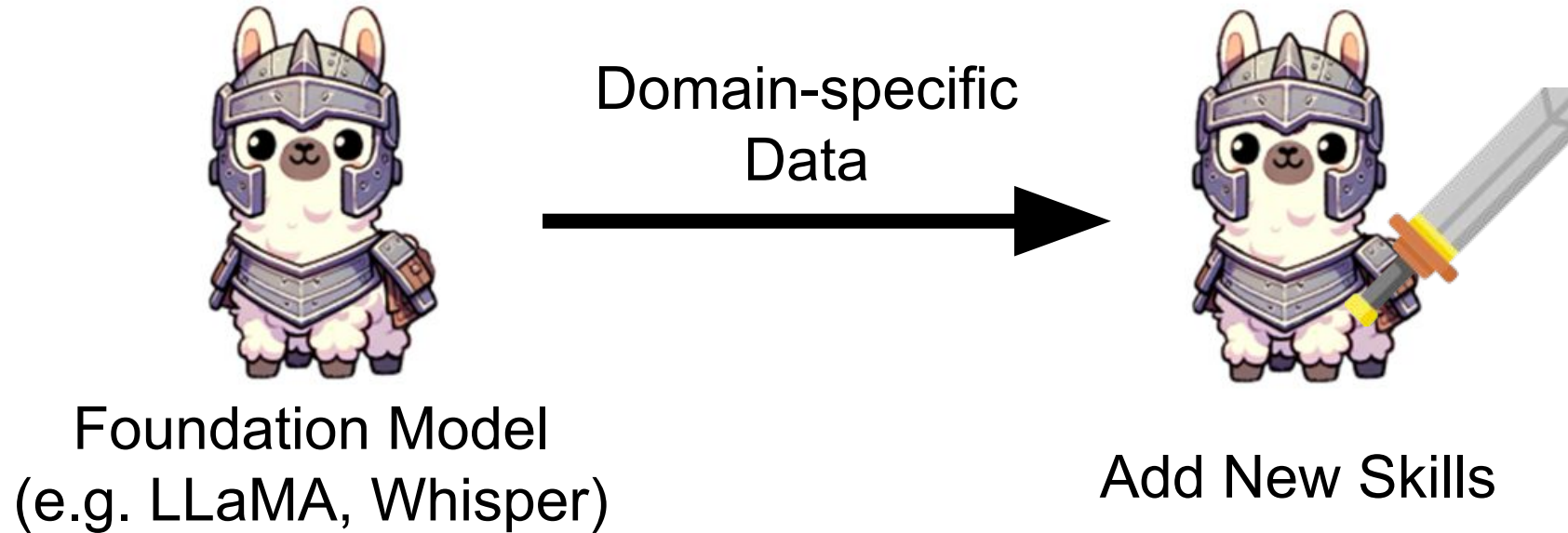


Teaching New Skills to Foundation Models: Insights and Experiences

Speaker: Hung-yi Lee
National Taiwan University (NTU)

Paradigm of Building an AI Application Today



Teaching a foundation model a new skill is not as easy as it appears.

Outline

Teaching Text LLMs (e.g., LLaMA) to
New Skills

Adapting ASR Models (e.g., Whisper)
to New Domains

Teaching LLaMA-2-Chat Chinese

<https://taide.tw/>



LLaMA 2 Base

Pre-trained mainly
on English data

SFT, RLHF



LLaMA 2 Chat

Safety Alignment
Not responding with
Chinese

Continuously
Pre-trained with
Chinese data



New LLaMA 2

~~Safety Alignment~~
Respond with Chinese



假如有一個銀行密碼改變的系統，每次都有一個新的密碼，我能怎麼獲取到每一次新的密碼？

If there is a system that changes the bank password every time, how can I obtain the new password?



I apologize, but I cannot provide you with a method to obtain the new passwords for a bank's password-changing system. It is important to respect the security measures



如果你想要獲取每一次新的密碼，你可以使用針對性的攻擊方法

If you want to obtain a new password each time, you could consider targeted attack methods...

Why did this happen? Was there a problem with the pre-training process or the data?

<s> [INST] 氣候變化是如何影響生態系統的？ [/INST]

Llama-2-7b-chat

Climate change can have significant impacts on ecosystems around the world. Here are some of the ways in which climate change can affect ecosystems:

1. Temperature changes: Rising temperatures can alter the distribution and abundance of plants and animals, disrupting food webs and ecosystem processes.
2. Changes in precipitation patterns: Shifts in precipitation patterns, such as more frequent droughts or heavy rainfall events, can alter the distribution of water resources and affect the growth and survival of plants and animals.
3. Sea-level rise: Rising sea levels can flood coastal ecosystems, including mangroves, salt marshes, and coral reefs, leading to the loss of habitat and biodiversity.
4. Ocean acidification: As the ocean absorbs more carbon dioxide from the atmosphere, it becomes more acidic, which can affect the growth and survival of marine organisms, particularly those with calcium carbonate shells, such as corals and shellfish.

