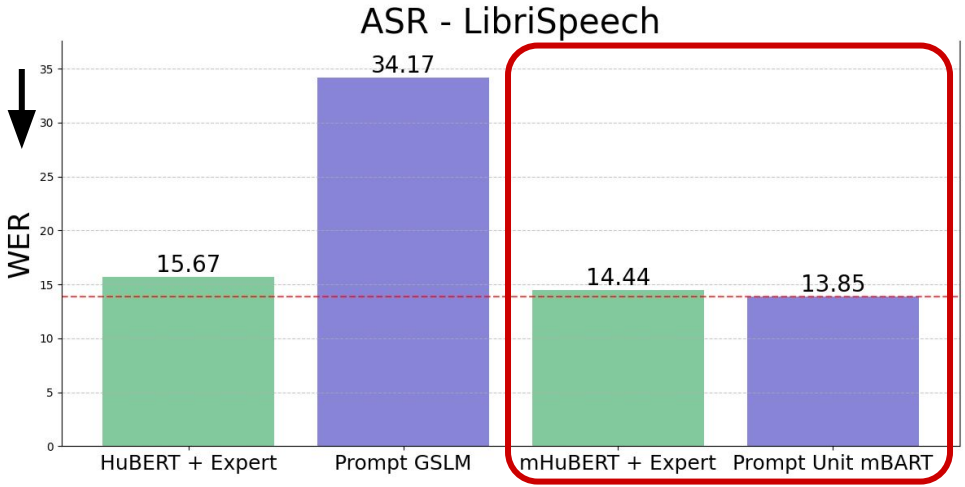


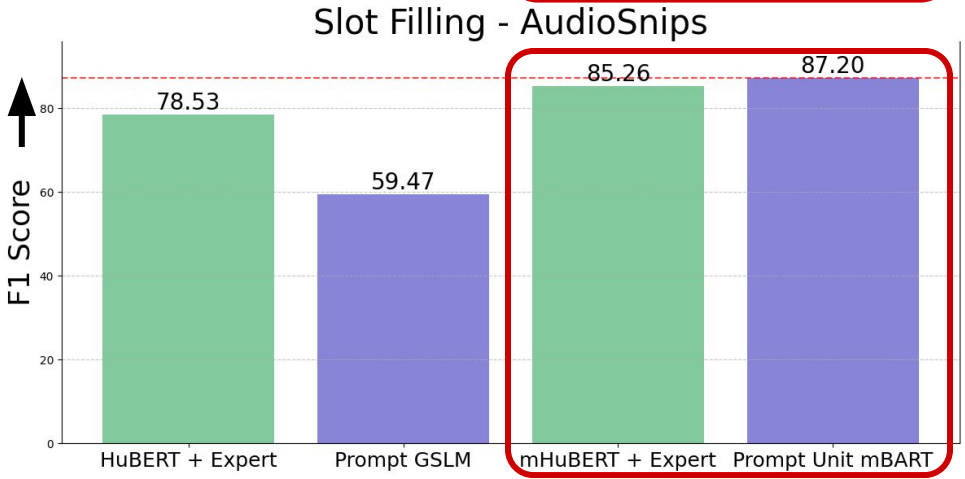


Scenario	Traniable Params.
HuBERT + Expert	2.9M
Prompt **Unit mBART**	2.6M

- Prompting Unit mBART outperforms the pre-train, fine-tune paradigm.

Sequence Generation Tasks

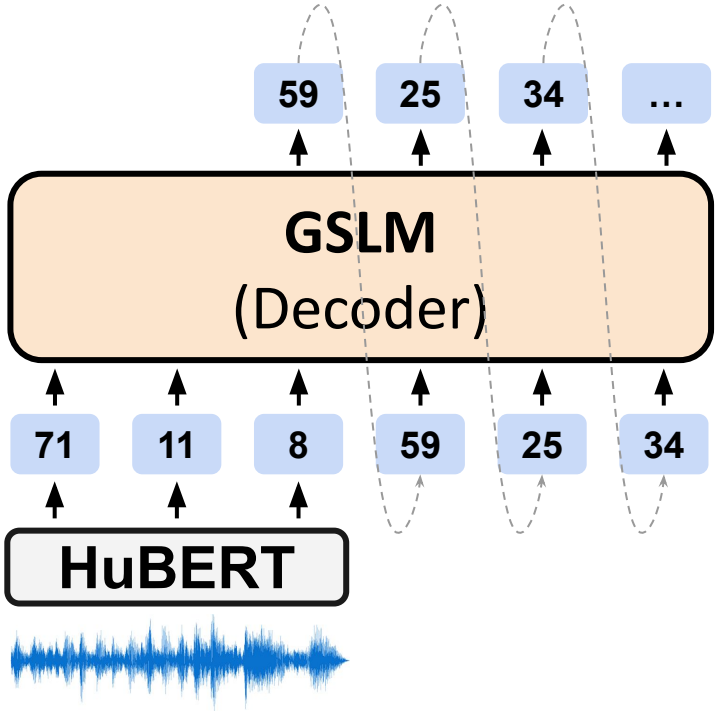




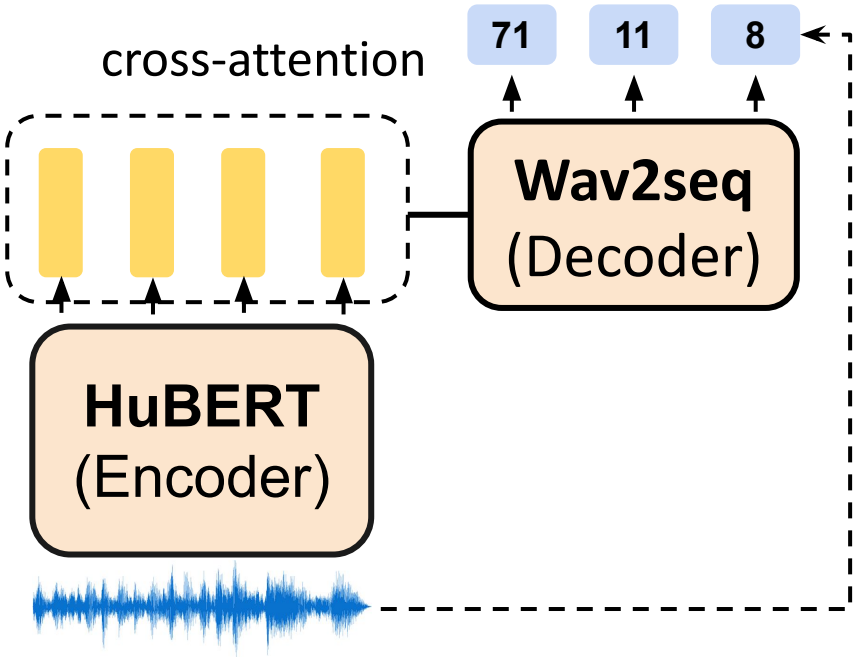
Scenario	Traniable Params.
HuBERT + Expert	2.9M
Prompt **Unit mBART**	2.6M

- Prompting Unit mBART outperforms the pre-train, fine-tune paradigm.
- For prompting, model architecture and pre-training task matter.
- Encoder-decoder model is better than decoder-only model?

Decoder-only vs. Encoder-Decoder Speech LM



Next-token prediction



Pseudo speech recognition

Prompting and adapter tuning for self-supervised encoder-decoder speech model, ASRU2023 (https://arxiv.org/abs/2310.02971)

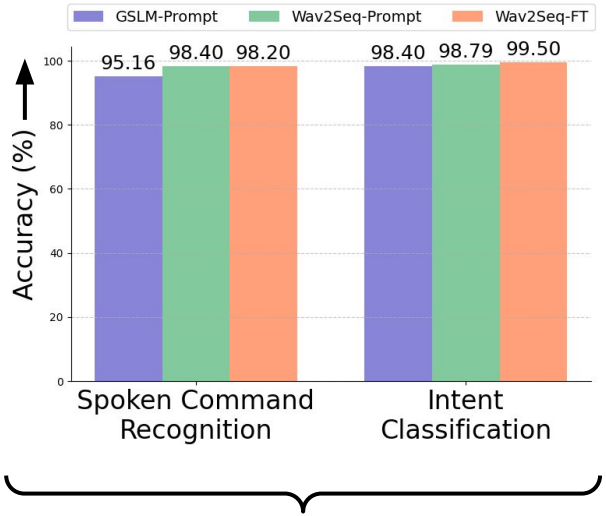
Decoder-only vs. Encoder-Decoder Speech LM

Model	Architecture	Params	Data	Task
GSLM	HuBERT Unit + 12-layer Transformer	~150M	LibriLight 60k hours	Next-token prediction
Wav2Seq	HuBERT + 6-layer Transformer	~150M	LibriSpeech 960 hours	Pseudo speech recognition

GSLM has more training data

Similar model size

Decoder-only vs. Encoder-Decoder Speech LM

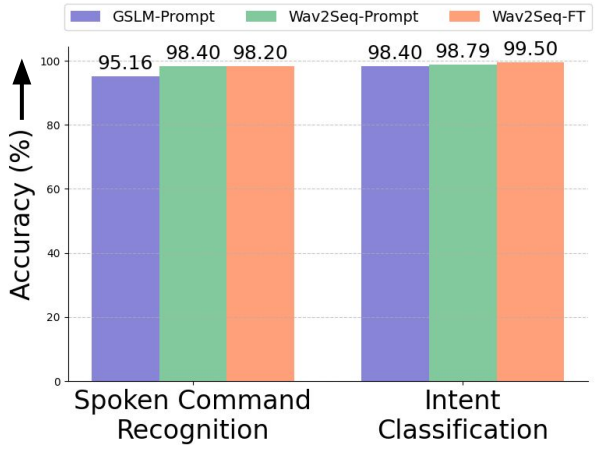


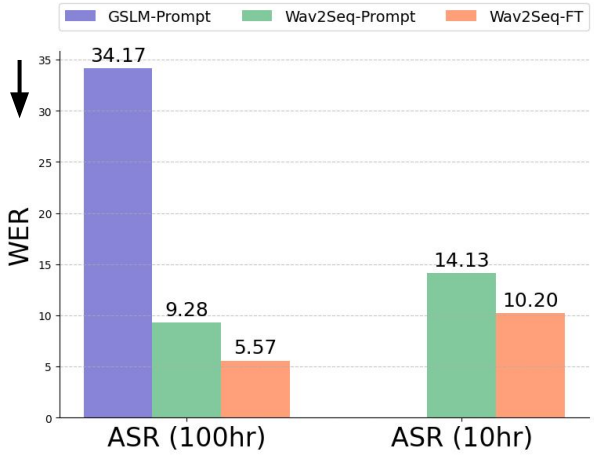
Speech Classification

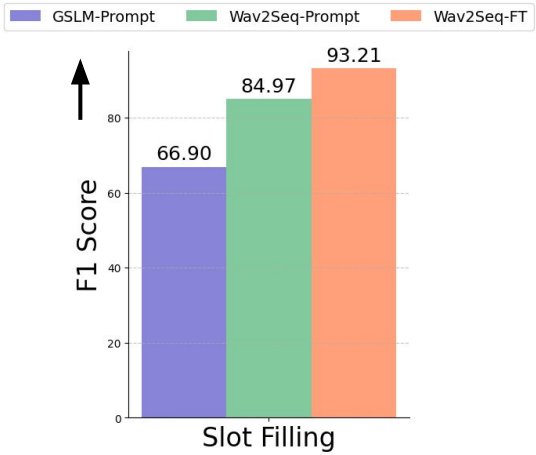
Comparable performance

Prompting and adapter tuning for self-supervised encoder-decoder speech model, ASRU2023 (https://arxiv.org/abs/2310.02971)

Decoder-only vs. Encoder-Decoder Speech LM







Speech Classification

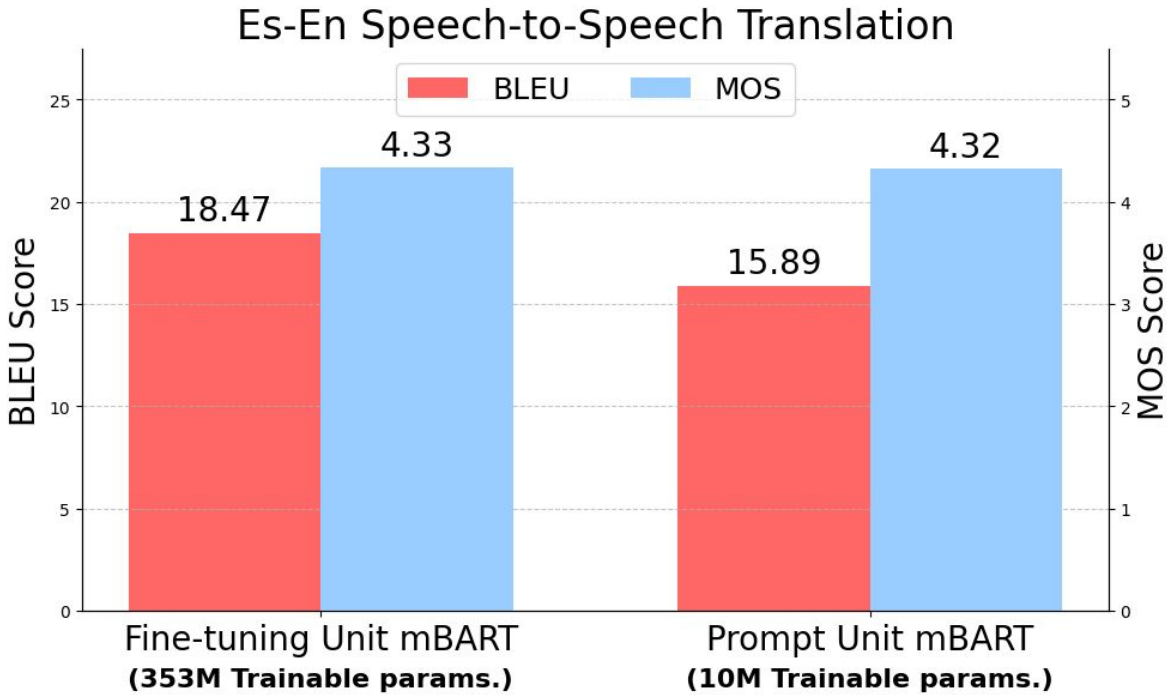
Sequence Generation

Comparable performance

Prompt Wav2Seq is much better than **prompt GSLM**

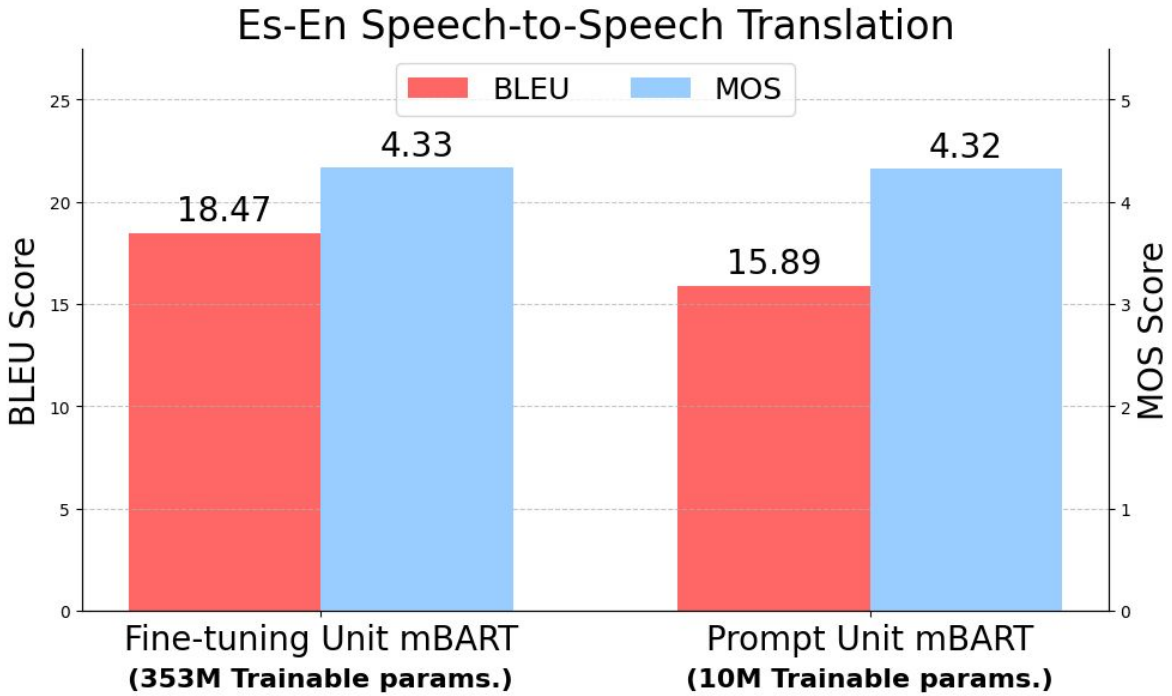
Prompting and adapter tuning for self-supervised encoder-decoder speech model, ASRU2023 (https://arxiv.org/abs/2310.02971)

Speech Generation



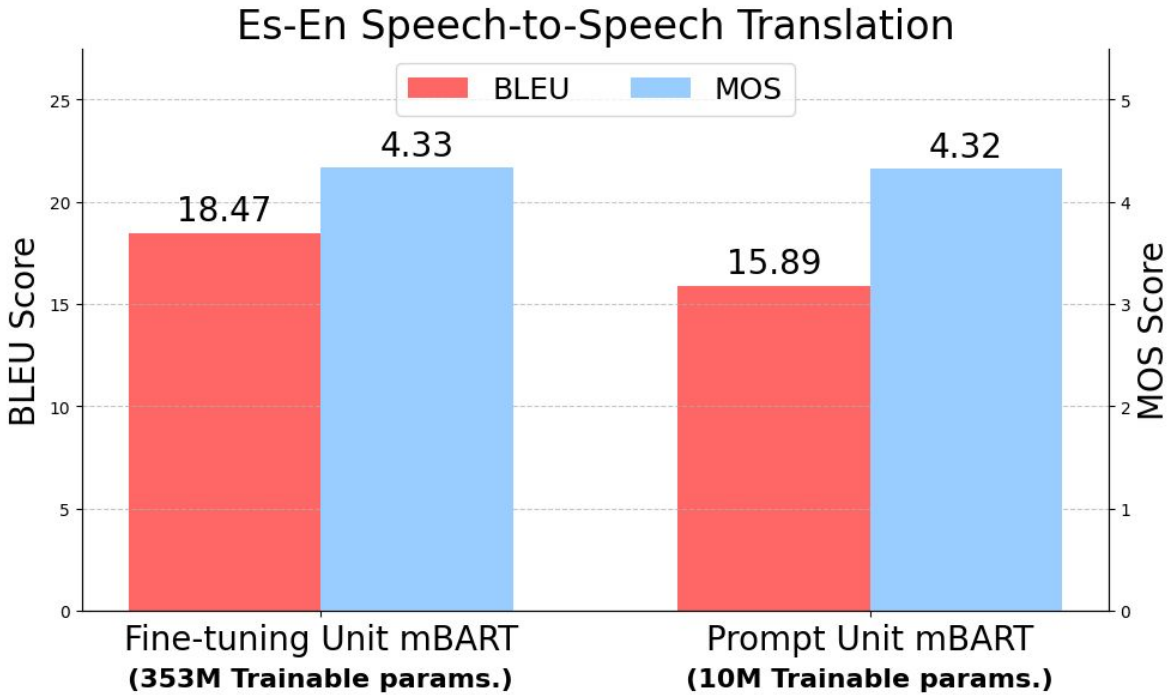
- Left: fine-tuning the whole Unit mBART (353M params.)
- Right: Prompting Unit mBART (10M params)

Speech Generation



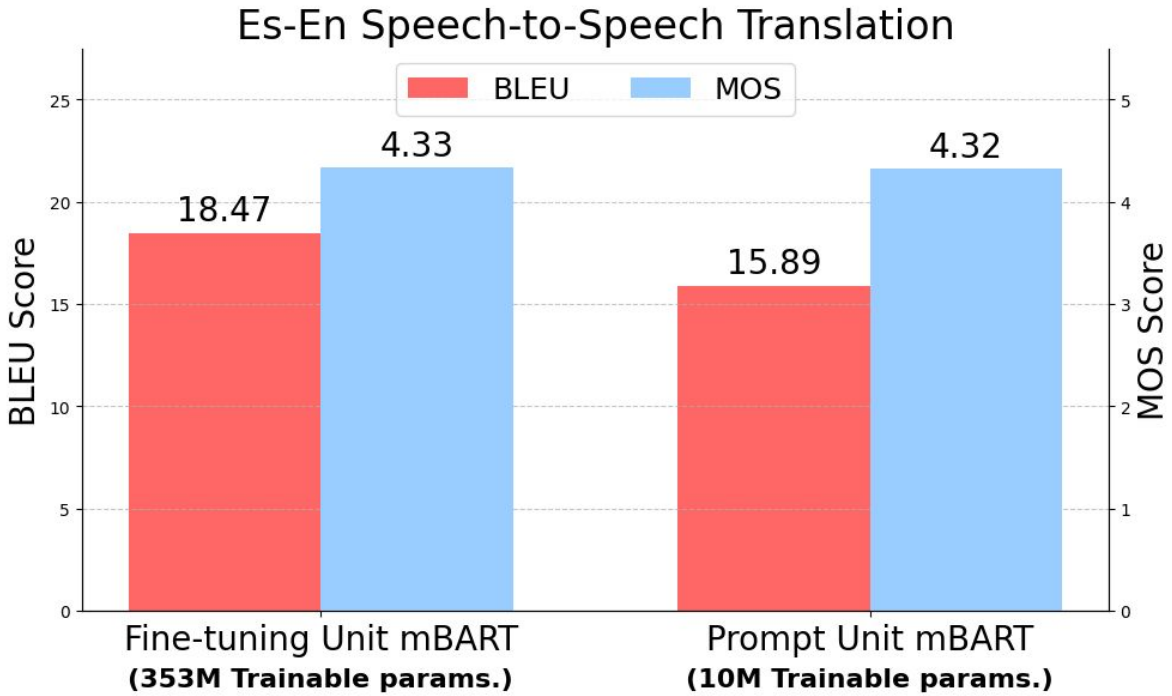
- BLEU score: Translation quality
- MOS score: Speech quality

Speech Generation



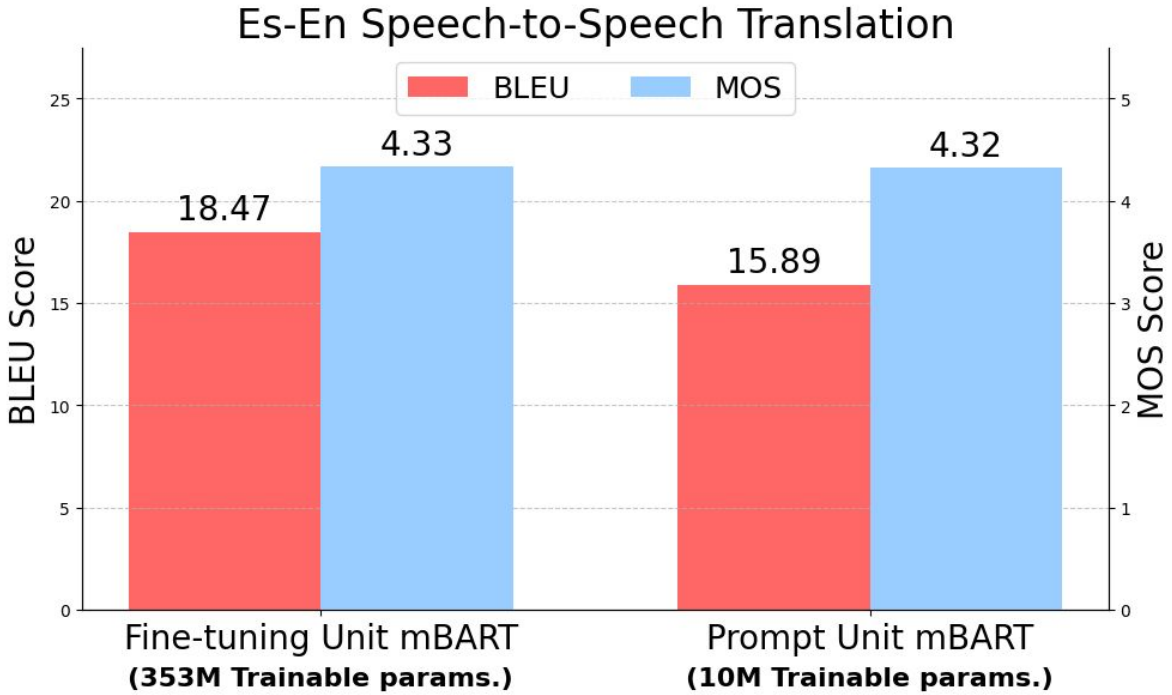
- Prompting Unit mBART has performance drop but uses much fewer trainable params.
- Both have similar MOS score.