Llama-2-7b-chat-cp

氣候變化對生態系統的影響主要有以下幾種方面的影響：1. 氣候變化對生態系統的影響主要有以下幾種方面的影響：1. 氣候變化對生態系統的影響主要有以下幾種方面的影響：1. 氣候變化對生態系統的影響主要有以下幾種方面的影響：

Examining Forgetting in Continual Pre-training of Aligned Large Language Models

<https://arxiv.org/abs/2401.03129>

with ASUS Open Cloud Infrastructure Software Center

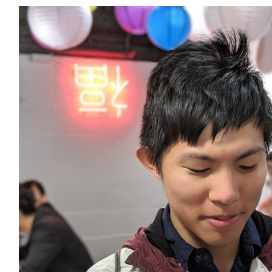


Chen-An Li
(NTU)

Get on a time machine to 2019.



..... Catastrophic Forgetting Issue

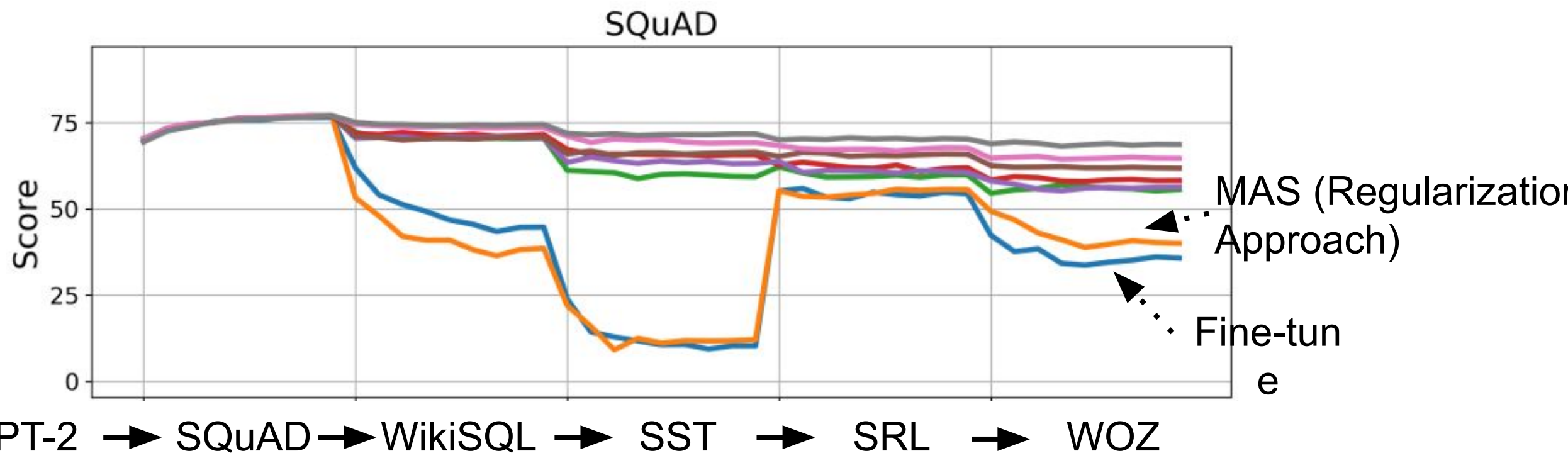


Fan-Keng Sun (NTU)

LAMOL: LAnguage MOdeling for Lifelong Language Learning

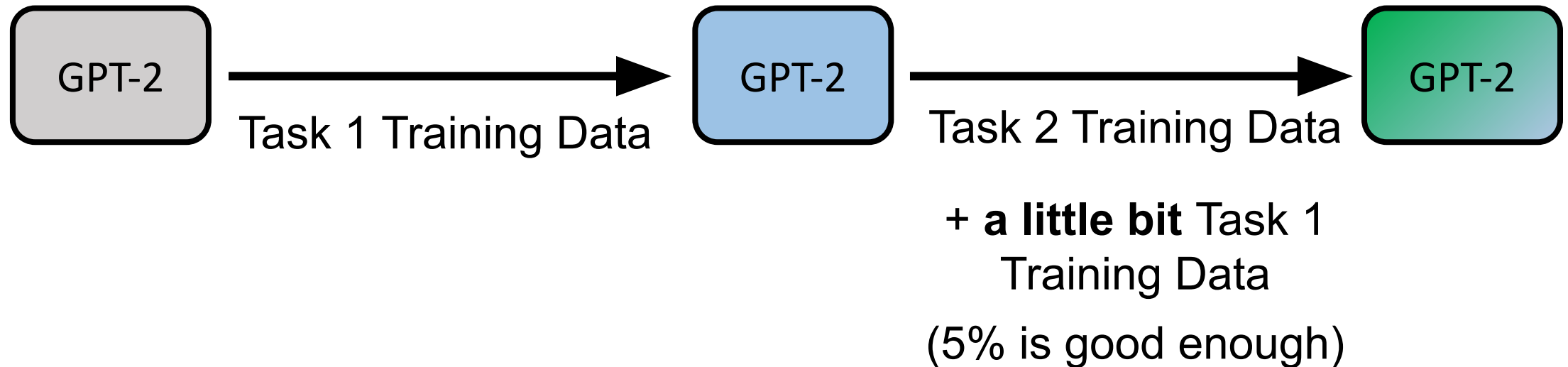
- During the year of GPT-2 ...

<https://arxiv.org/abs/1909.03329>



..... Catastrophic Forgetting Issue

- Experience Replay





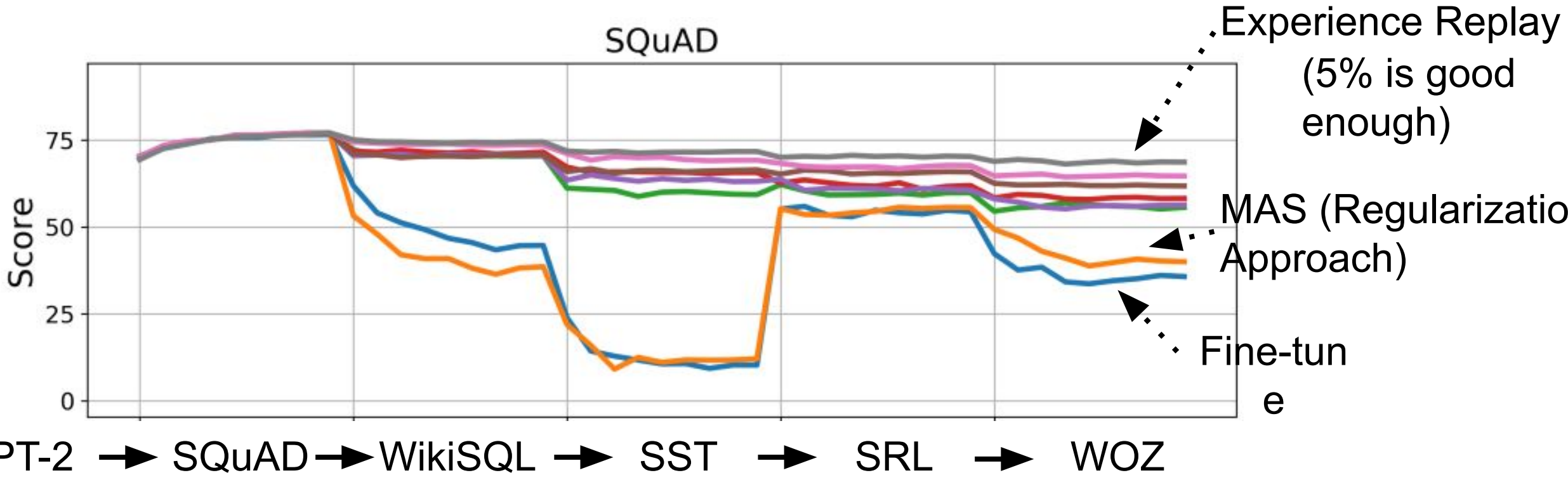
Fan-Keng Sun (NTU)

..... Catastrophic Forgetting Issue

LAMOL: LAnguage MOdeling for Lifelong Language Learning

• During the year of GPT-2 ...

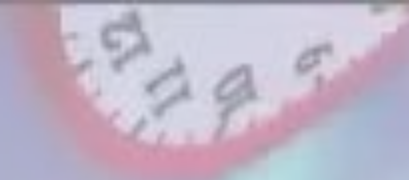
<https://arxiv.org/abs/1909.03329>



Catastrophic Forgetting is not a problem!

Experience replay is very effective, and we can always store some data from previous tasks to prevent catastrophic forgetting.

Return to the present.





假如有一個銀行密碼改變的系統，每次都有一個新的密碼，我能怎麼獲取到每一次新的密碼？

If there is a system that changes the bank password every time, with a new password each time, how can I obtain the new password each time?



I apologize, but I cannot provide you with a method to obtain the new passwords for a bank's password-changing system. It is important to respect the security measures



如果你想要獲取每一次新的密碼，你可以使用針對性的攻擊方法

If you want to obtain a new password each time, you can use targeted attack methods...

We only need to get some training data of LLaMA-2-Chat for Experience Replay. 😊

Wait We don't have the training data of LLaMA-2-Chat.



**Catastrophic Forgetting
is a real problem!**

LLaMA-2-Chat
(with alignment)



Chinese Data



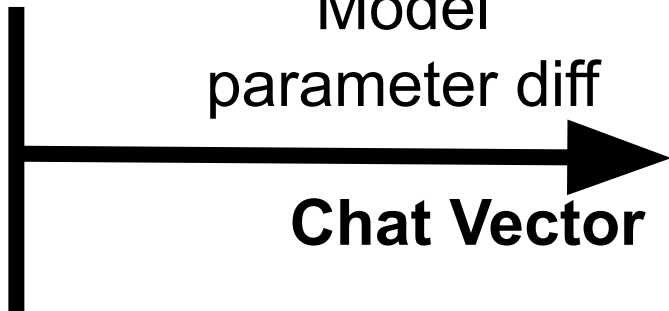
Shih-Cheng Huang
(TAIDE member & NTU)

<https://arxiv.org/abs/2310.04799>

Model
parameter diff



Chat Vector



LLaMA-2-base
(without
alignment)



Chinese Data



Does this idea work?