Speech Generation



- Fine-tuning HuBERT/mHuBERT fails
- GSLM also fails

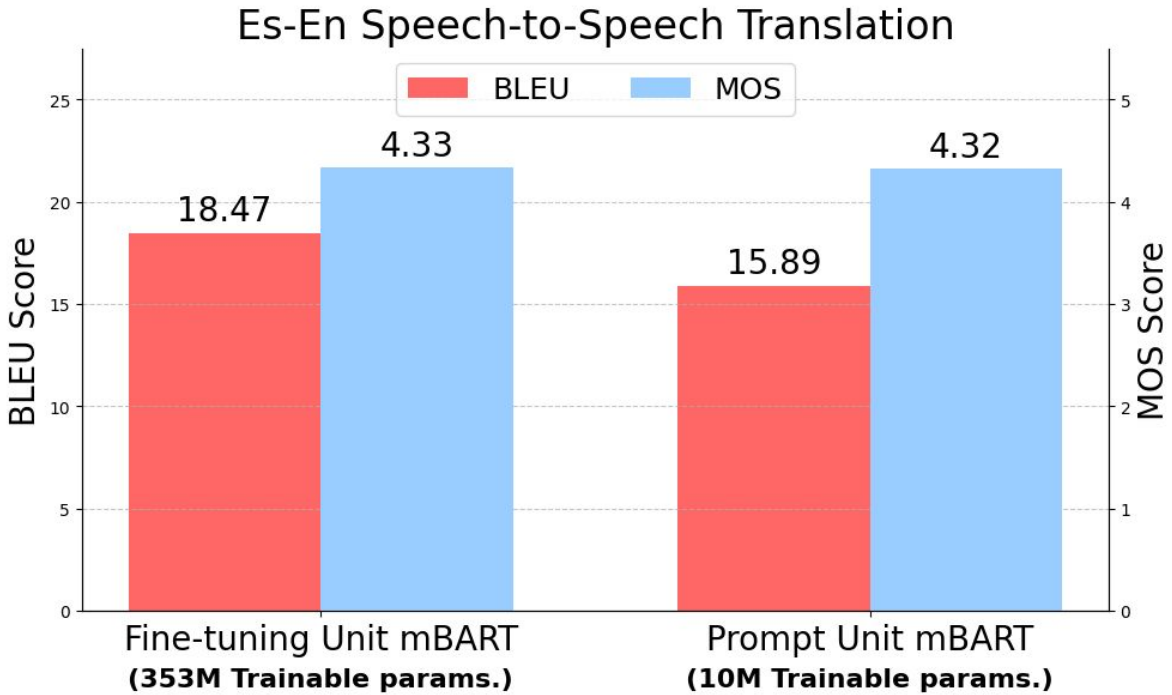
Speech Generation



- speech-to-speech translation is challenging, often require auxiliary tasks

Direct speech-to-speech translation with a sequence-to-sequence model (https://arxiv.org/abs/1904.06037)

Speech Generation



Demo for speech translation, speech continuation, and speech inpainting: https://kwchang.org/SpeechPrompt/speechgen.html

Fully Utilize the information in Discrete Units

- Until now, we use random mapping to bridge the units and the labels.
  - Speech classification tasks
  - Sequence generation tasks

Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table
(Verbalizer)

Fully Utilize the information in Discrete Units

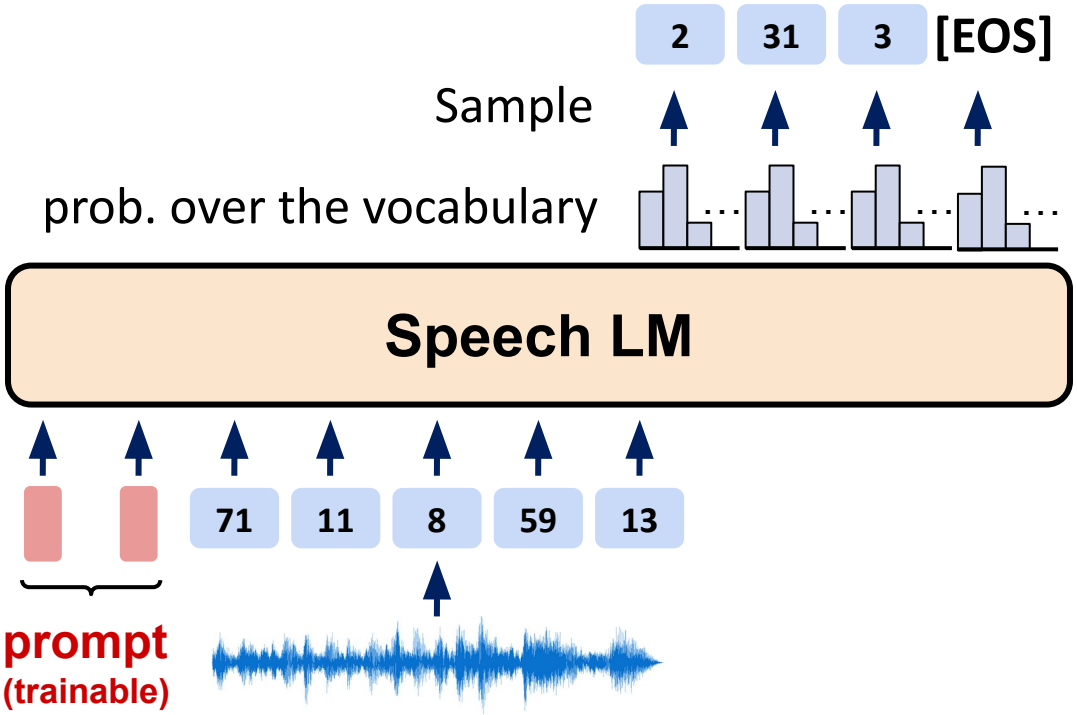
- Until now, we use random mapping to bridge the units and the labels.
  - Speech classification tasks
  - Sequence generation tasks

Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)

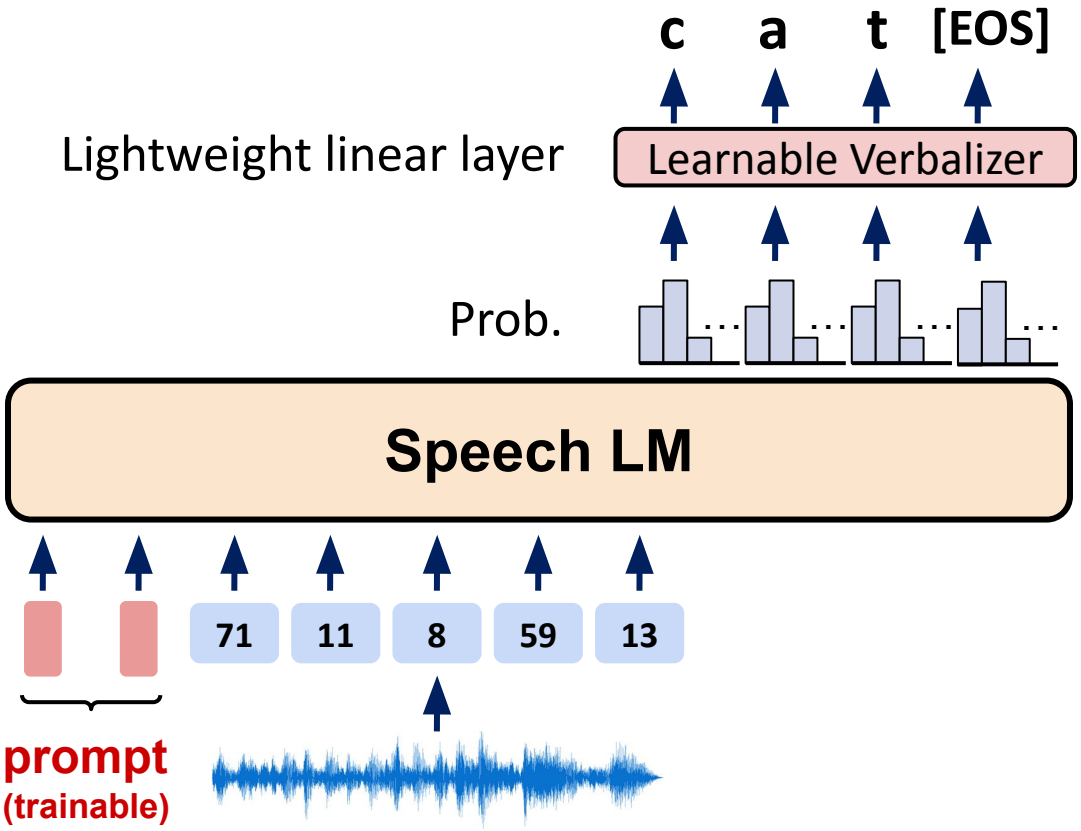
Contains rich information

Fully Utilize the information in Discrete Units

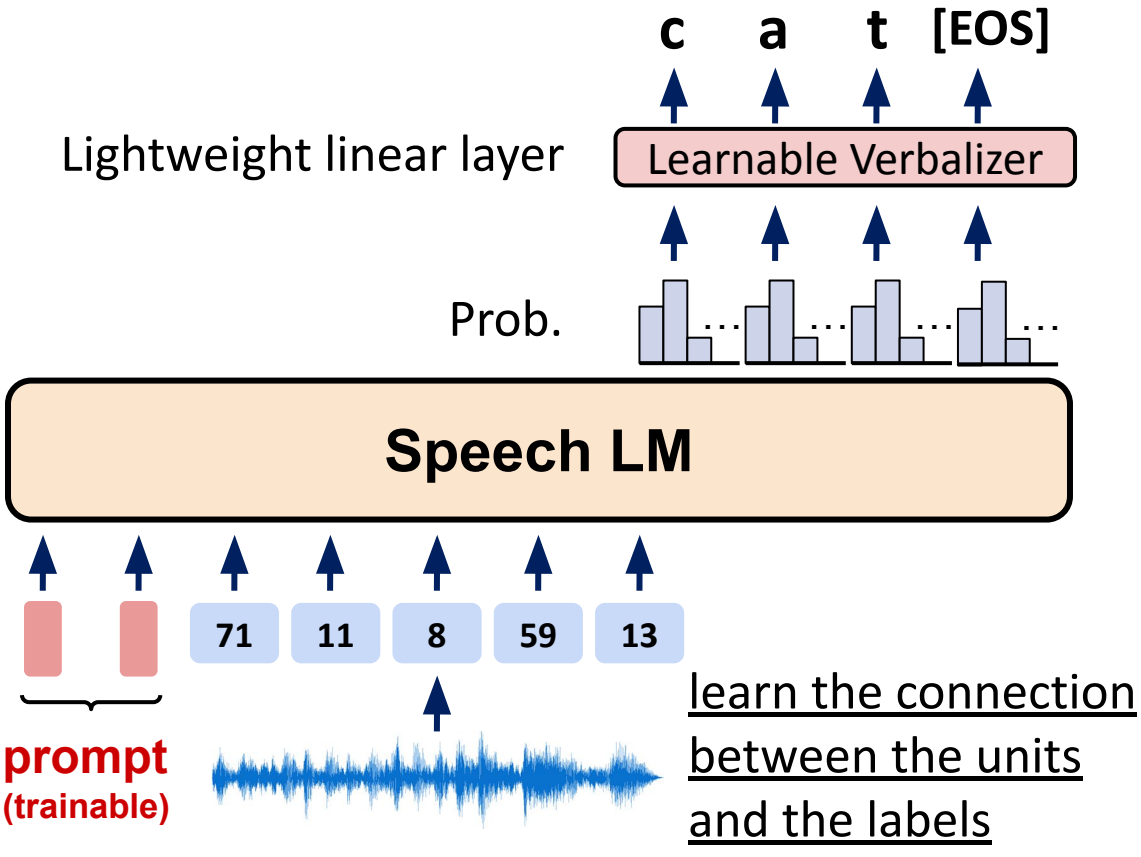


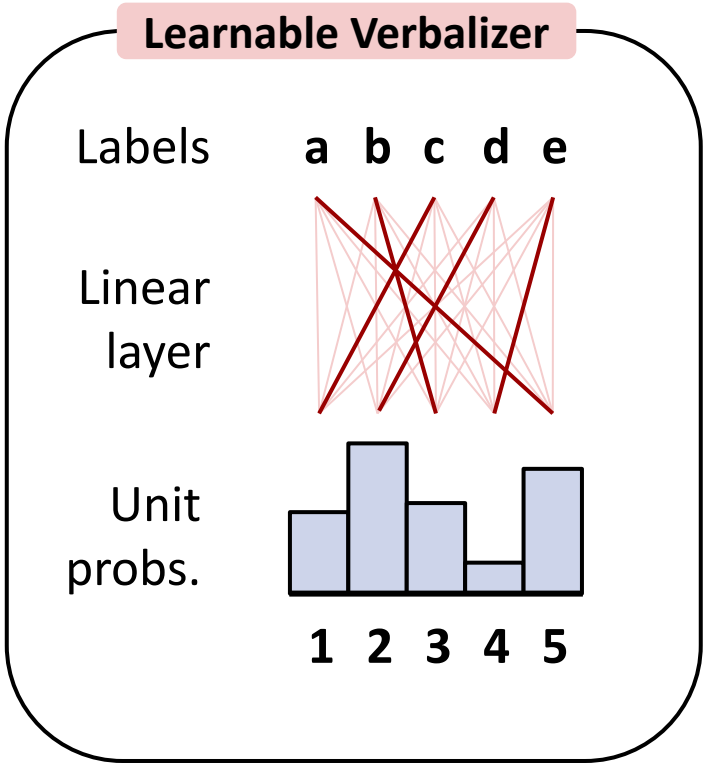
At each timestep, the LM generates the probability distribution over the vocabulary

Fully Utilize the information in Discrete Units



Fully Utilize the information in Discrete Units



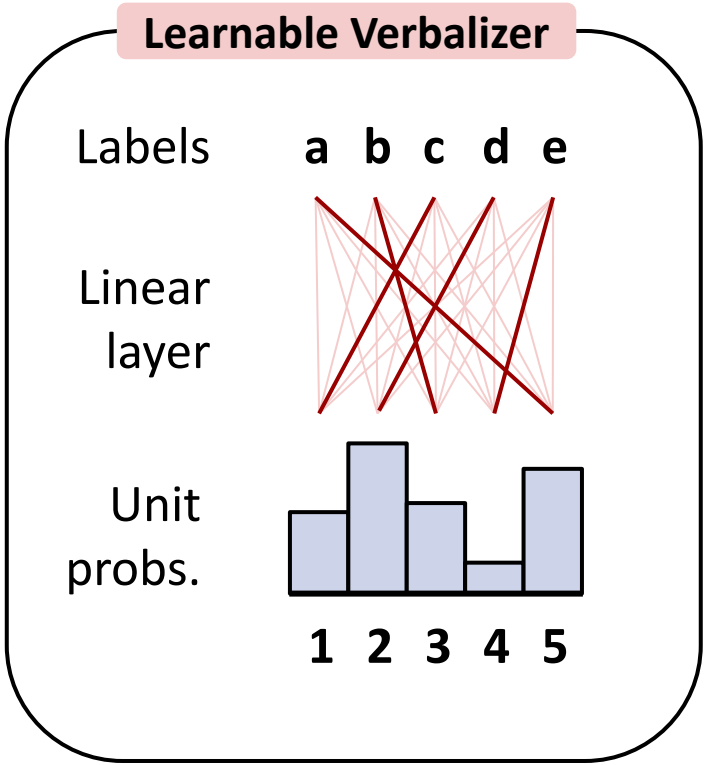


Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	‘B’	‘H’	‘V’
Unit	290	470	577

The unit that has the largest weight in the linear layer for one specific label.



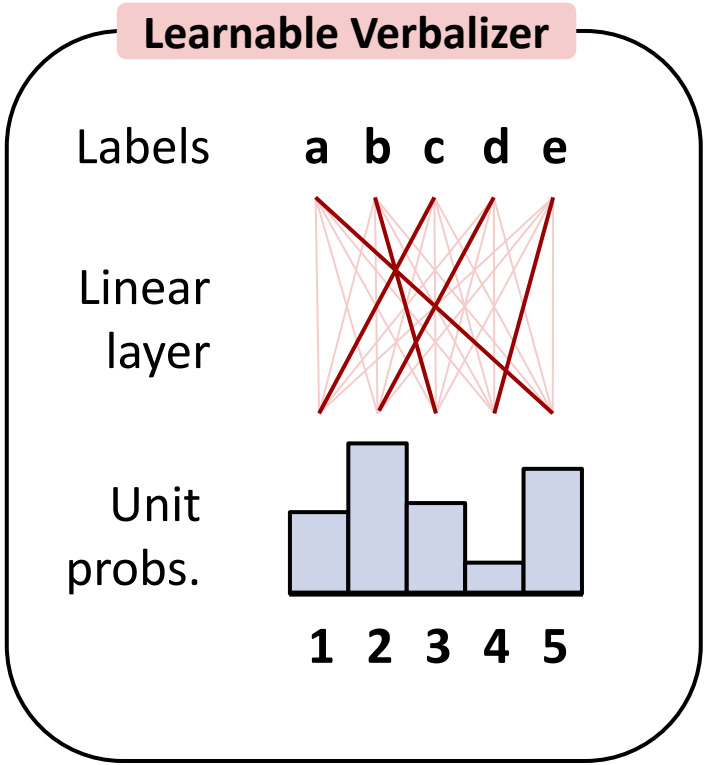
Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	‘B’	‘H’	‘V’
Unit	290	470	577

The unit that has the largest weight in the linear layer for one specific label.

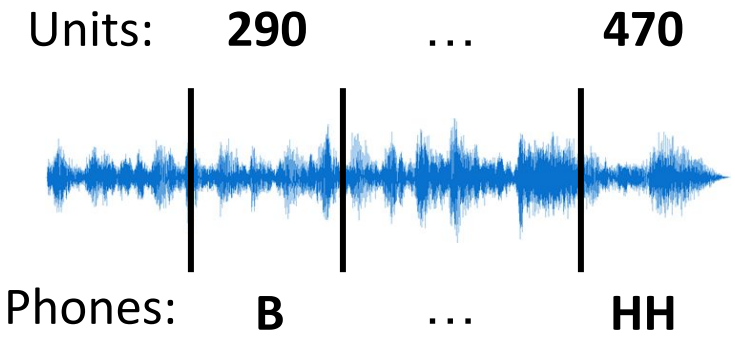
What is the meaning of these discrete units?