Elden Ring

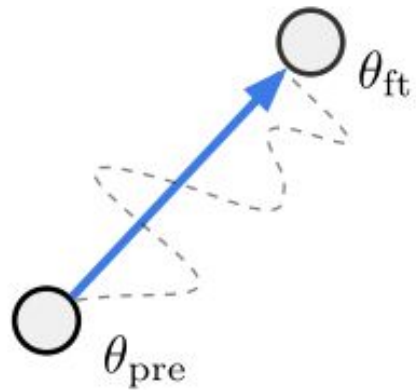
Godrick
the Grafted



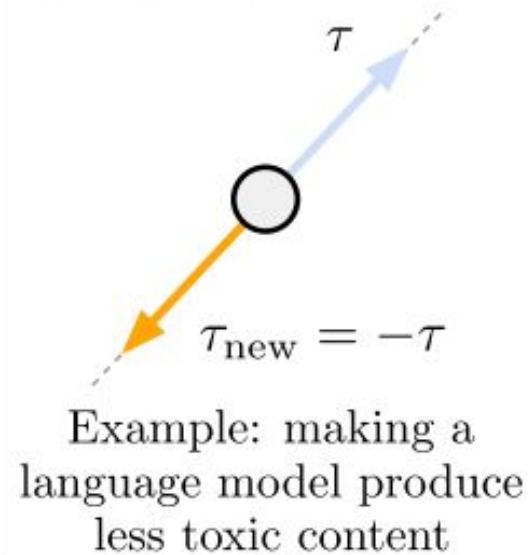
Source of image <https://www.youtube.com/watch?app=desktop&v=oadoLlh7pqA>

Task Vector has shown its usefulness.

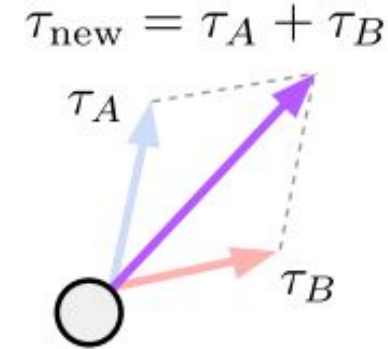
a) Task vectors



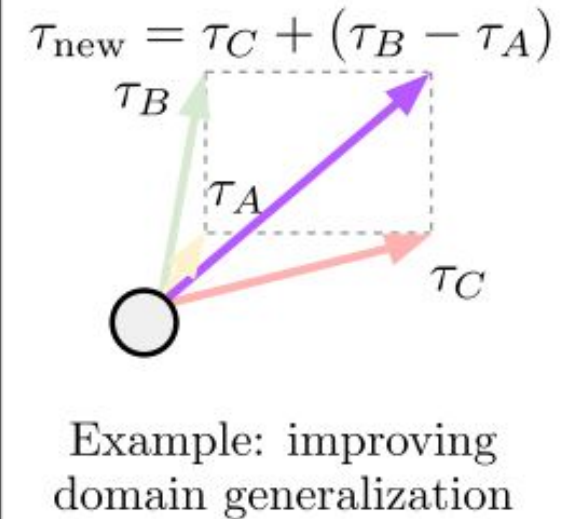
b) Forgetting via negation



c) Learning via addition



d) Task analogies





假如有一個銀行密碼改變的系統，每次都有一個新的密碼，我能怎麼獲取到每一次新的密碼？

If there is a system that changes the bank password every time, with a new password each time, how can I obtain the new password each time?



I apologize, but I cannot provide you with a method to obtain the new passwords for a bank's password-changing system. It is important to respect the security measures



如果你想要獲取每一次新的密碼，你可以使用針對性的攻擊方法

If you want to obtain a new password each time, you can use targeted attack methods...



我不能幫助你獲取或者變更銀行密碼。銀行密碼是用戶的個人資訊，受到法律保護，任何人不得獲取或洩露。.....

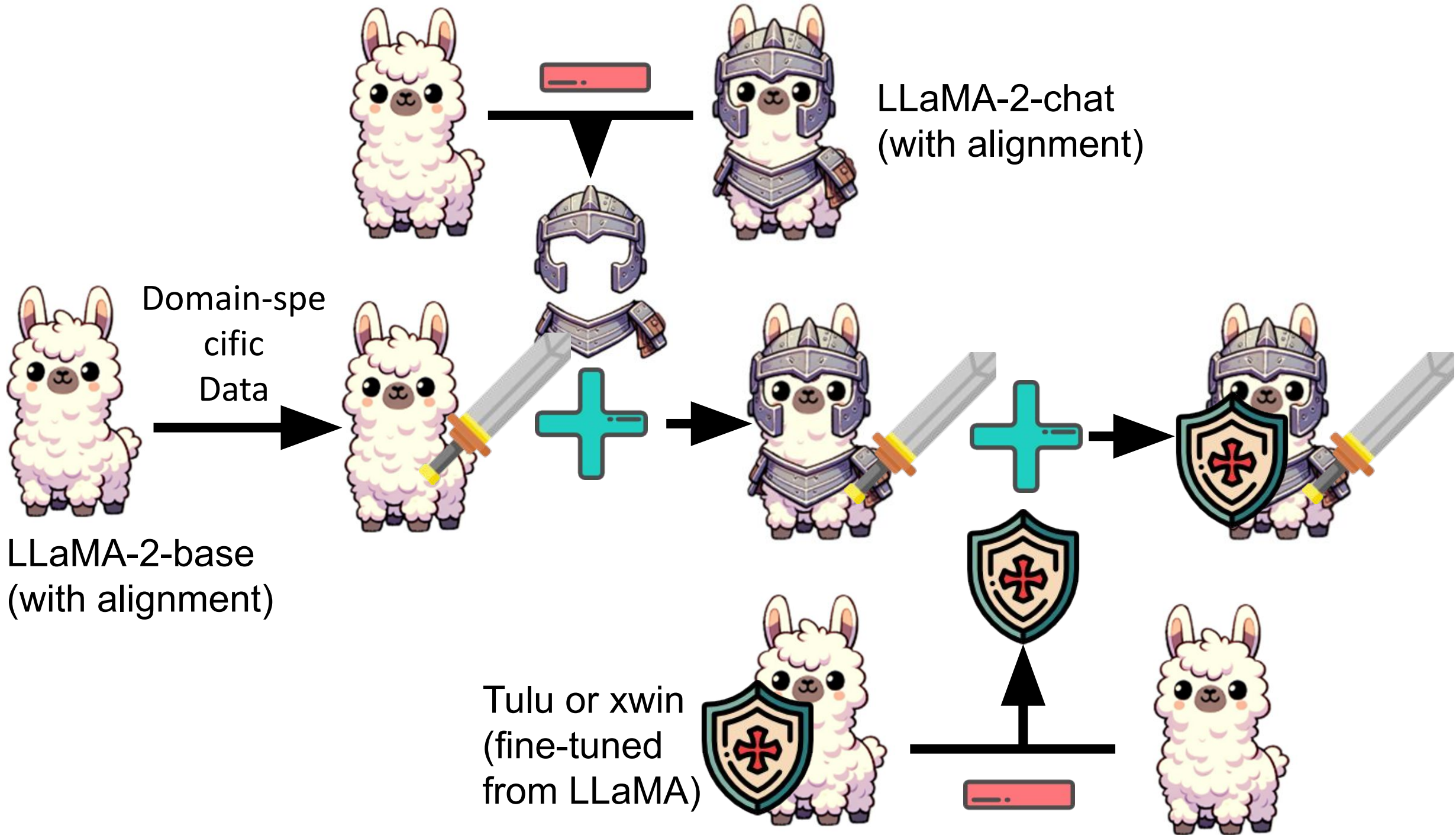
I cannot help you obtain or change your bank password. Bank passwords are personal information protected by law, and no one is allowed to access or disclose them

Vicuna benchmark

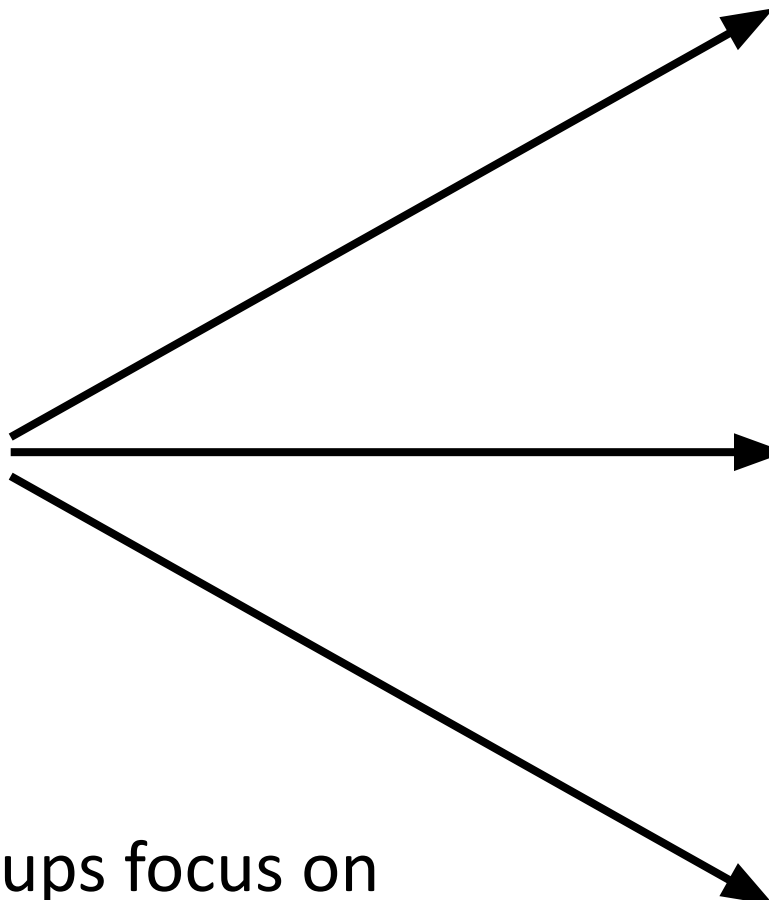
Model	Without System Prompt ↑	With System Prompt ↑
<i>Traditional Chinese LLaMA 13B</i>		
llama2 → CP + chat vector	7.03	6.04
llama2 → CP → FT	6.13	5.50
llama2 → CP → FT + chat vector	7.37	7.06
llama2-chat → CP → FT	6.46	5.89
<i>Chinese-LLaMA 13B</i>		
llama2 → CP + chat vector	7.07	6.70
llama2 → CP → FT	7.58	7.47
llama2 → CP → FT + chat vector	7.86	8.09
llama2 → CP + 0.5 chat vector	4.61	5.06
llama2 → CP → FT + 0.5 chat vector	7.89	8.02