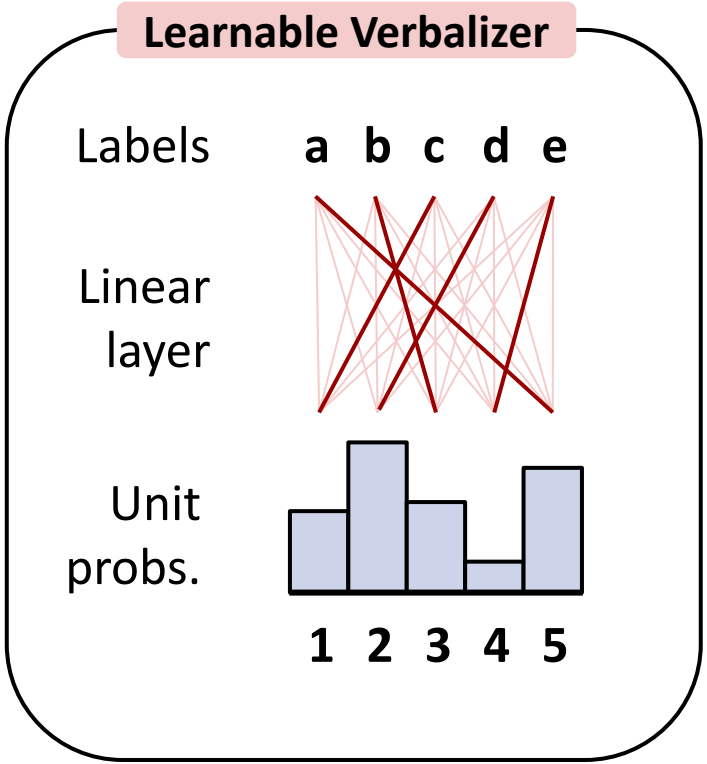
Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	‘B’	‘H’	‘V’
Unit	290	470	577



Forced alignment on LibriSpeech

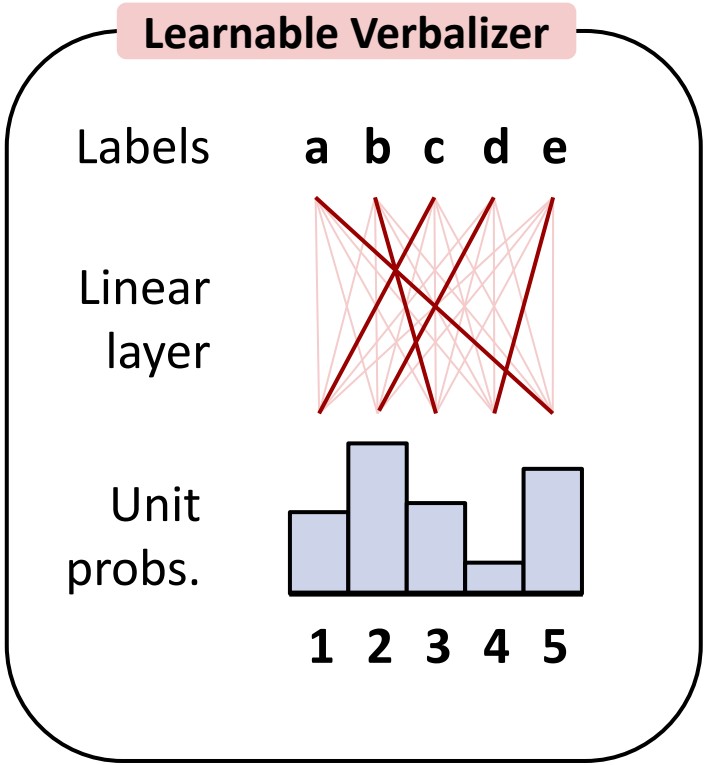


Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	‘B’	‘H’	‘V’
Unit	290	470	577
Phoneme	B, V	HH, T	V, B

- Top 2 phonemes correlated to the units
- The learnable verbalizer can automatically find the units for the labels

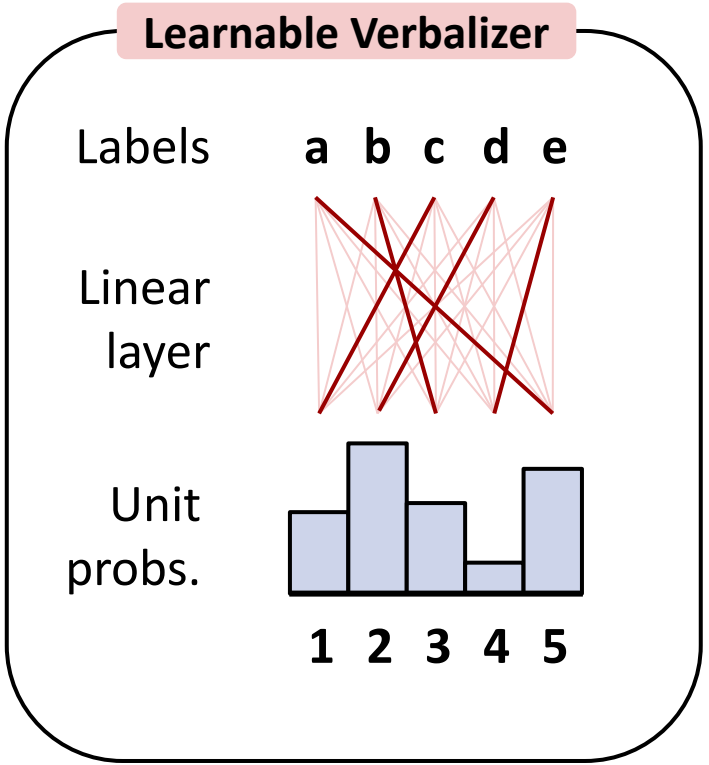


Learnable Verbalizer - A Case Study

For prompting Unit mBART in Phoneme Recognition (PR)

Label	‘F’	‘K’	‘TH’
Unit	958	487	918
Phoneme	F, V	K, G	TH, DH

- Top 2 phonemes correlated to the units
- The learnable verbalizer can automatically find the units for the labels



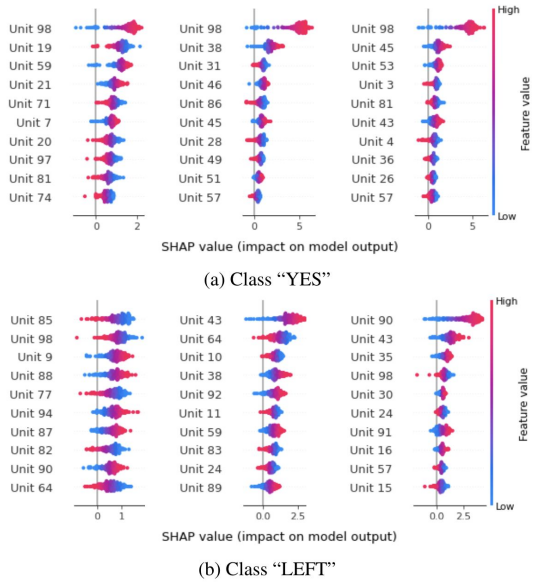
Learnable Verbalizer - A Case Study

For prompting Unit mBART in Phoneme Recognition (PR)

Label	‘F’	‘K’	‘TH’
U			
Pho			

Learnable Verbalizer

There are some units more suitable serving as labels

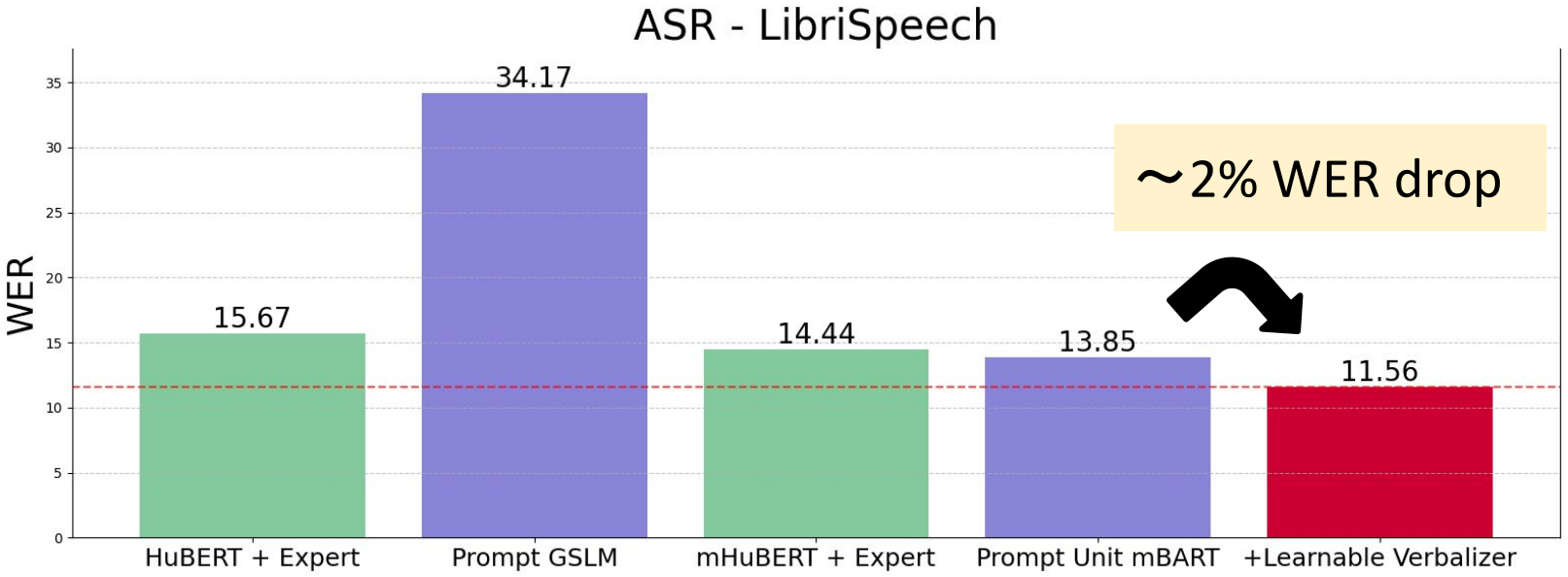


(a) Class “YES”

(b) Class “LEFT”

SpeechPrompt v2: Prompt Tuning for Speech Classification Tasks (https://arxiv.org/abs/2303.00733)

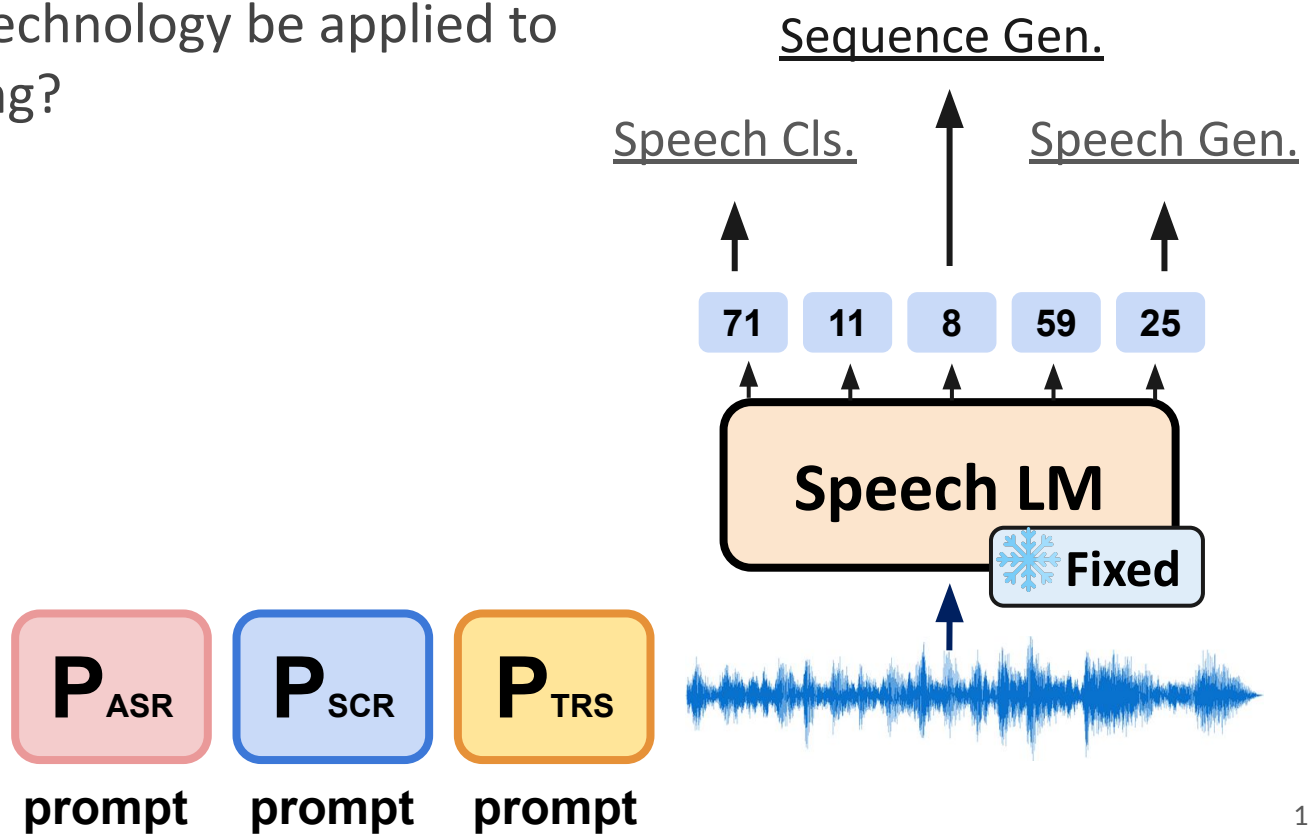
Learnable Verbalizer - A Case Study



- Performance improvement with learnable verbalizer.
- With additional parameters less than **0.03M** (**~1%** of the prompt parameters)

Prompting Paradigm

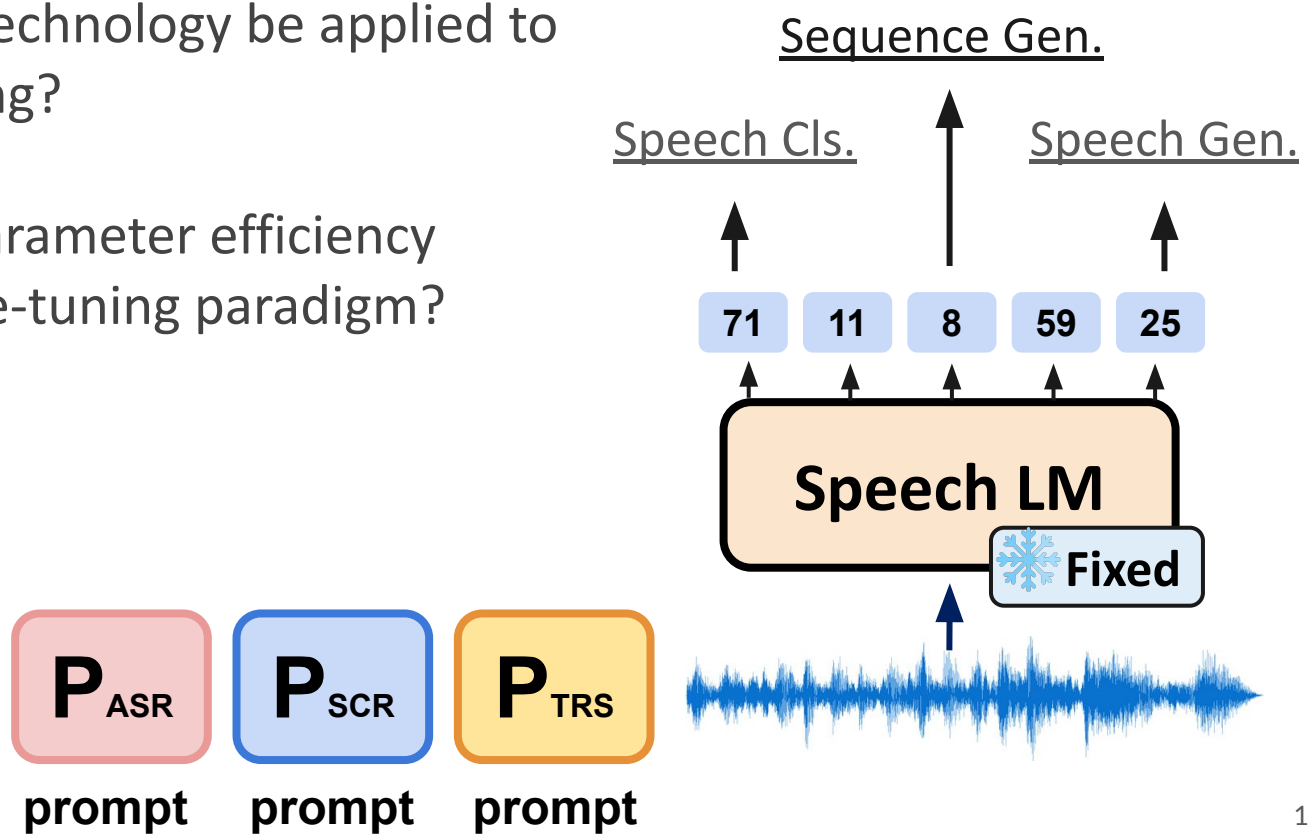
1. Can prompting technology be applied to speech processing?



Prompting Paradigm

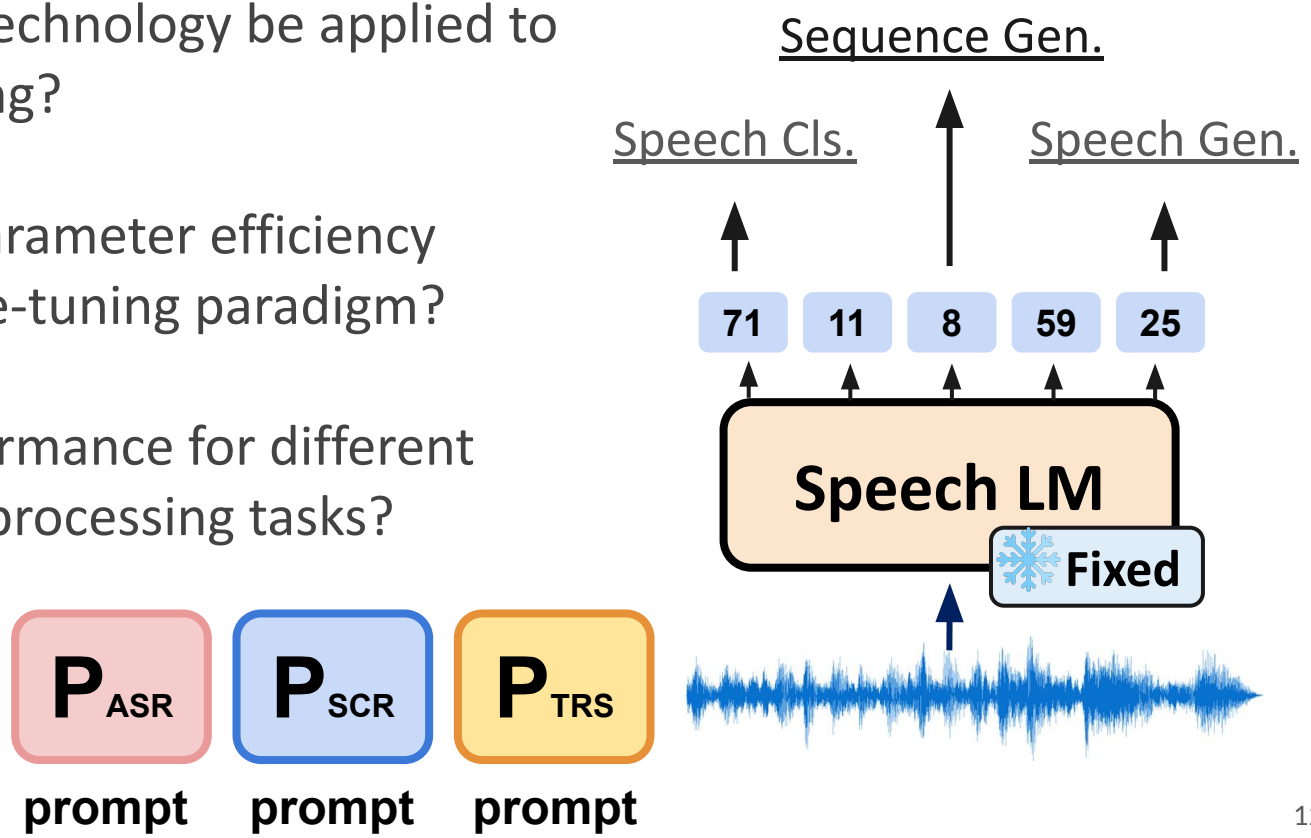
1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency compared to fine-tuning paradigm?



Prompting Paradigm

1. Can prompting technology be applied to speech processing?
2. Can it achieve parameter efficiency compared to fine-tuning paradigm?
3. How is the performance for different kinds of speech processing tasks?

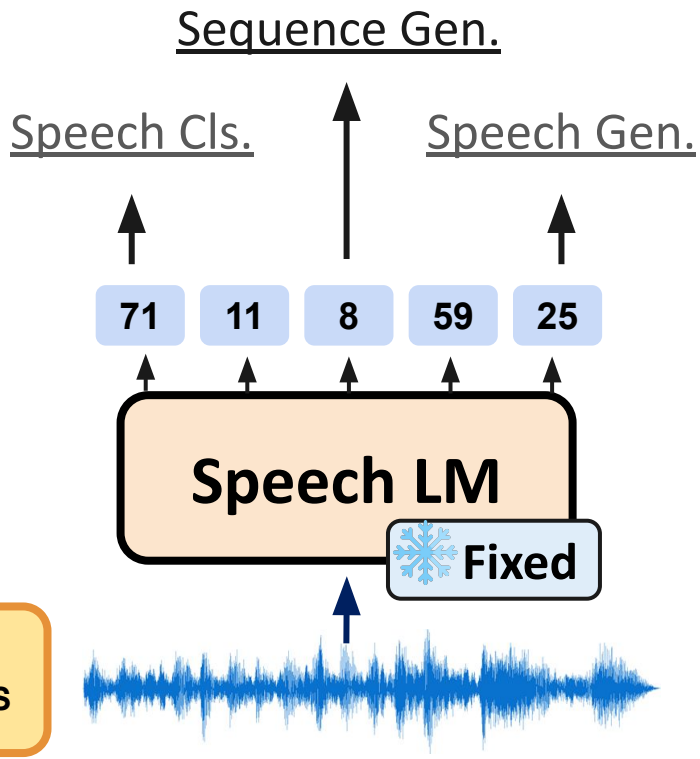


Prompting Paradigm

1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency compared to fine-tuning paradigm?