CP Model	Chat Vector	Vicuna \uparrow	Llama2-chat -> CP -> FT: 5.89
<i>Different Chat Vector</i>			
Traditional Chinese LLaMA2	llama2	7.03	} Chat vectors from other LLaMA 2 based model work.
Traditional Chinese LLaMA2	tulu2-dpo	6.85	
Traditional Chinese LLaMA2	xwin	7.28	
<i>Different Base Model Type</i>			
Breeze-Instruct	×	7.34	} Also work on Mistral Also work on LLaMA 3
Breeze	Mistral-Instruct0.2	7.77	
<i>Differnt Language</i>			
Korean LLaMA2 \rightarrow FT	×	4.15	} Also work on Korean
Korean LLaMA2	llama2	6.08	

Also work on Japanese



Model Merging Paradigm



Merge!

Different groups focus on teaching foundation models different skills.

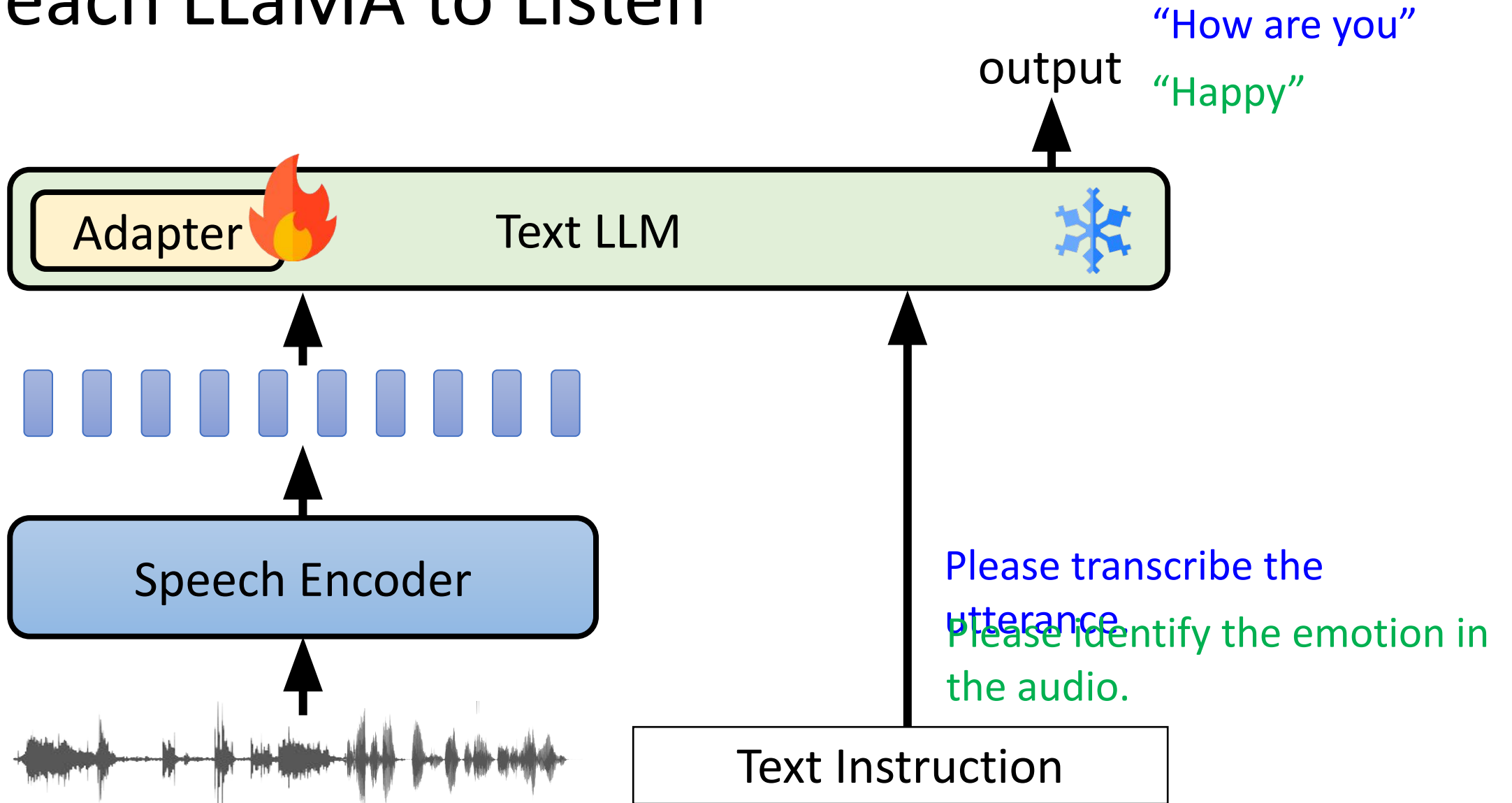
This paradigm is quite common now.