3. How is the performance for different kinds of speech processing tasks?



Limitation:

Still require training for a specifc task

In-Context Learning

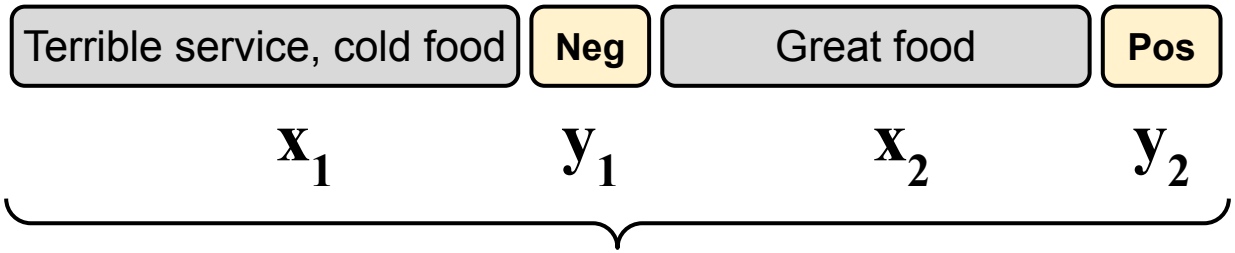
Predicting based on the demonstrations

In-Context Learning

Predicting based on the demonstrations

Language Model

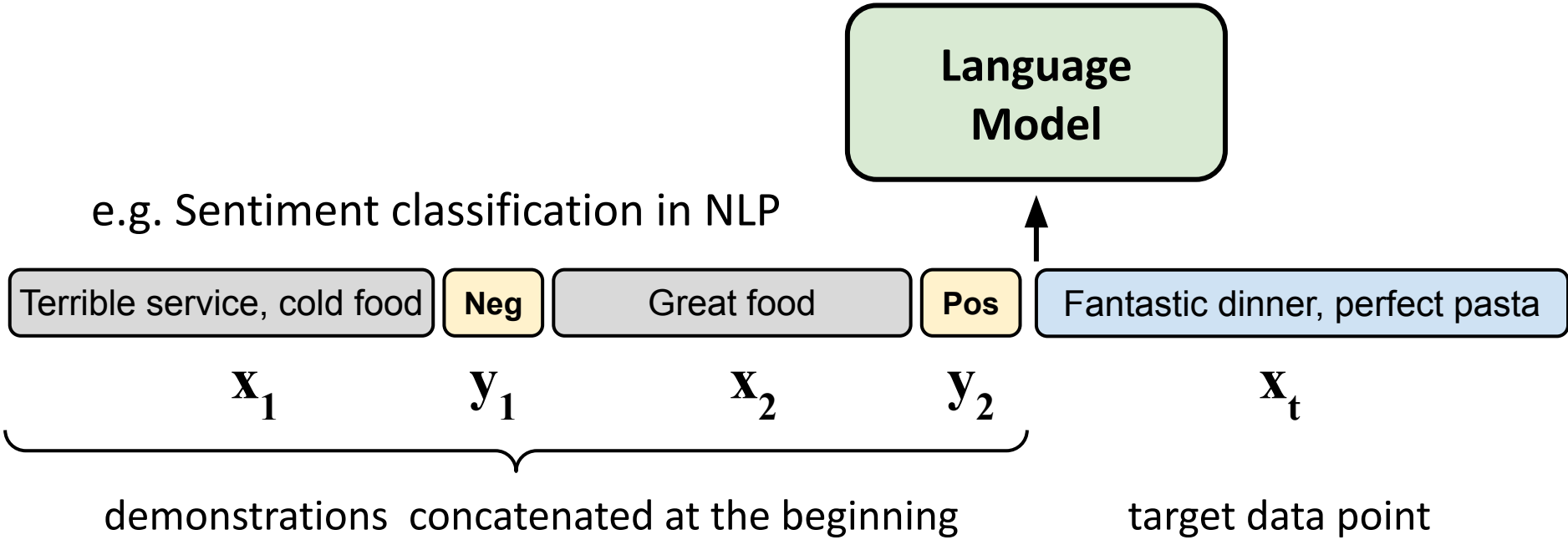
e.g. Sentiment classification in NLP



demonstrations concatenated at the beginning

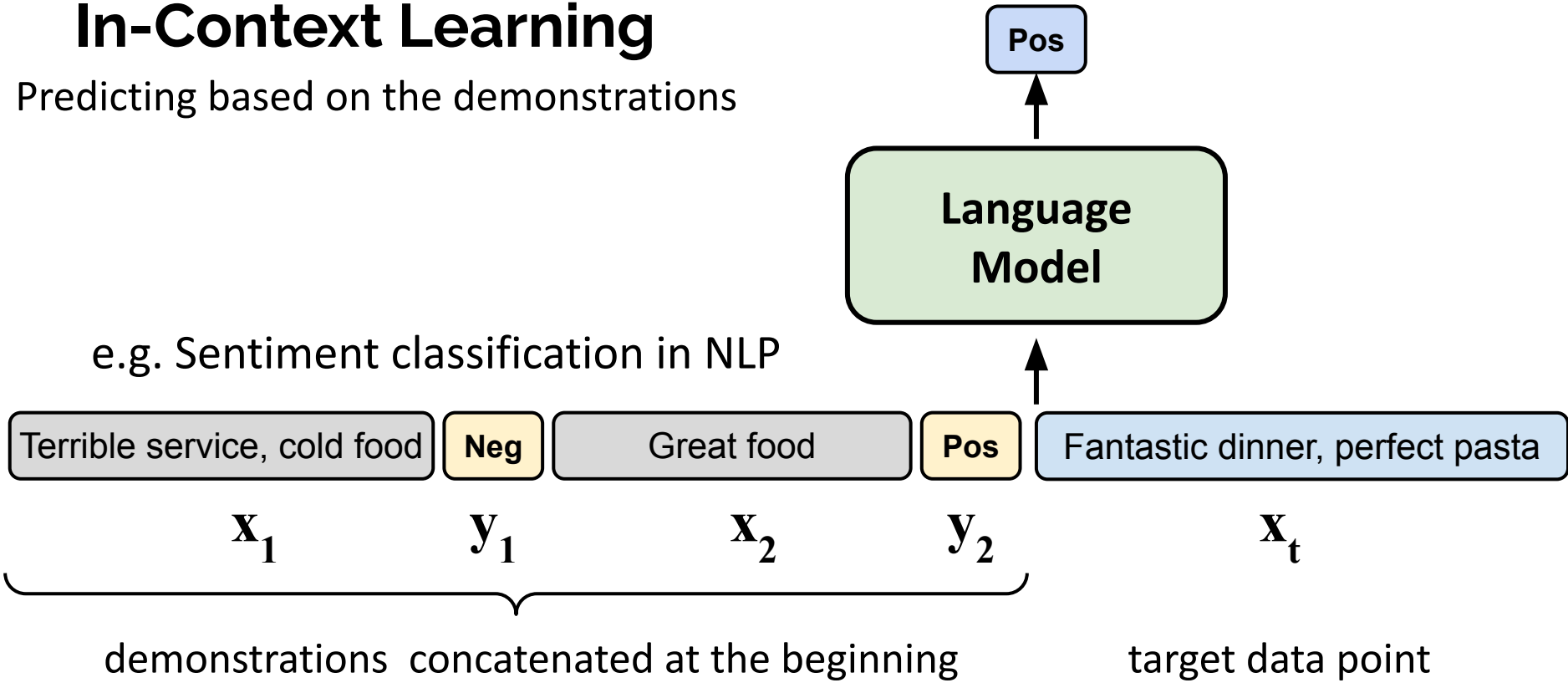
In-Context Learning

Predicting based on the demonstrations



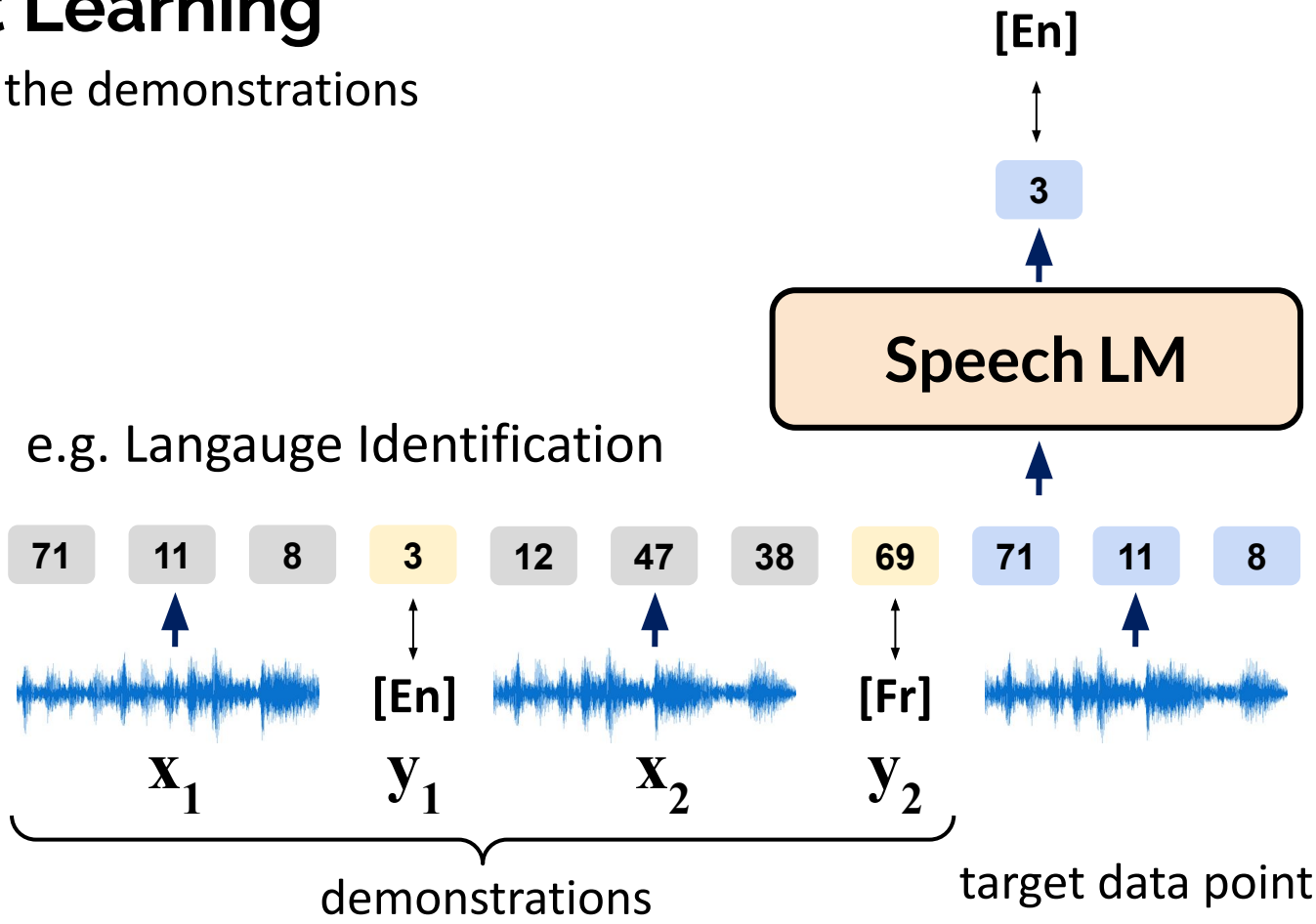
In-Context Learning

Predicting based on the demonstrations



In-Context Learning

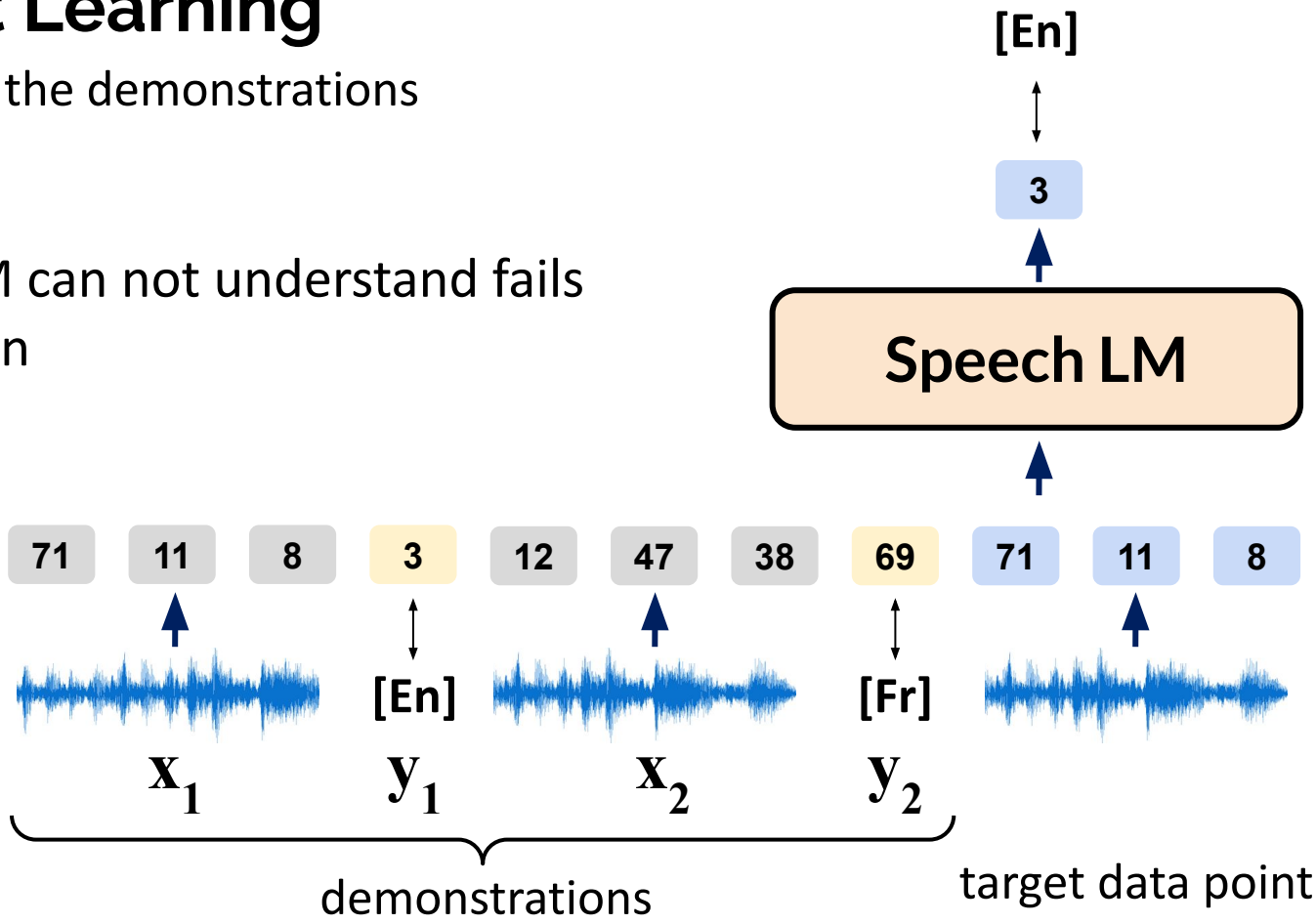
Predicting based on the demonstrations



In-Context Learning

Predicting based on the demonstrations

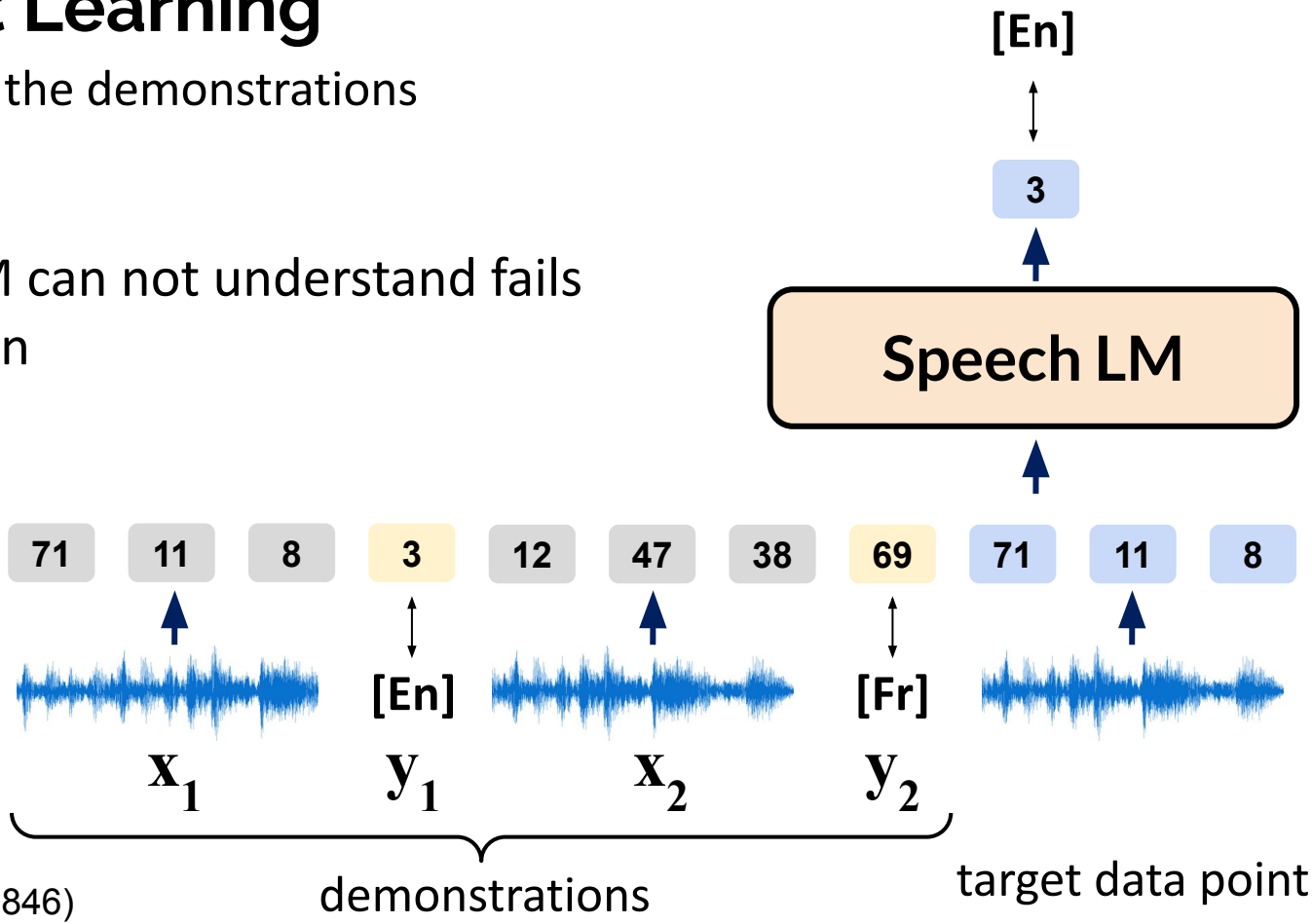
The original GSLM can not understand fails to make prediction



In-Context Learning

Predicting based on the demonstrations

The original GSLM can not understand fails to make prediction



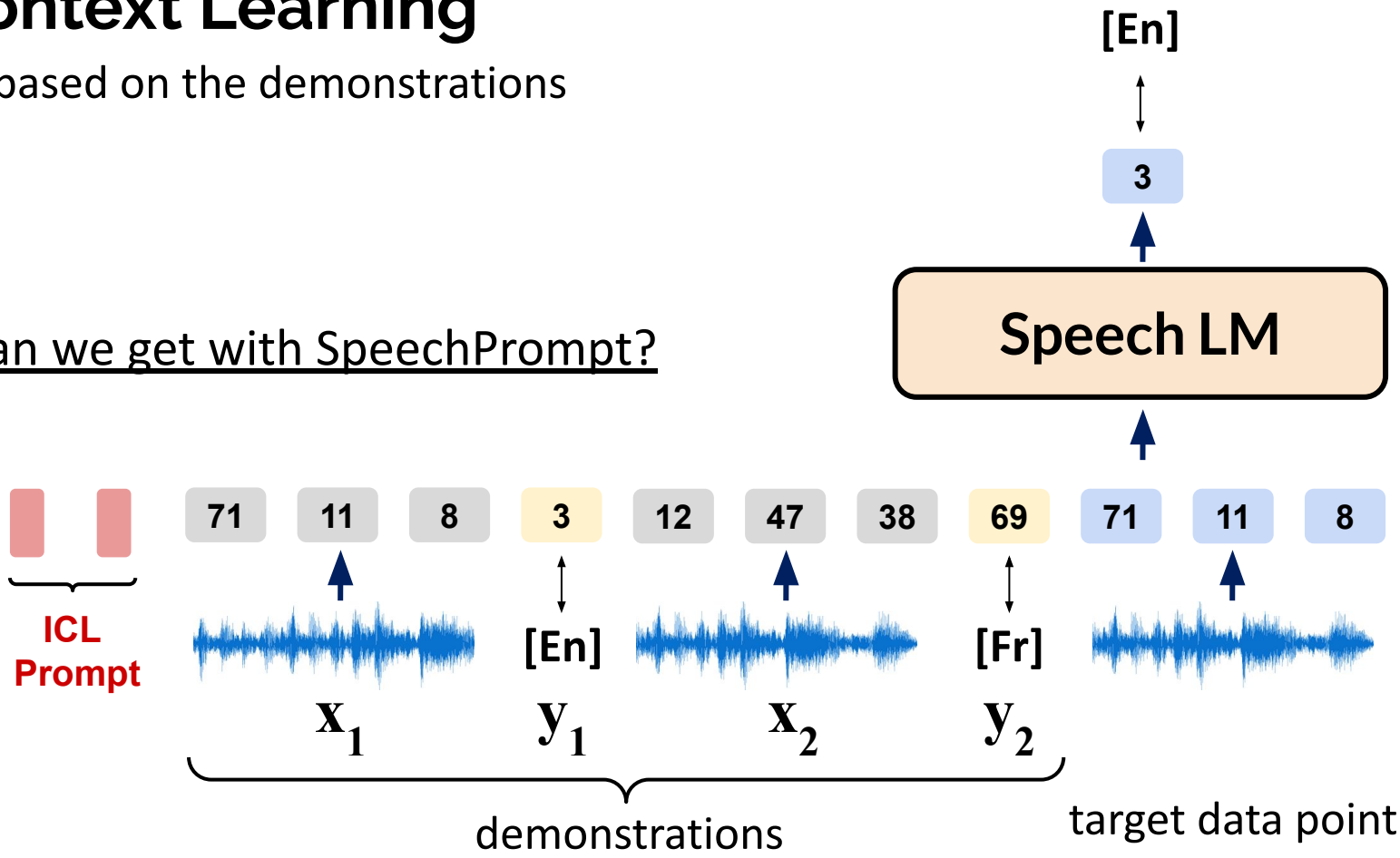
LLM can take care of random labels

Larger language models do in-context learning differently (https://arxiv.org/abs/2303.03846)

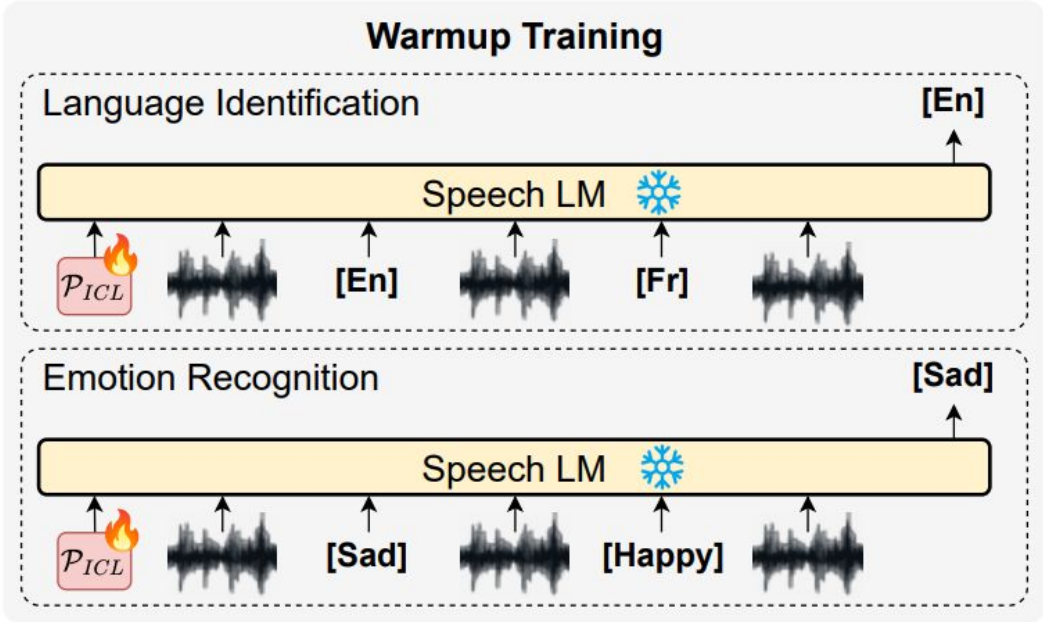
In-Context Learning

Predicting based on the demonstrations

How far can we get with SpeechPrompt?



- **Warmup Training:**
  Learn ICL prompts to enable the speech LM with ICL capability.

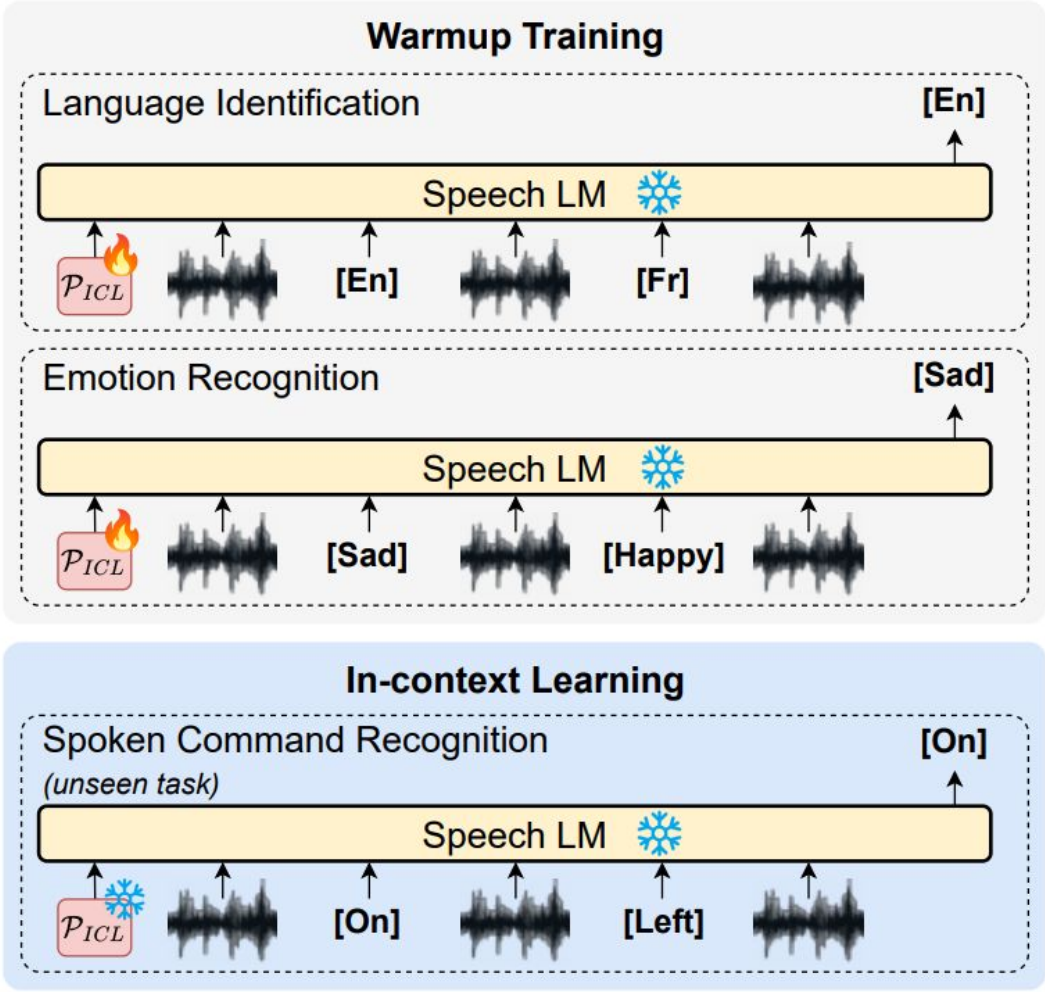


Exploring In-Context Learning of Textless Speech Language Model for Speech Classification Tasks, Interspeech 2024. (https://arxiv.org/abs/2310.12477)

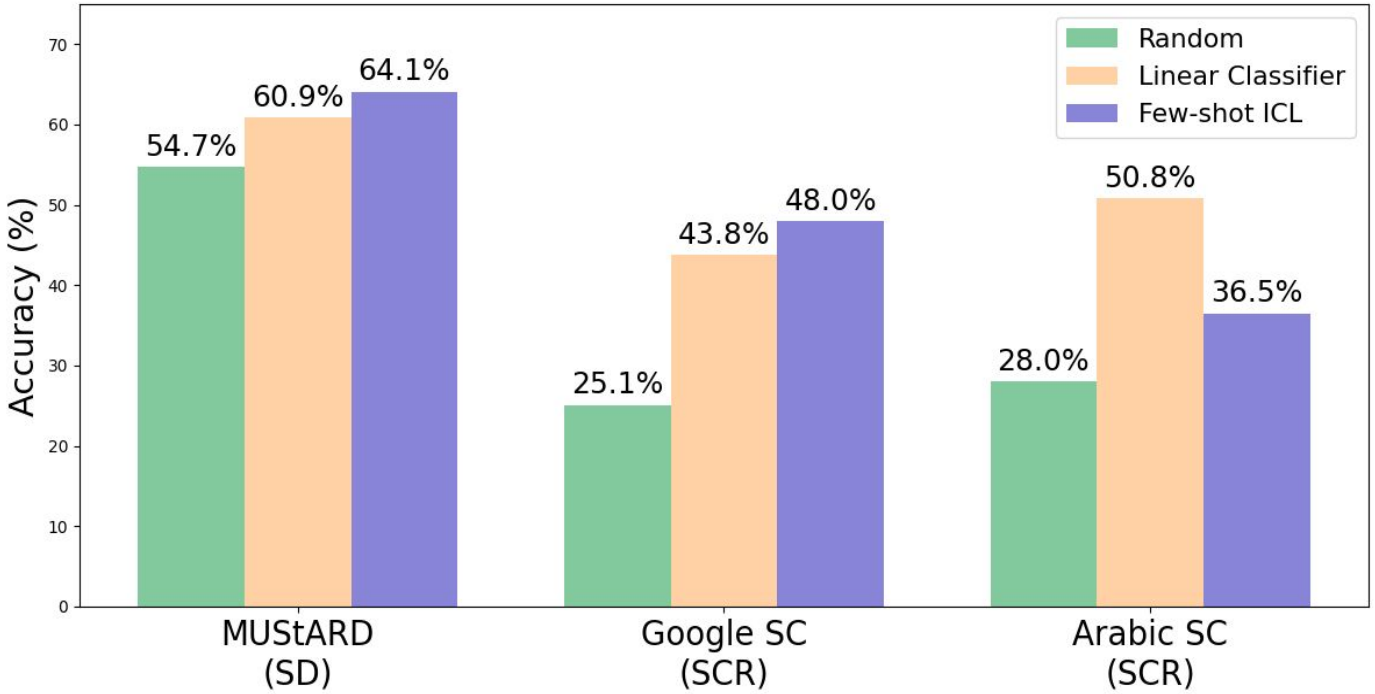
- **Warmup Training:**
  Learn ICL prompts to enable the speech LM with ICL capability.

ICL-Speech LM

- **In-context Learning**
  - The LM is fixed
  - The prompt is fixed
  - The task is unseen



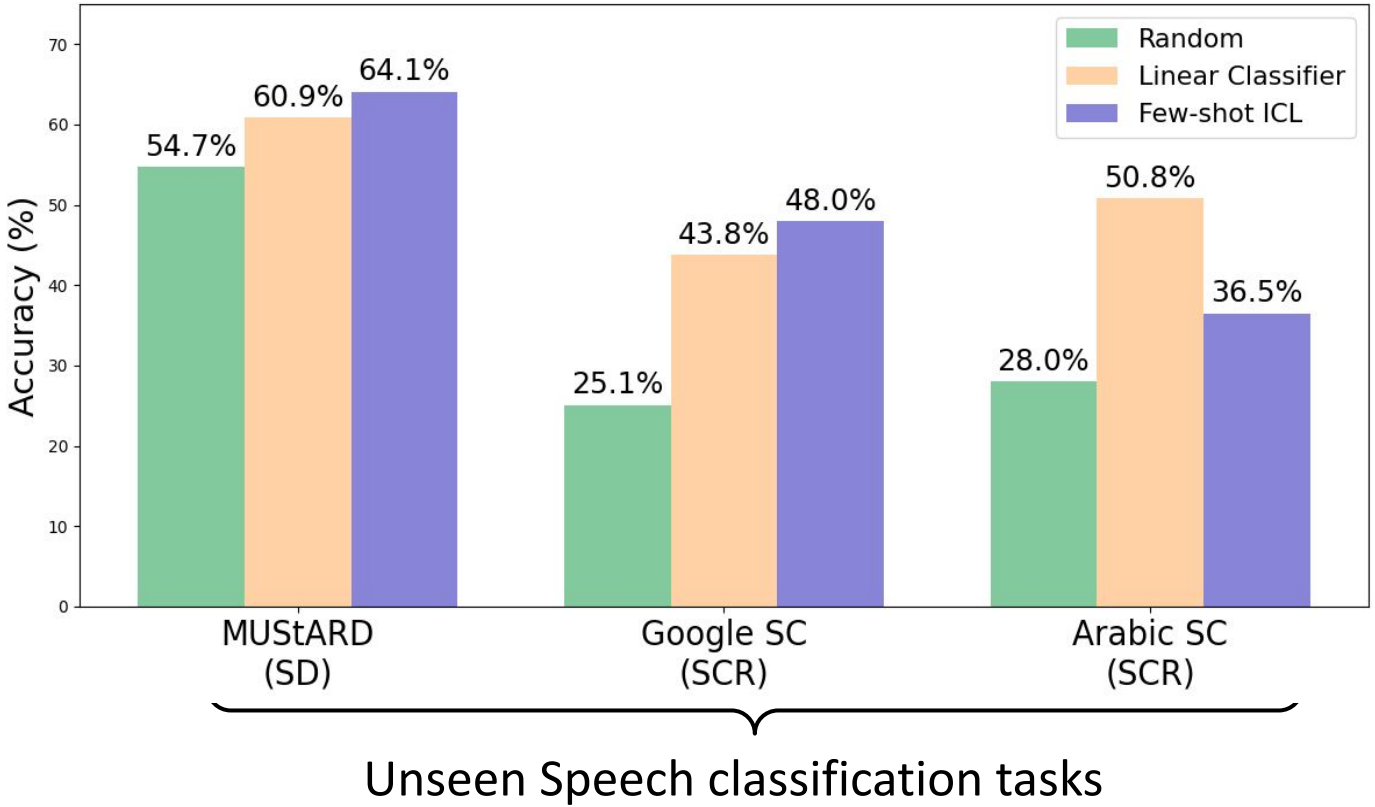
In-Context Learning on Unseen Tasks



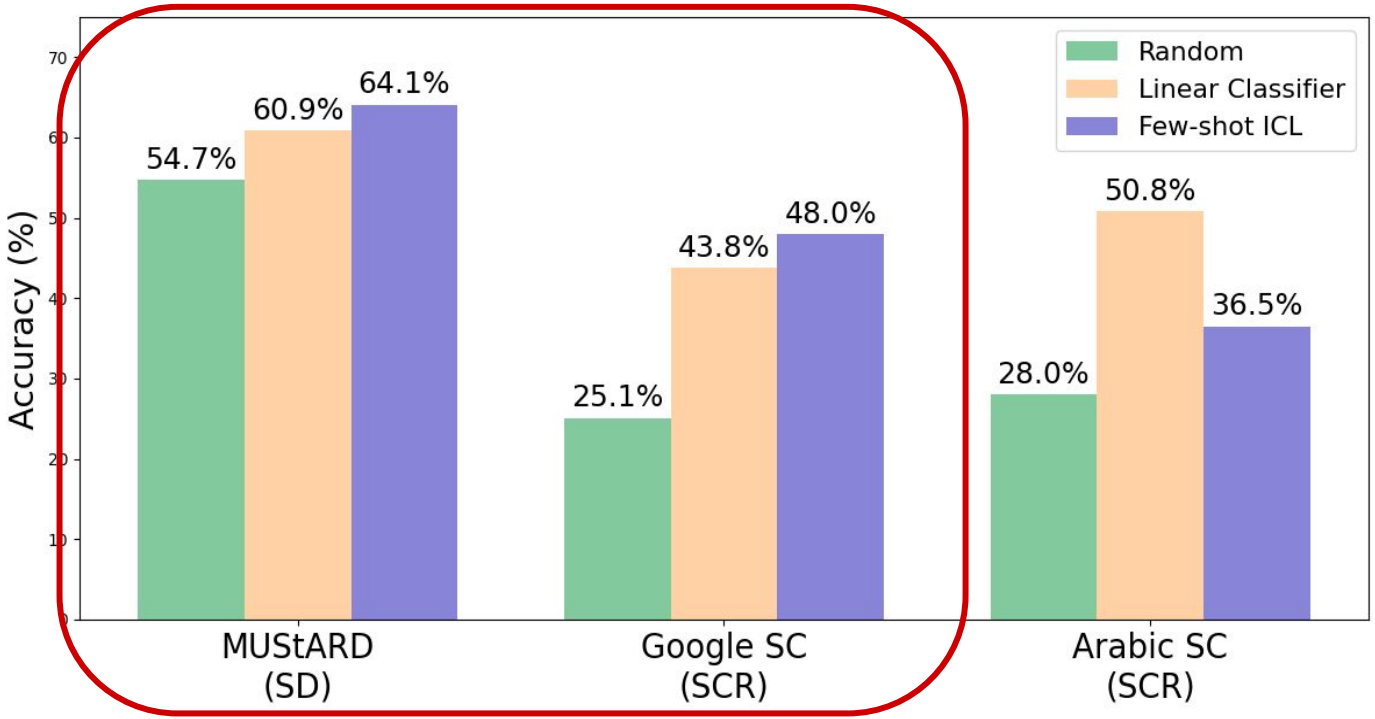
Warmup training:

Mandarin SCR, Lithuanian SCR, Language ID, Emotion Recognition

In-Context Learning on Unseen Tasks

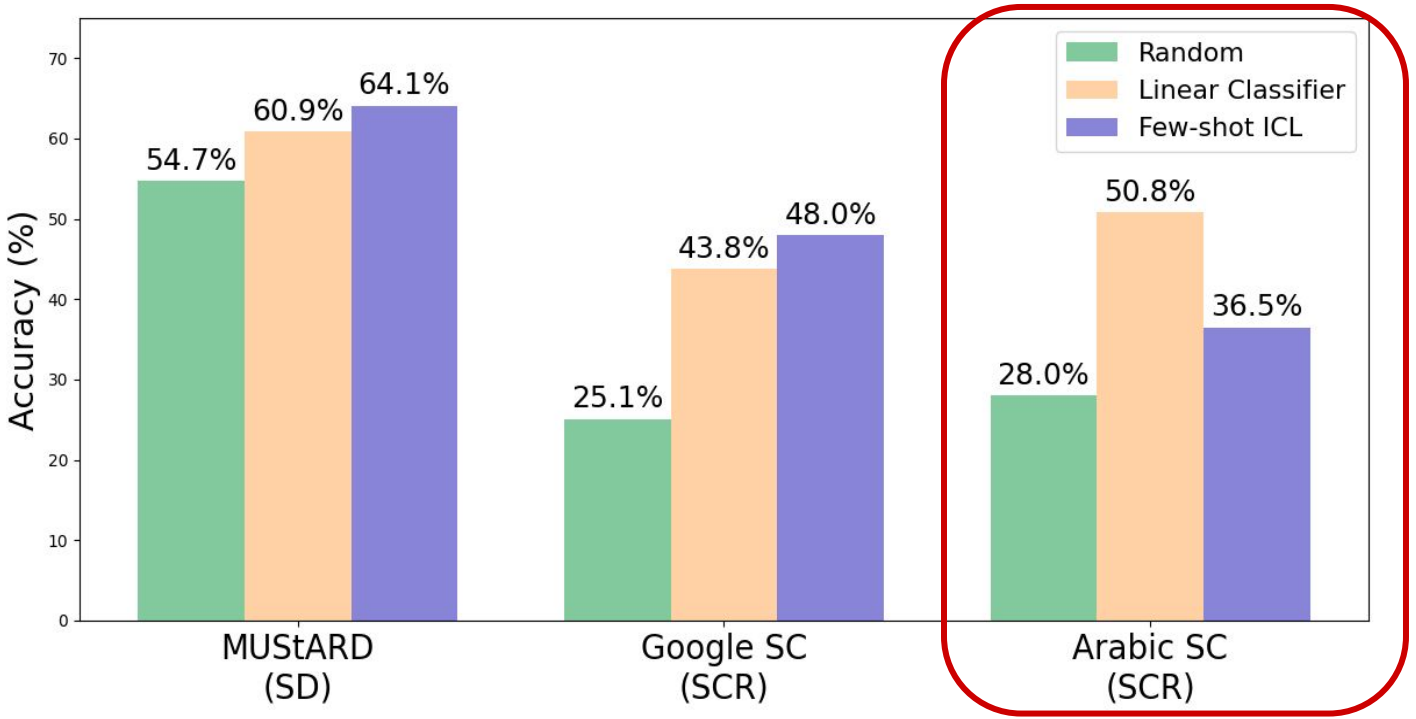


In-Context Learning on Unseen Tasks



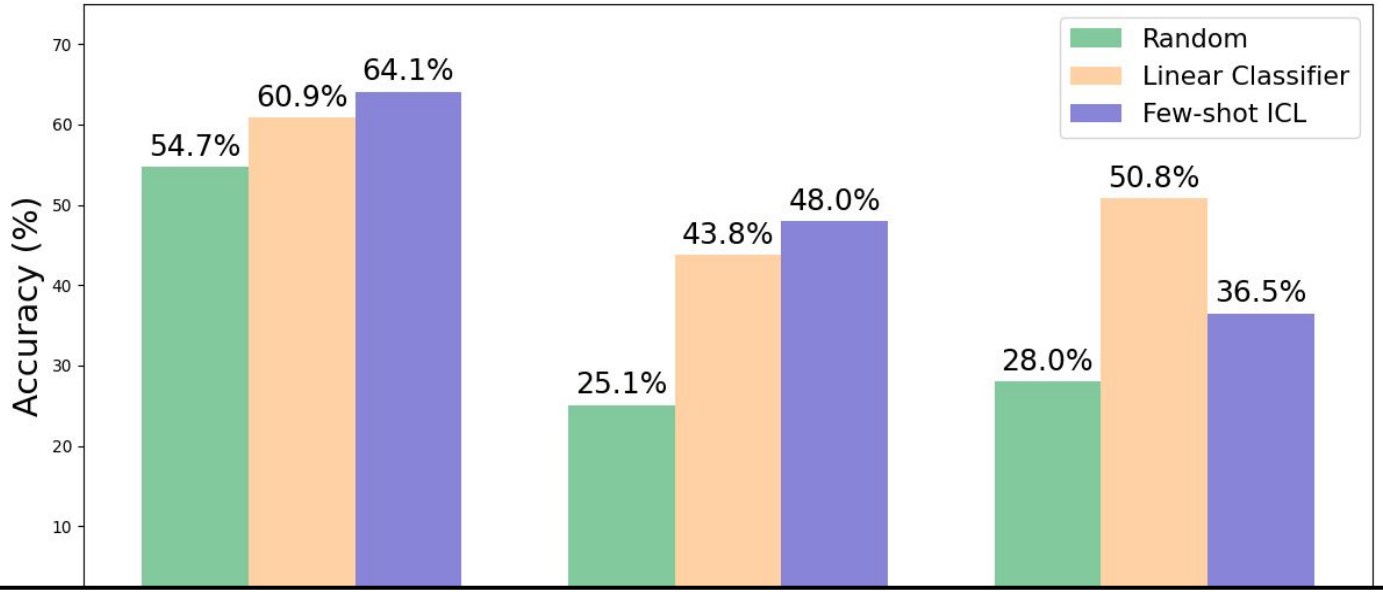
- GSLM can perform In-context Learning outperforming random guessing and linear classifier

In-Context Learning on Unseen Tasks



- Underperform linear classifer probably due to cross-lingual setting

In-Context Learning on Unseen Tasks



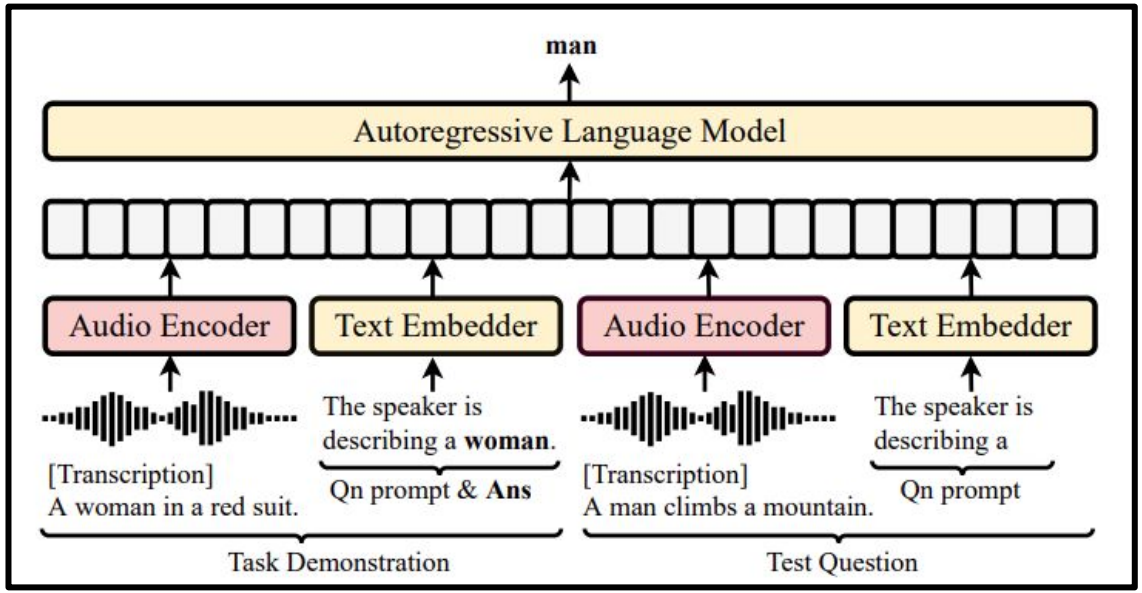
There's still a big performance gap between the simple supervised models.
Surprising to get a non trivial result.