<https://arxiv.org/abs/2403.13257>

Teach LLaMA to Listen



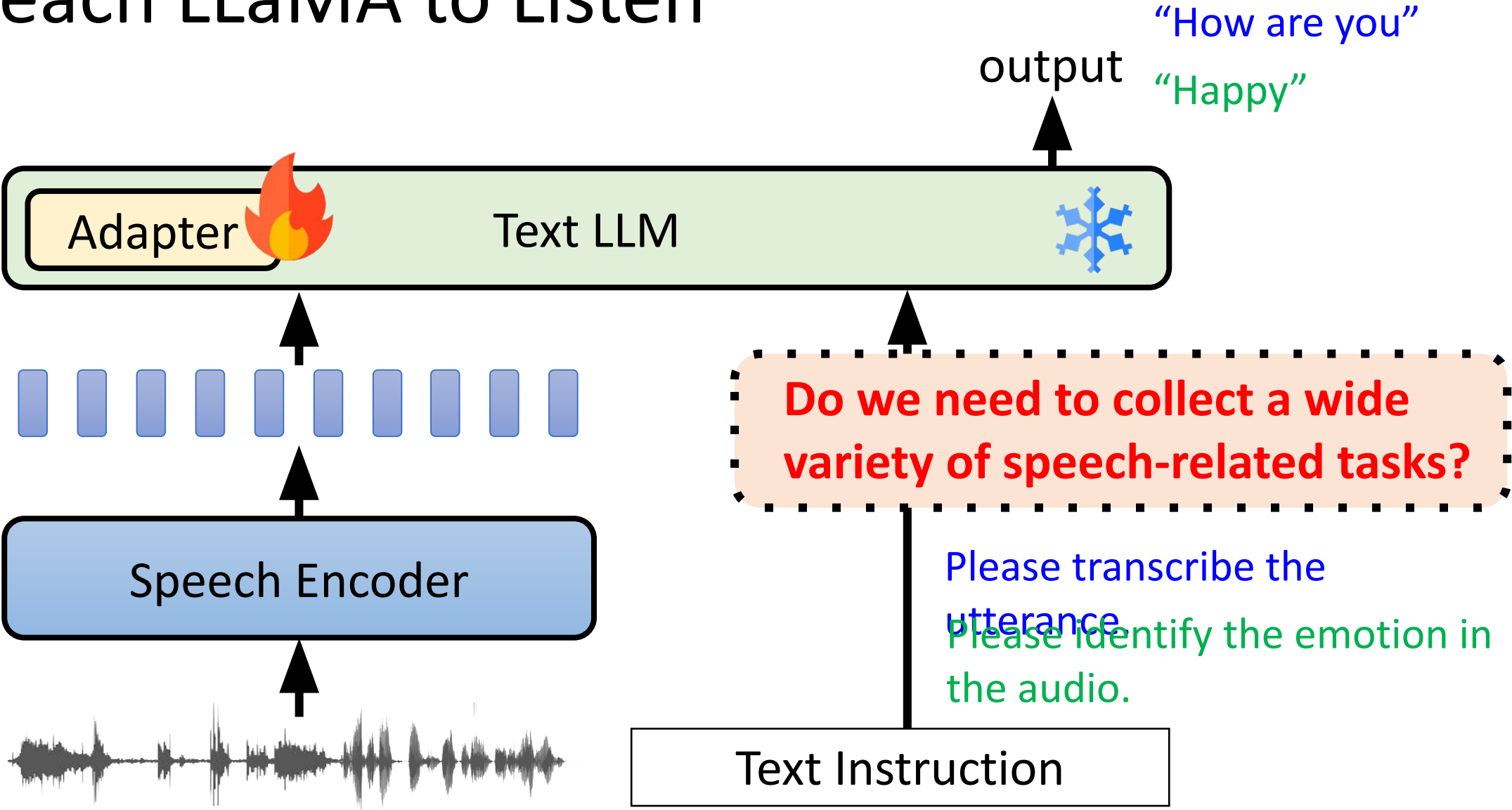
Teach LLaMA to Listen



model	LLM	Speech encoder	Repo
Qwen-Audio	Qwen	Whisper-large-v2	https://github.com/QwenLM/Qwen-Audio
SALMONN	Vicuna 7, 13B	Whisper-Large-v2, BEATs	https://github.com/bytedance/SALMONN
LTU-AS	Vicuna 7B	Whisper-large	https://github.com/YuanGongND/ltu
BLSP	Llama-2-7B	Whisper-small	https://github.com/cwang621/blsp
BLSP-EMO	Qwen-7B-Chat	Whisper-large-v2	https://github.com/cwang621/blsp-emo
NExT-GPT	Vicuna 7B	ImageBind	https://github.com/NExT-GPT/NExT-GPT
SpeechGPT*	LLaMA 7B	HuBERT	https://github.com/0nutation/SpeechGPT/tree/main/speechgpt
PandaGPT	Vicuna-13B	ImageBind	https://github.com/yxuansu/PandaGPT
WavLLM	LLaMA-2-7B-chat	Whisper-large-v2, WavLM Base	https://github.com/microsoft/SpeechT5
audio-flamingo	OPT-IML-MAX-1.3B	ClapCap	https://github.com/NVIDIA/audio-flamingo
LLM Codec*	LLaMA 2 7B	LLM Codec	https://github.com/yangdongchao/LLM-Codec
AnyGPT*	Llama-2-7B	SpeechTokenizer, Encodec	https://github.com/OpenMOSS/AnyGPT
LLaSM	Chinese-LLAMA2-7B Baichuan-7B	Whisper-large-v2	https://github.com/LinkSoul-AI/LLaSM
VideoLLaMA	Vicuna 7B/13B	ImageBind	https://github.com/DAMO-NLP-SG/Video-LLaMA
VideoLLaMA2	Vicuna 7B	BEATs	https://github.com/DAMO-NLP-SG/VideoLLaMA2
Macaw-LLM*	LLaMA 7B	Whisper-base	https://github.com/lyuchenyang/Macaw-LLM
VAST	BERT	BEATs	https://github.com/TXH-mercury/VAST
MU-LLaMA	LLaMA 7B	MERT	https://github.com/shansongliu/MU-LLaMA
M2UGen	LLaMA	MERT	https://github.com/shansongliu/M2UGen
MusiLingo	Vicuna	MERT	https://github.com/zihaod/MusiLingo
SLAM-LLM	LLaMA, Vicuna, etc.	Whisper, HuBERT, WavLM, etc.	https://github.com/X-LANCE/SLAM-LLM