- GPT-3 ~ 170B parameters
- GSLM ~ 150M parameters + prompts (0.2M)

Related Works

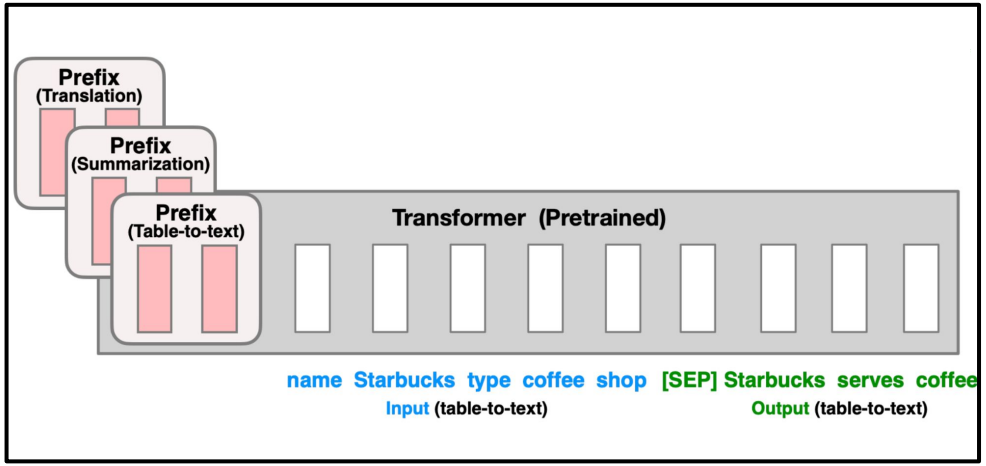
WavPrompt

- Pioneer of prompting for speech processing
- Audio Encoder + Text language model (GPT-2)
- In-context Learning for spoken language understanding tasks



WAVPROMPT: Towards Few-Shot Spoken Language Understanding with Frozen Language Models, *Heting Gao, Junrui Ni, Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson* (Interspeech 2022, https://arxiv.org/abs/2203.15863)

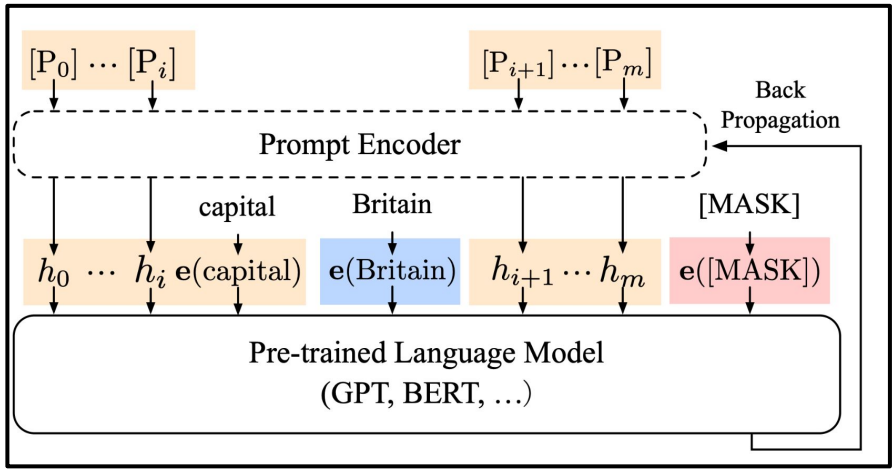
Prefix-Tuning



- Deep prompts for Natural Language Generation

Prefix-Tuning: Optimizing Continuous Prompts for Generation, Xiang Lisa Li, Percy Liang
(ACL 2021, https://arxiv.org/abs/2101.00190)

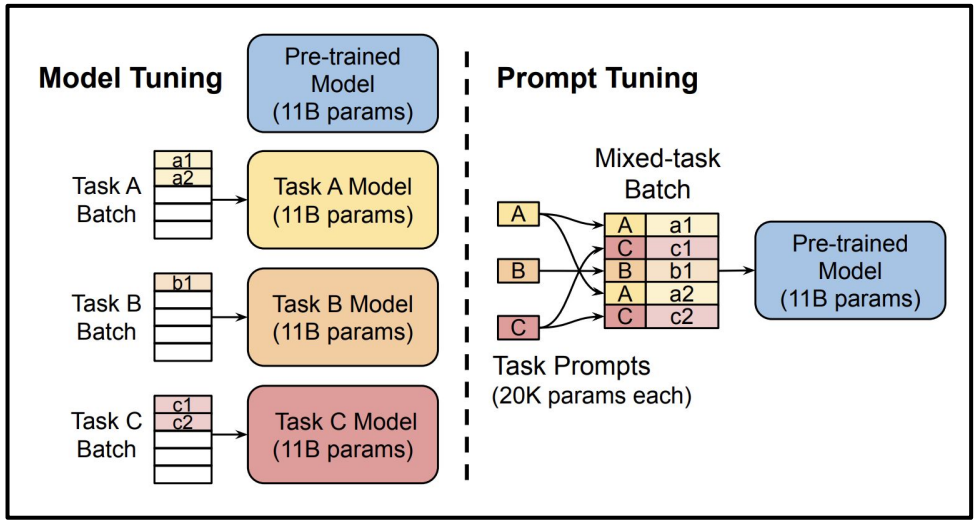
P-Tuning



- Deep prompts for Natural Language Understanding

GPT Understands, Too, Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, Jie Tang
(https://arxiv.org/abs/2103.10385)

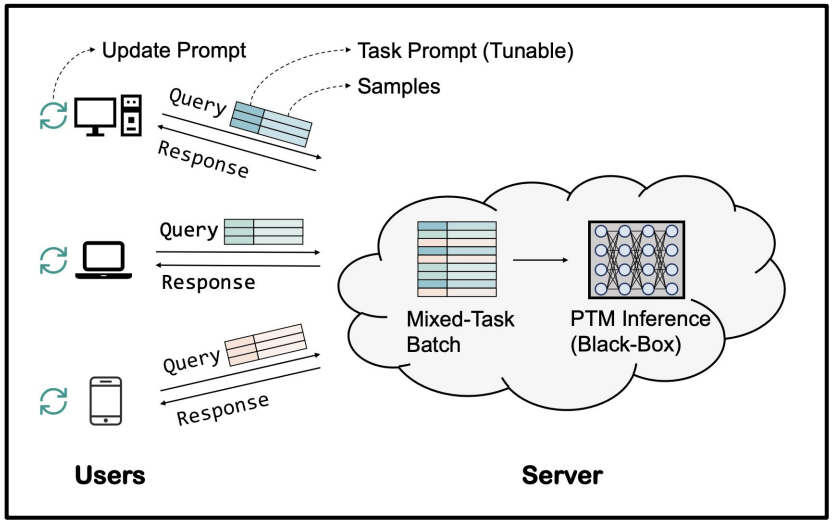
The Power of Scale for Parameter-Efficient Prompt Tuning



- Mixed-task Batching

The Power of Scale for Parameter-Efficient Prompt Tuning, Brian Lester, Rami Al-Rfou, Noah Constant (EMNLP 2021, https://arxiv.org/abs/2104.08691)

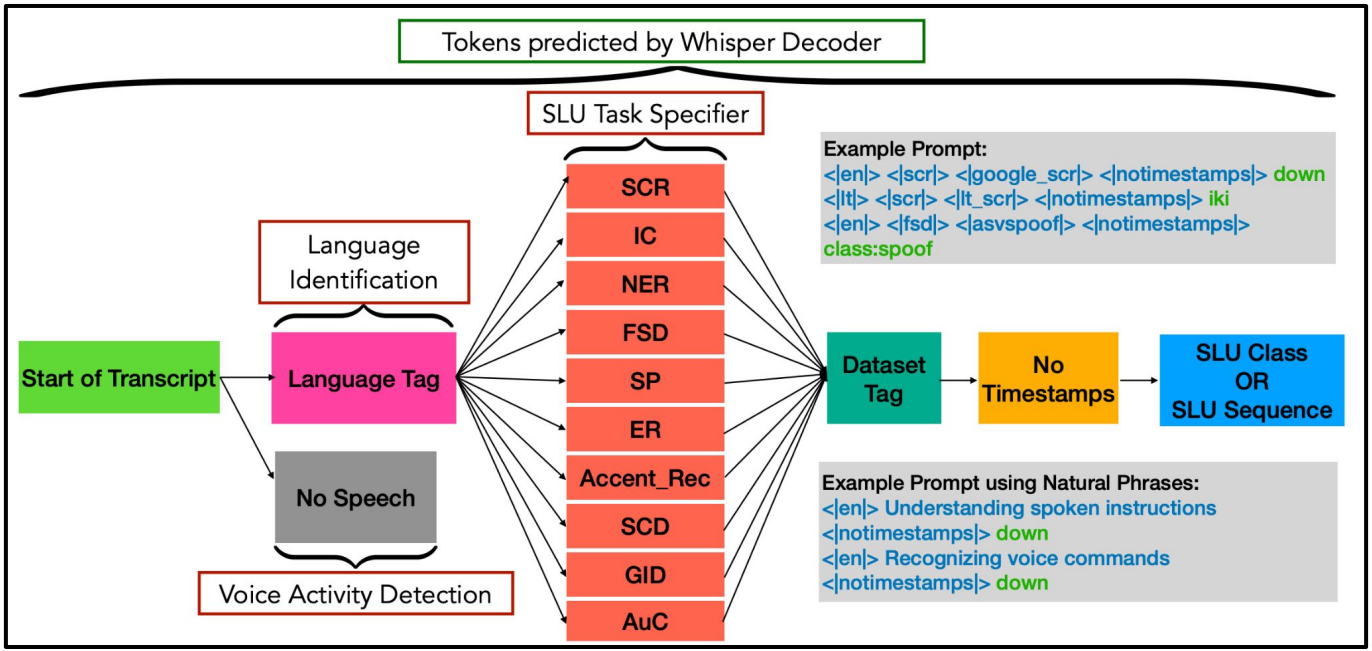
Black-box tuning



- Language-Model as a Service (LMaaS)

Black-Box Tuning for Language-Model-as-a-Service, Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, Xipeng Qiu (ICML 2022, https://arxiv.org/abs/2201.03514)

UniverSLU



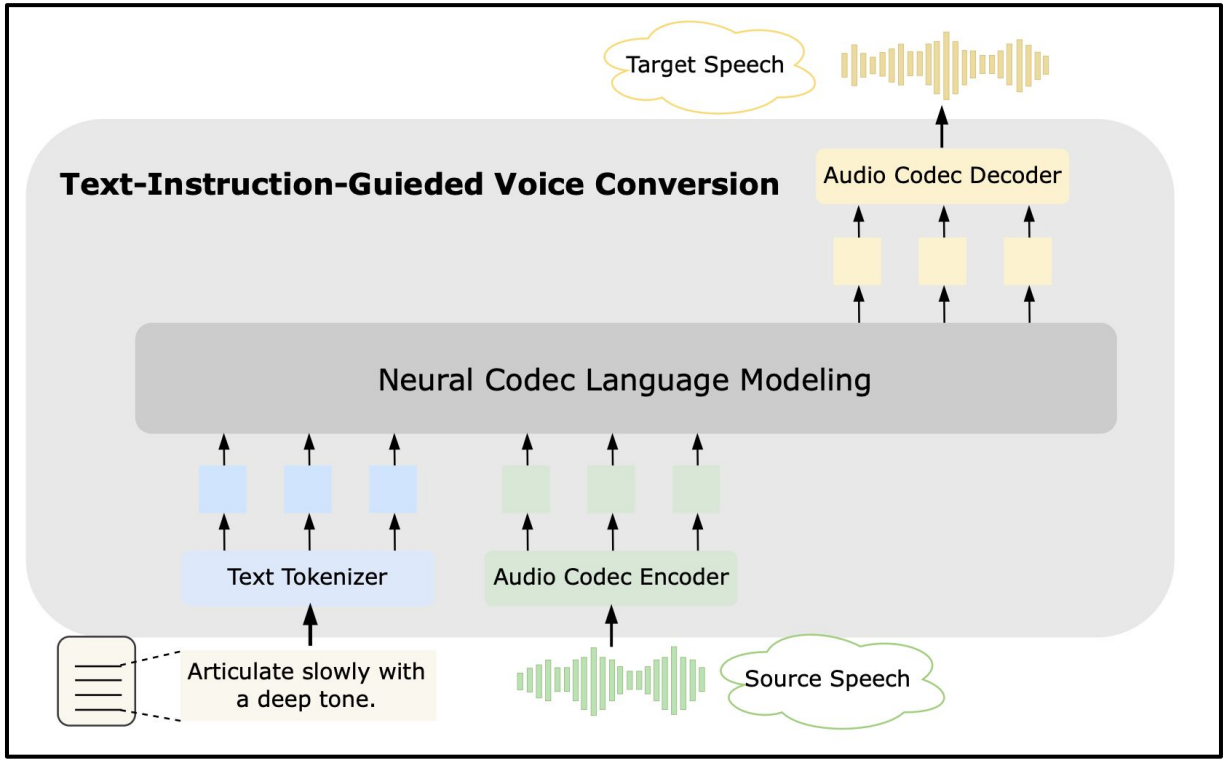
- Design task specifier for prompting Whisper
- For SLU, promoting Whisper is better than SpeechPrompt

UniverSLU: Universal Spoken Language Understanding for Diverse Tasks with Natural Language Instructions, Siddhant Arora, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan Sharma, Yosuke Kashiwagi, Emiru Tsunoo, Karen Livescu, Shinji Watanabe (NAACL 2024, https://arxiv.org/abs/2310.02973)

Natural Language as Instruction

- Natural language instruction for speech LM
- Condition the speech LM on instruction and an input

Instruct-VC

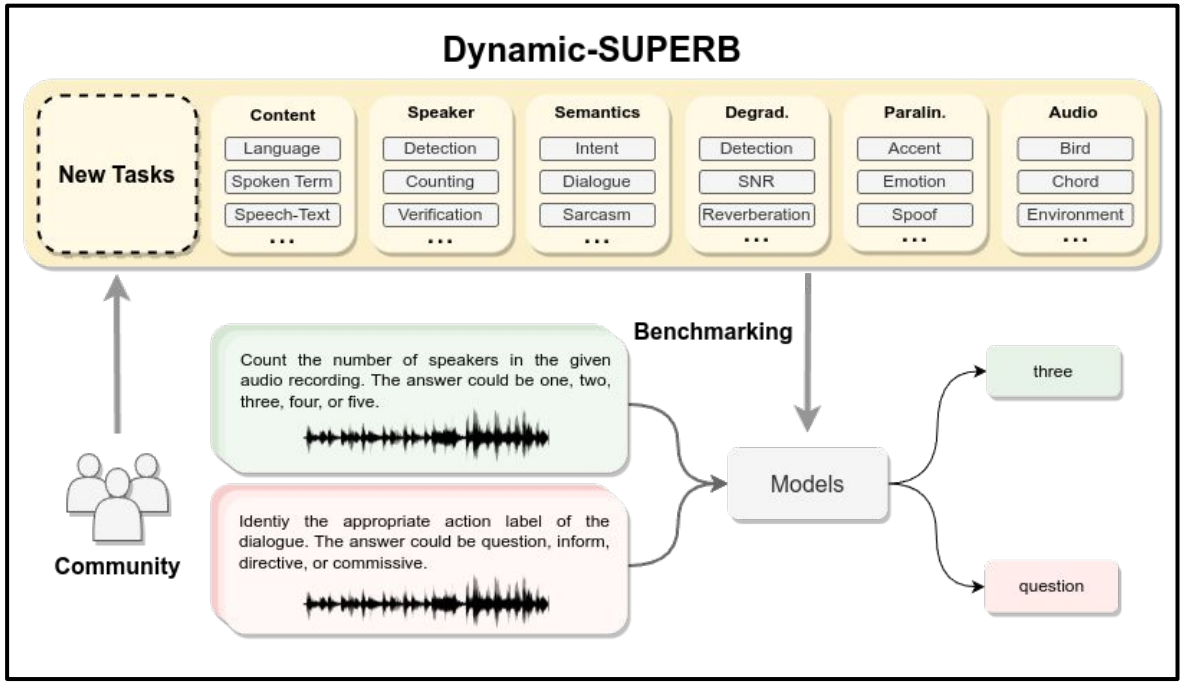


Towards General-Purpose Text-Instruction-Guided Voice Conversion, Chun-Yi Kuan, Chen An Li, Tsu-Yuan Hsu, Tse-Yang Lin, Ho-Lam Chung, Kai-Wei Chang, Shuo-yiin Chang, Hung-yi Lee (ASRU 2023, https://arxiv.org/abs/2309.14324) demo: https://text-guided-vc.github.io/asru2023_demo/

Natural Language as Instruction

- Instruction-tuning benchmark for speech models
- 55 tasks: Audio, content, paralinguistics, semantic, degradation, speaker

Dynamic-SUPERB

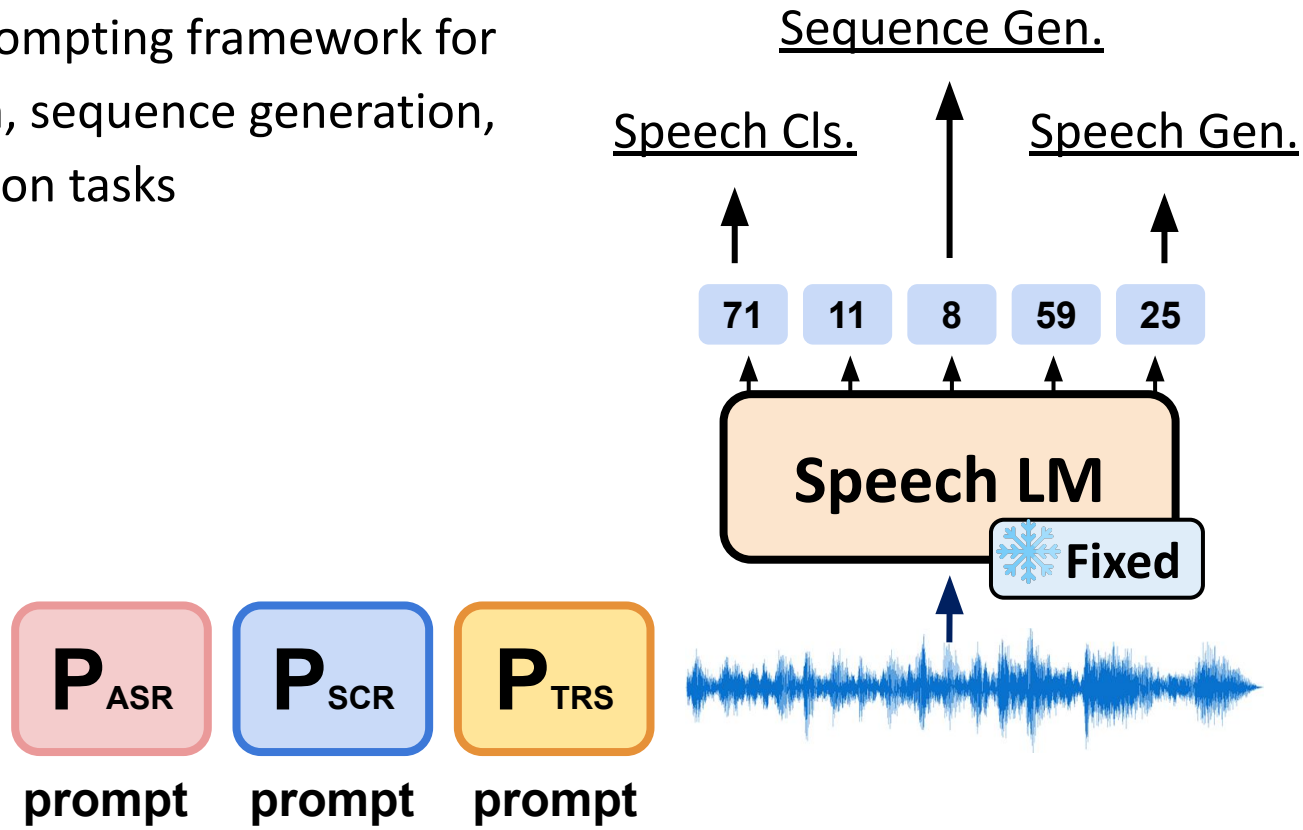


Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech, Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, Hung-yi Lee (ICASSP 2024, https://arxiv.org/abs/2309.09510) website: https://dynamic-superb.github.io/

Conclusion

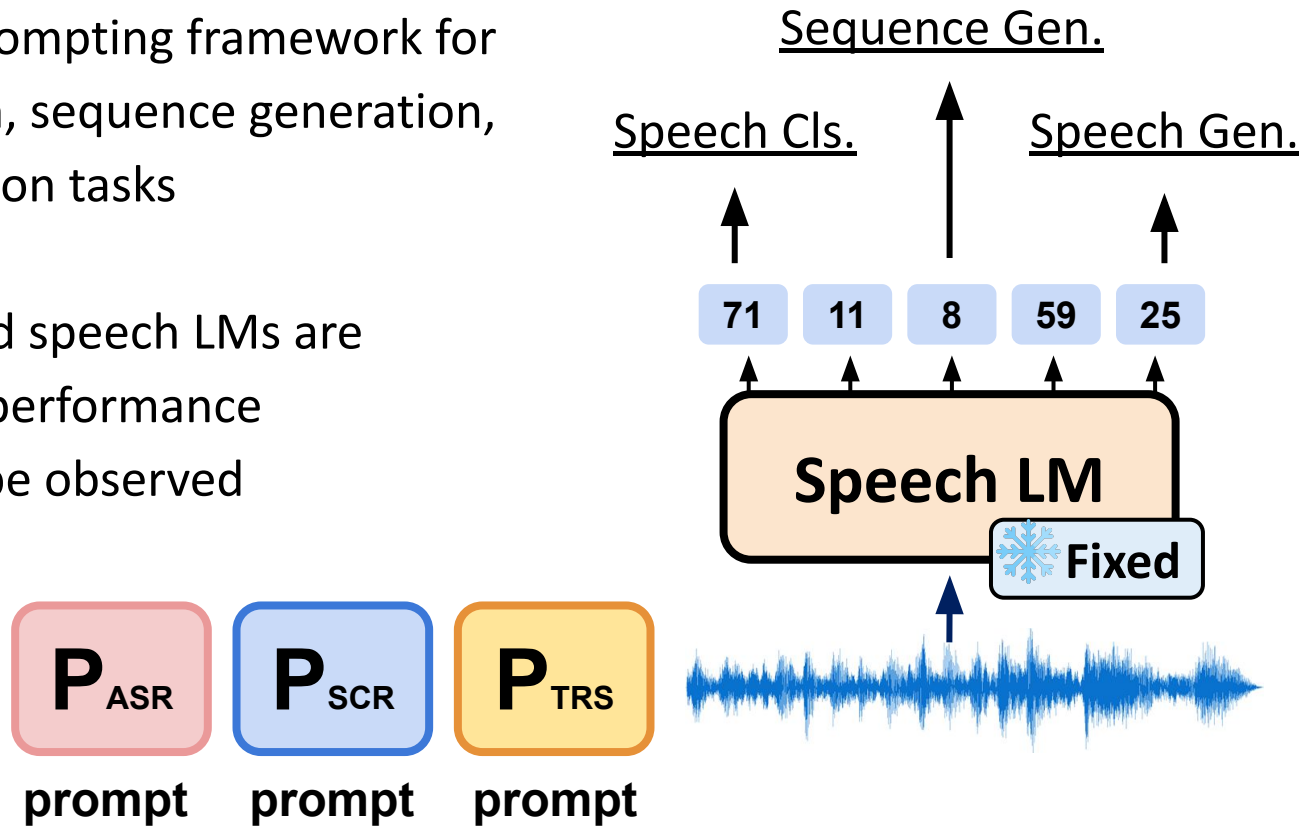
Conclusion

- Achieve a unified prompting framework for speech classification, sequence generation, and speech generation tasks

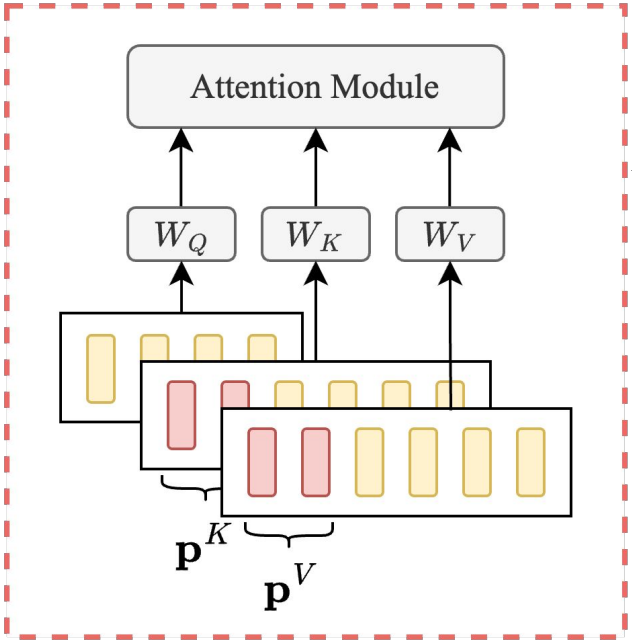


Conclusion

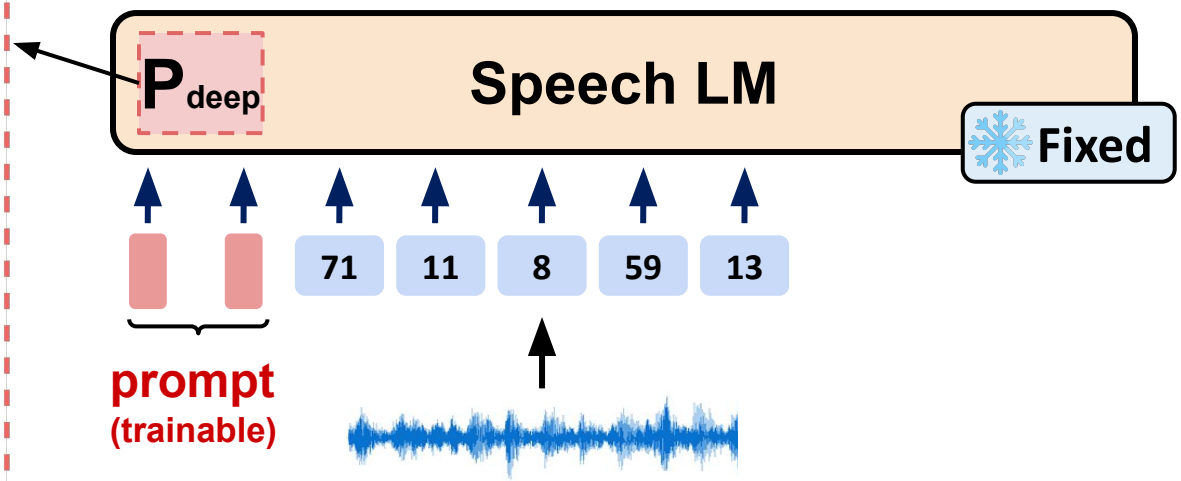
- Achieve a unified prompting framework for speech classification, sequence generation, and speech generation tasks
- With more advanced speech LMs are developed, further performance improvements can be observed



Future Works

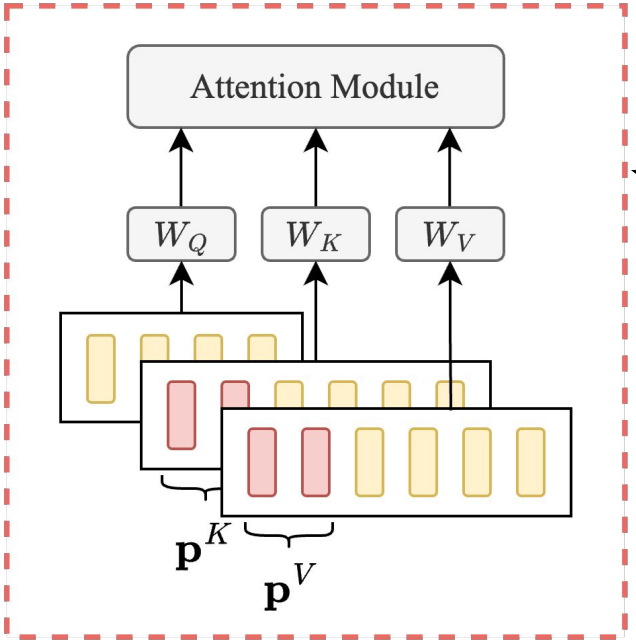


Deep Prompts guiding the attention mechnism

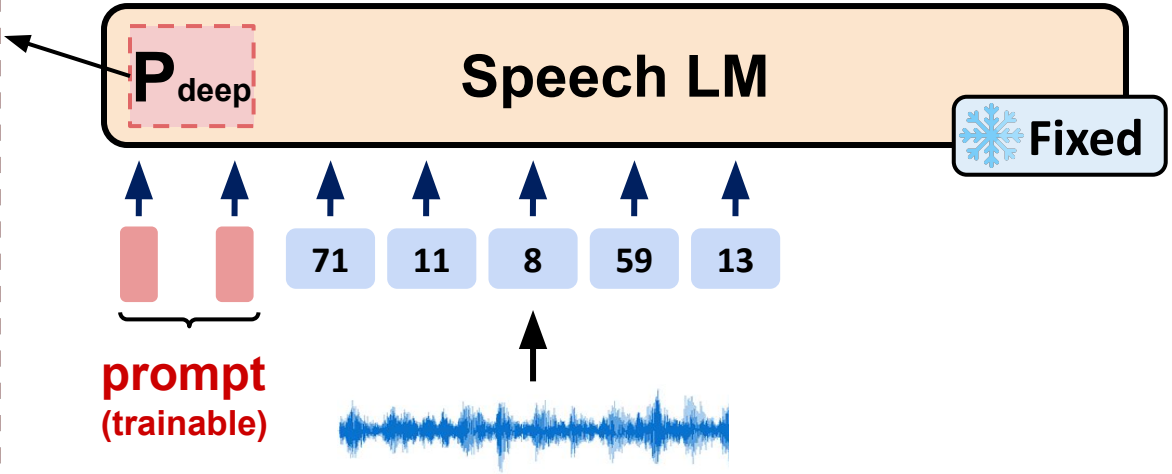


Input Prompts

- Prompting is an effective and efficient method to guide the speech LM



Deep Prompts guiding the attention mechnism



Input Prompts

- Prompting is an effective and efficient method to guide the speech LM
- Limitation: (1) Don't understand natural language (2) The generated speech may lack competitive semantic quality.

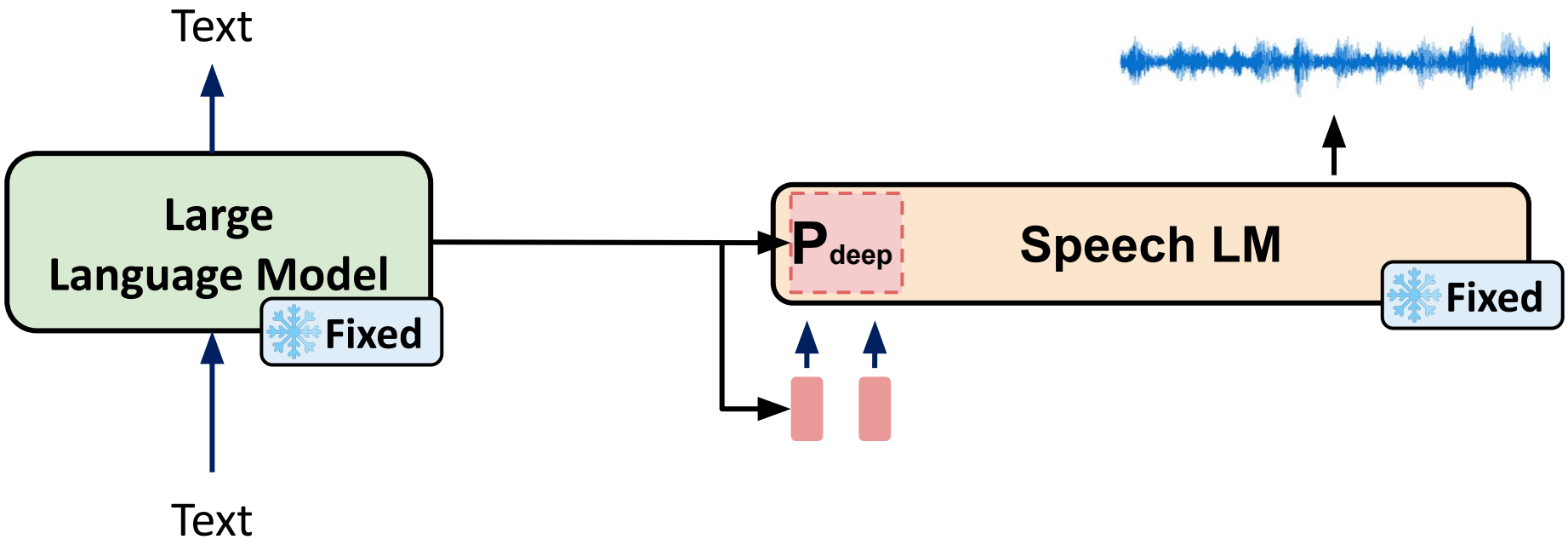
LLM Guided Speech Generation

Large Language Model

- Possible solution: Introducing LLM
- Strong natural langauge understanding and generation capability

LLM Guided Speech Generation

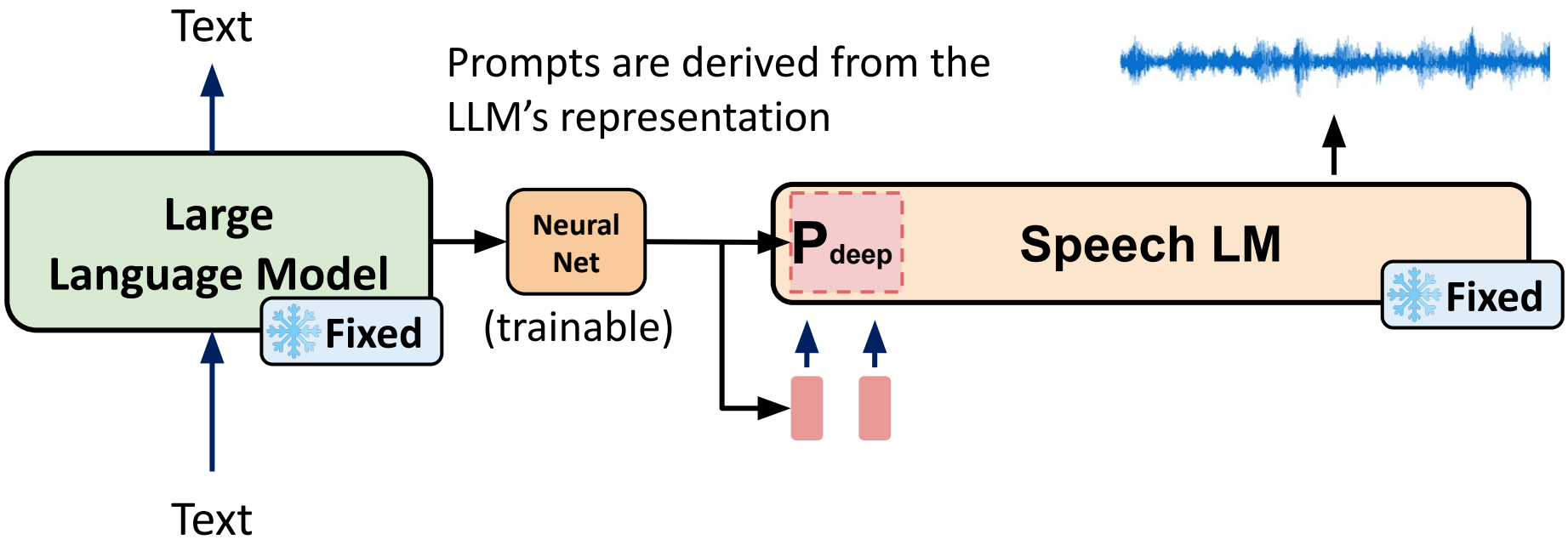
Prompts are derived from LLMs



Maintain the generation ability of both the text LM and speech LM

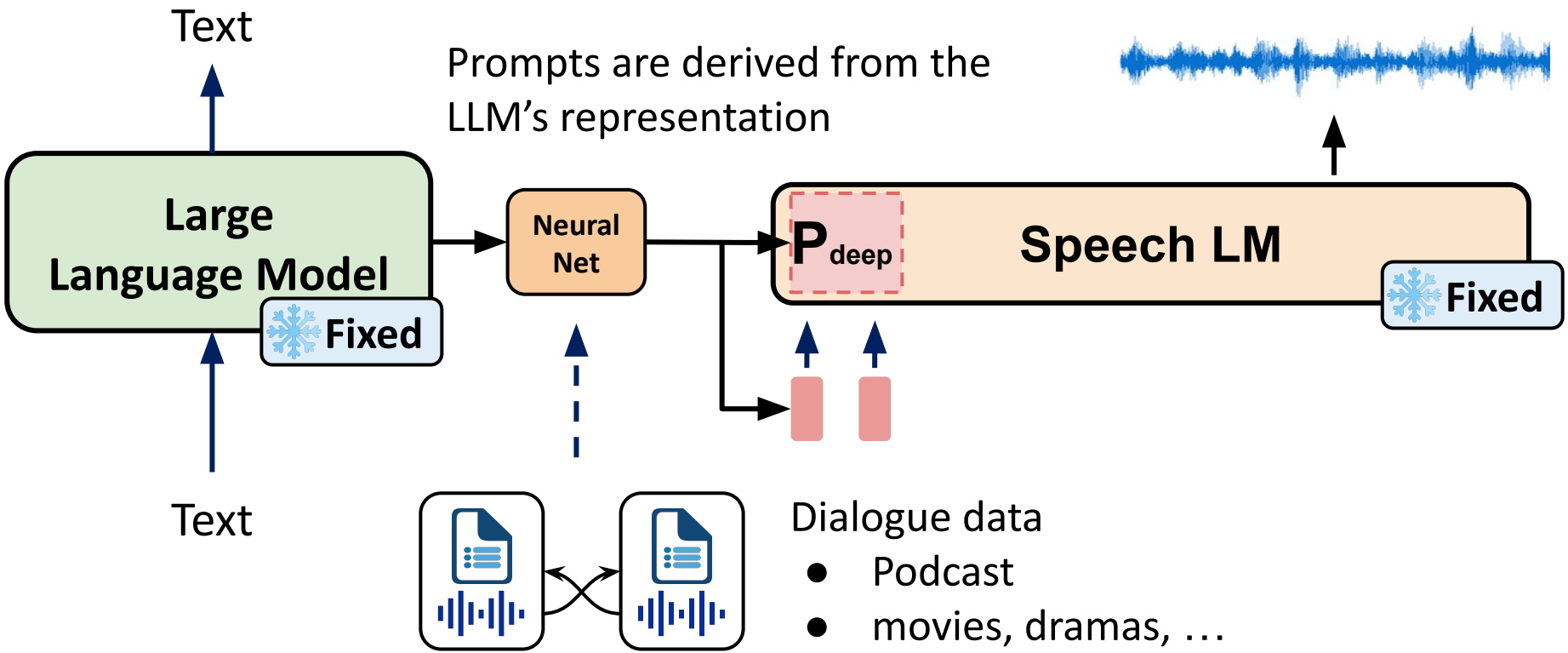
LLM Guided Speech Generation

Prompts are derived from LLMs



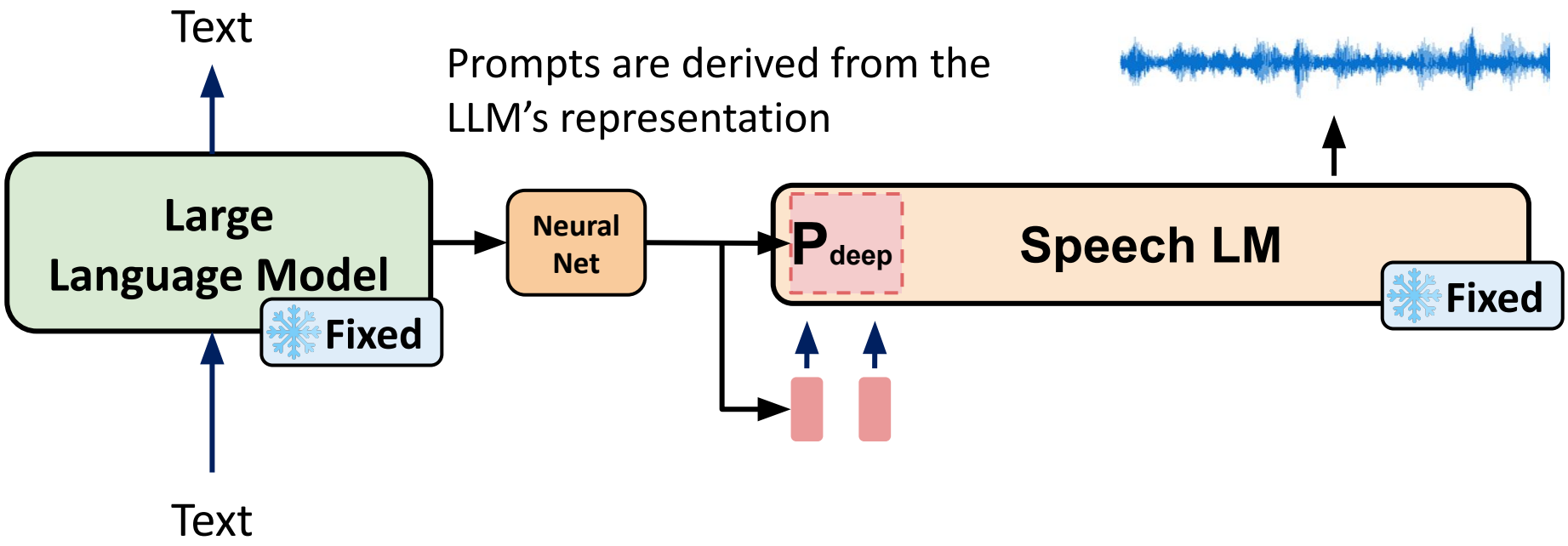
LLM Guided Speech Generation

Prompts are derived from LLMs



LLM Guided Speech Generation

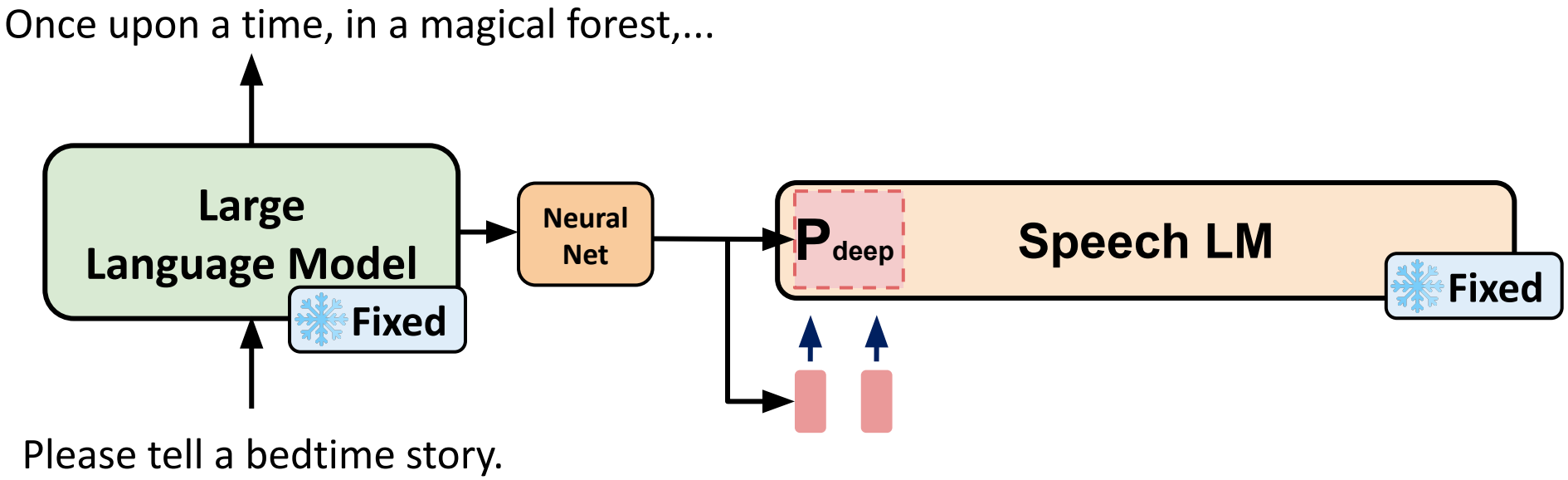
Prompts are derived from LLMs



Zipper: A Multi-Tower Decoder Architecture for Fusing Modalities (https://arxiv.org/pdf/2405.18669)

LLM Guided Speech Generation

e.g. Speech version Chatbot

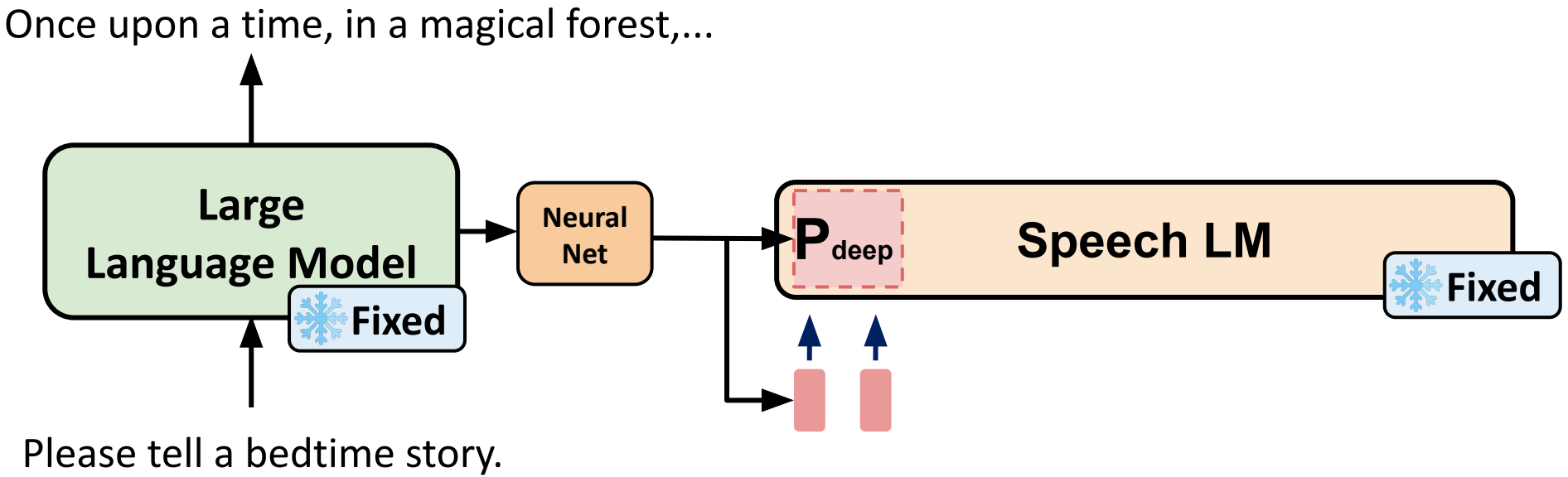


LLM Guided Speech Generation

e.g. Speech version Chatbot

LLM focuses on:

- **Content**

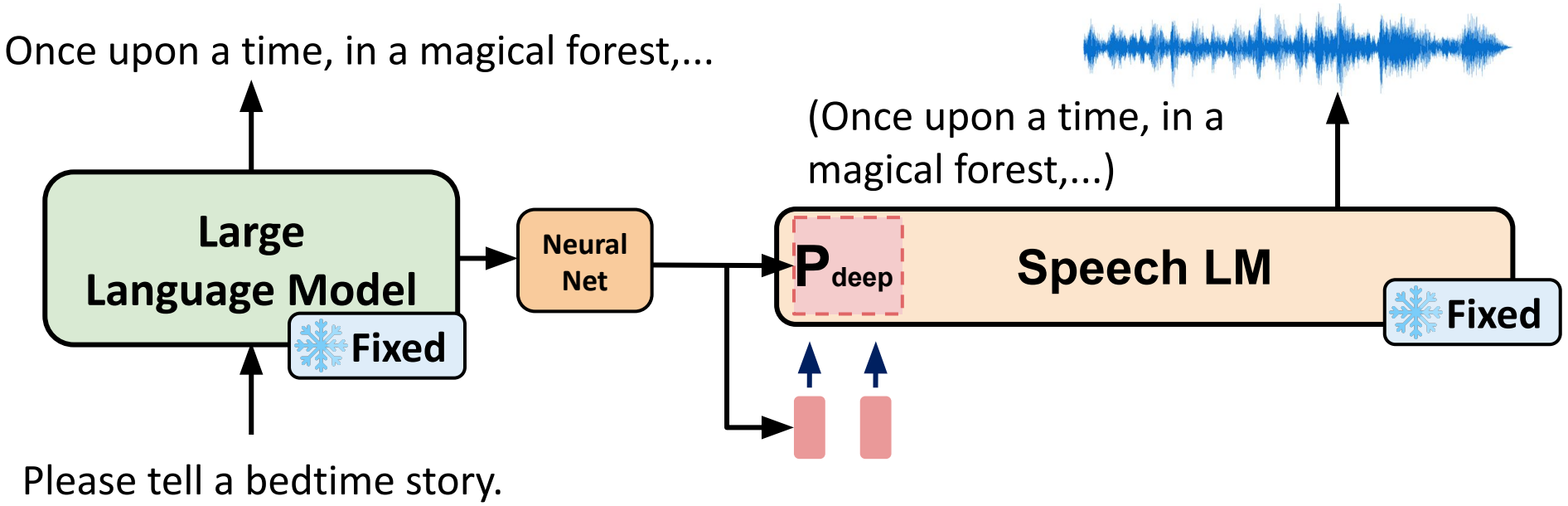


LLM Guided Speech Generation

e.g. Speech version Chatbot

LLM focuses on:

- **Content**



LLM Guided Speech Generation

e.g. Speech version Chatbot

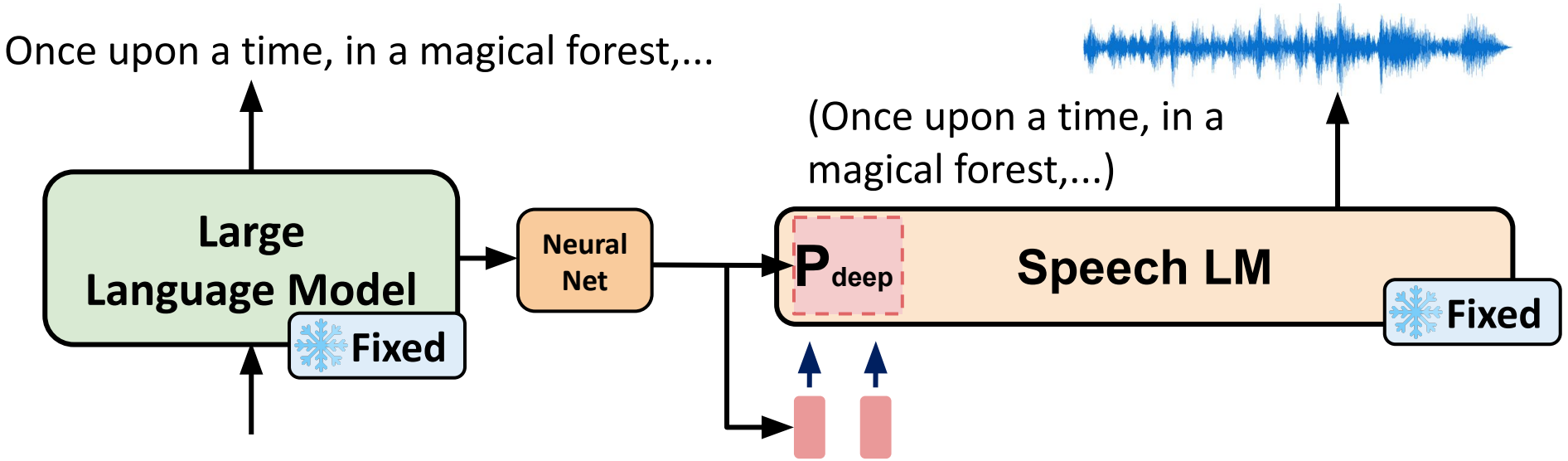
Speech LM focuses on :

- **Prosody, intonation,...**

LLM focuses on:

- **Content**

Soft and gentle tone



LLM Guided Speech Generation

e.g. Speech LM with Persona

Speech LM focuses on :

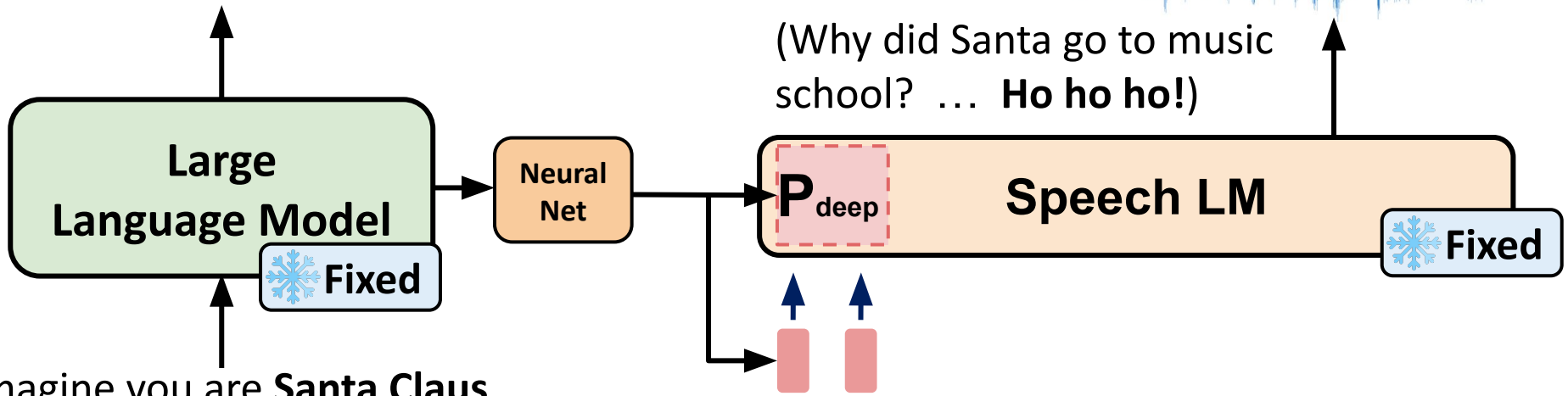
- **Prosody, intonation,...**
- **Laughter, sigh, …**

LLM focuses on:

- **Content**

Why did Santa go to music school? To improve his "wrap" skills! **Ho ho ho!**

(Why did Santa go to music school? … **Ho ho ho!**)



Imagine you are **Santa Claus**. Please tell me a joke.

References

[1] SpeechPrompt: Prompting Speech Language Models for Speech Processing Tasks

(Accepted to IEEE/ACM Transactions on Audio Speech and Language Processing, TASLP)

Kai-Wei Chang, *Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee*

[2] SpeechPrompt: An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks

(Interspeech 2022)

Kai-Wei Chang, *Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee*

[3] SpeechPrompt v2: Prompt Tuning for Speech Classification Tasks

(arXiv Preprint)

Kai-Wei Chang, *Yu-Kai Wang, Hua Shen, Iu-thing Kang, Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee*

[4] SpeechGen: Unlocking the Generative Power of Speech Language Models with Prompts

(arXiv Preprint)

Kai-Wei Chang, *Haibin Wu, Yuan-Kuei Wu, Hung-yi Lee*

[5] Prompting and Adapter Tuning for Self-supervised Encoder-Decoder Speech Model

(ASRU 2023)

Kai-Wei Chang, *Ming-Hsin Chen, Yun-Ping Lin, Jing Neng Hsu, Chien-yu Huang, Shang-Wen Li, Hung-Yi Lee*

[6] Exploring In-Context Learning of Textless Speech Language Model for Speech Classification Tasks

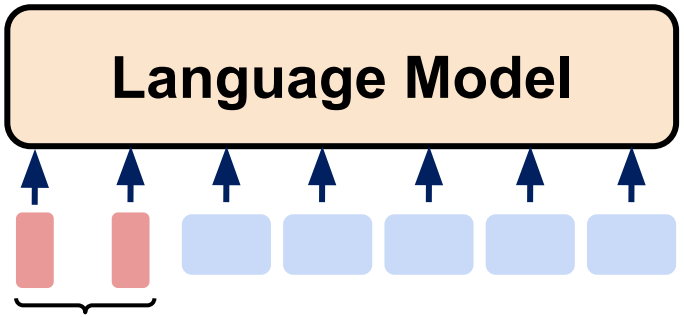
(Interspeech 2024)

Kai-Wei Chang, *Ming-Hao Hsu, Shang-Wen Li, Hung-yi Lee*

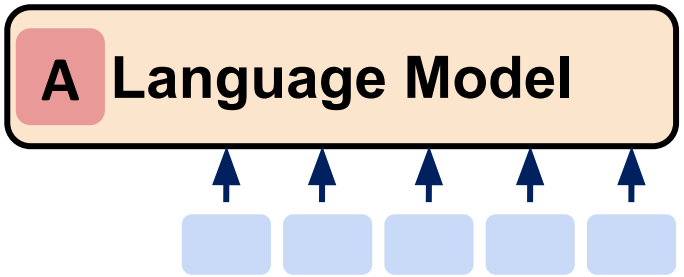
Thanks for your listening

FAQ

1. Prompting vs. Adapter Tuning

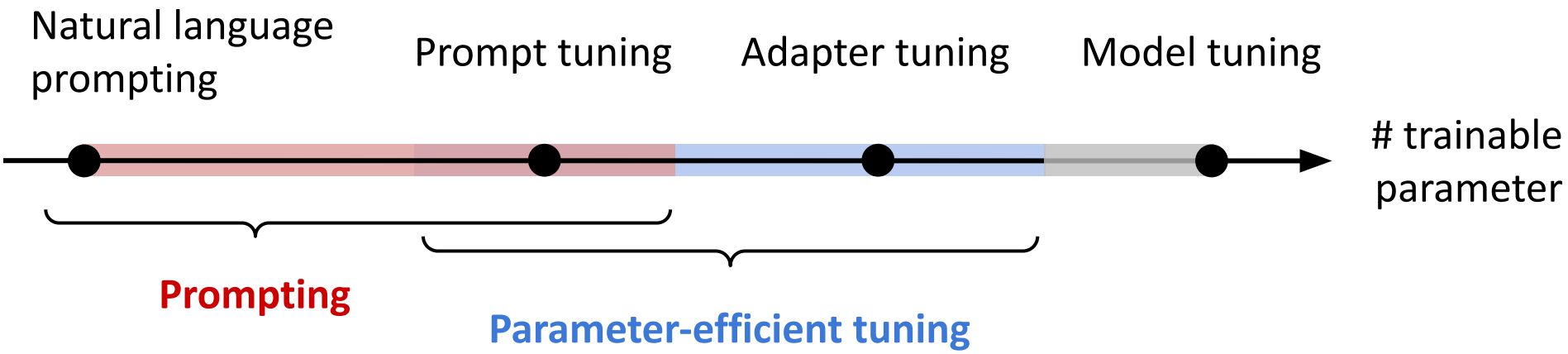


- Contain less trainable parameters
- Put at the input of the model
- Easier to achieve mixed-task batching



- Contain more trainable parameters
- Adapter modifies the model architecture

1. Prompting vs. Adapter Tuning

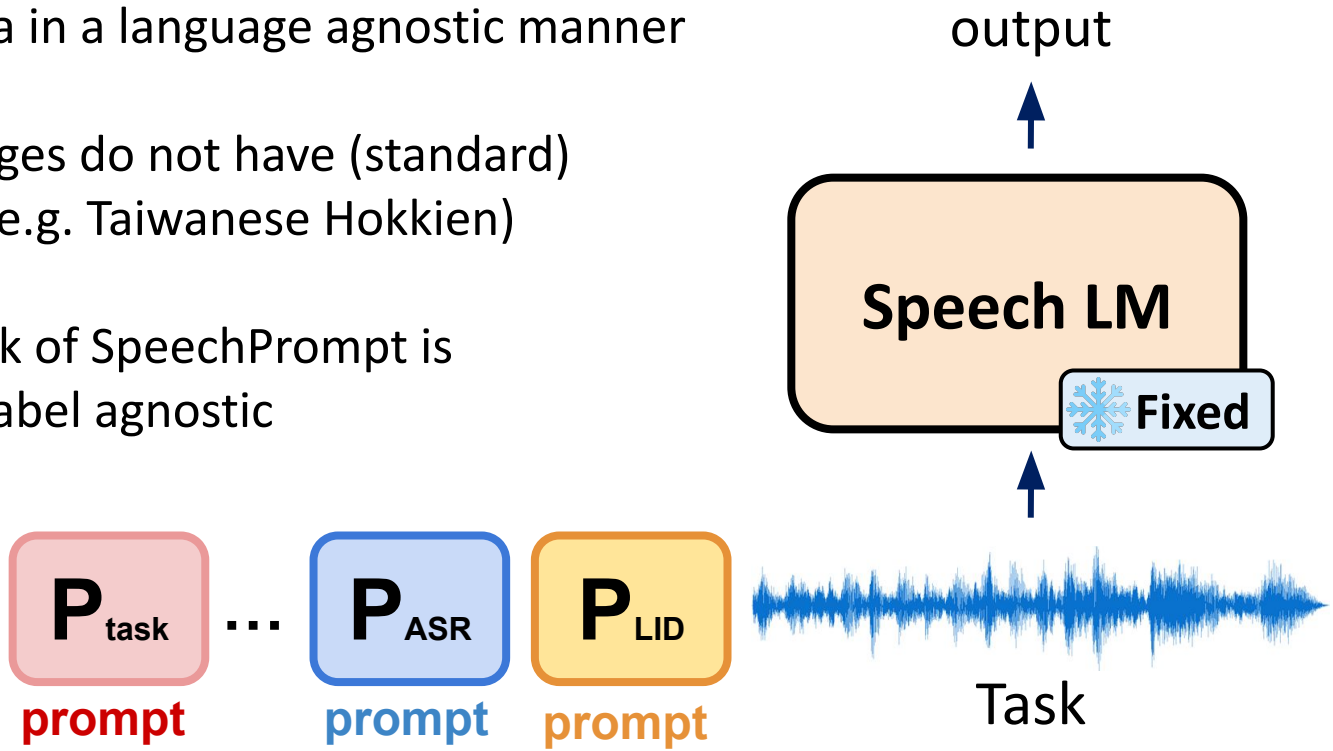


2. Why Textless Langauge Model

1. Speech LM can be pre-trained with lots unlabeled data in a language agnostic manner

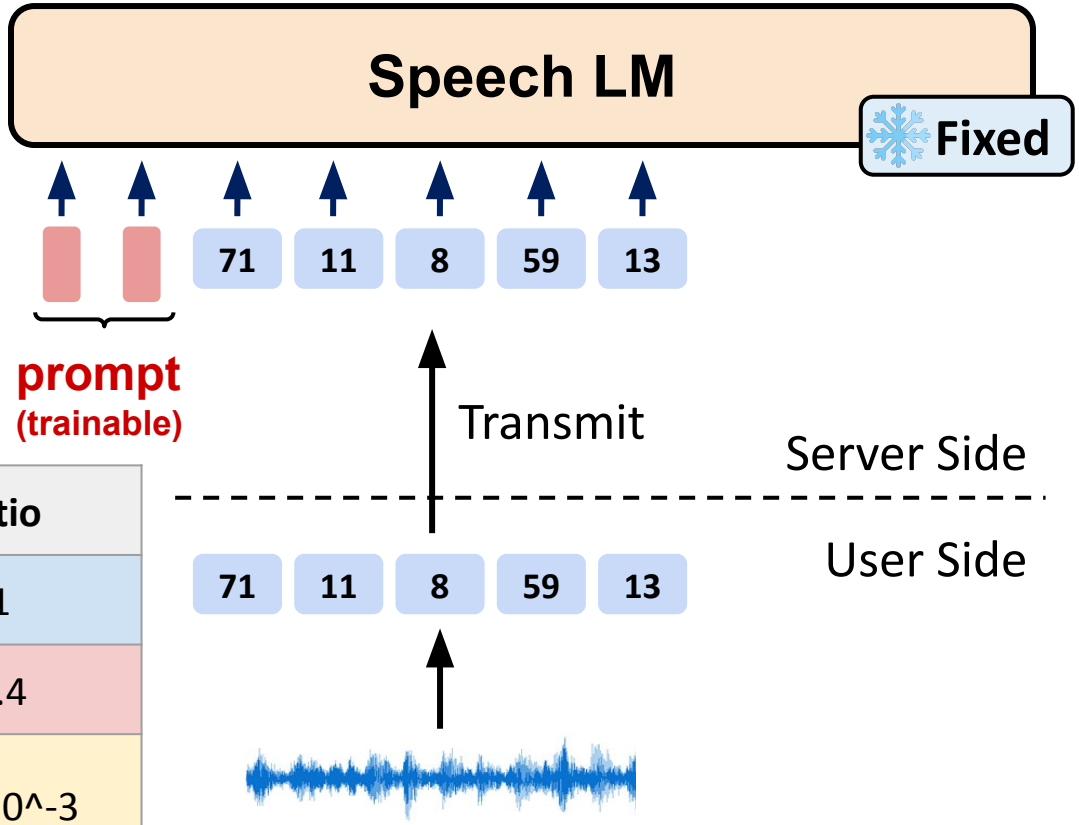
2. Lots of languages do not have (standard) written form (e.g. Taiwanese Hokkien)

3. The framework of SpeechPrompt is downstream label agnostic



2. Why Textless Langauge Model

- <u>Speech LM takes discrete units as input</u>
- storage and transmission efficiency



Data format	Data size (bits)	ratio
Raw waveform	16 x 16000 x T	1
HuBERT repre.	32 x 1024 x 50 x T	6.4
Discrete units (1,000 clusters)	10 x 50 x T	2 x 10^-3