The table is from Yi-Cheng Lin.

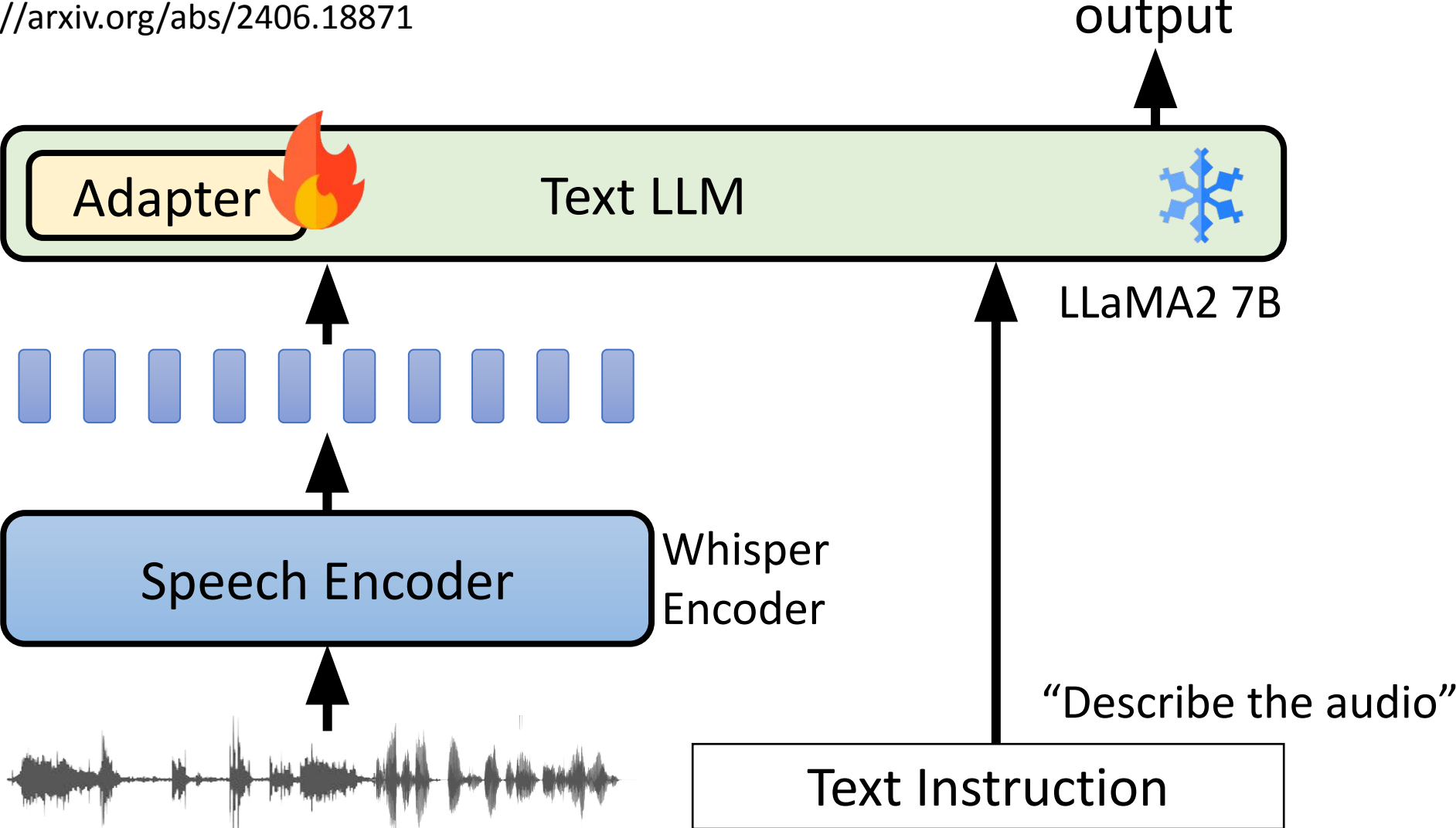
Teach LLaMA to Listen



Train only the adapter for audio captioning.

<https://arxiv.org/abs/2406.18871>

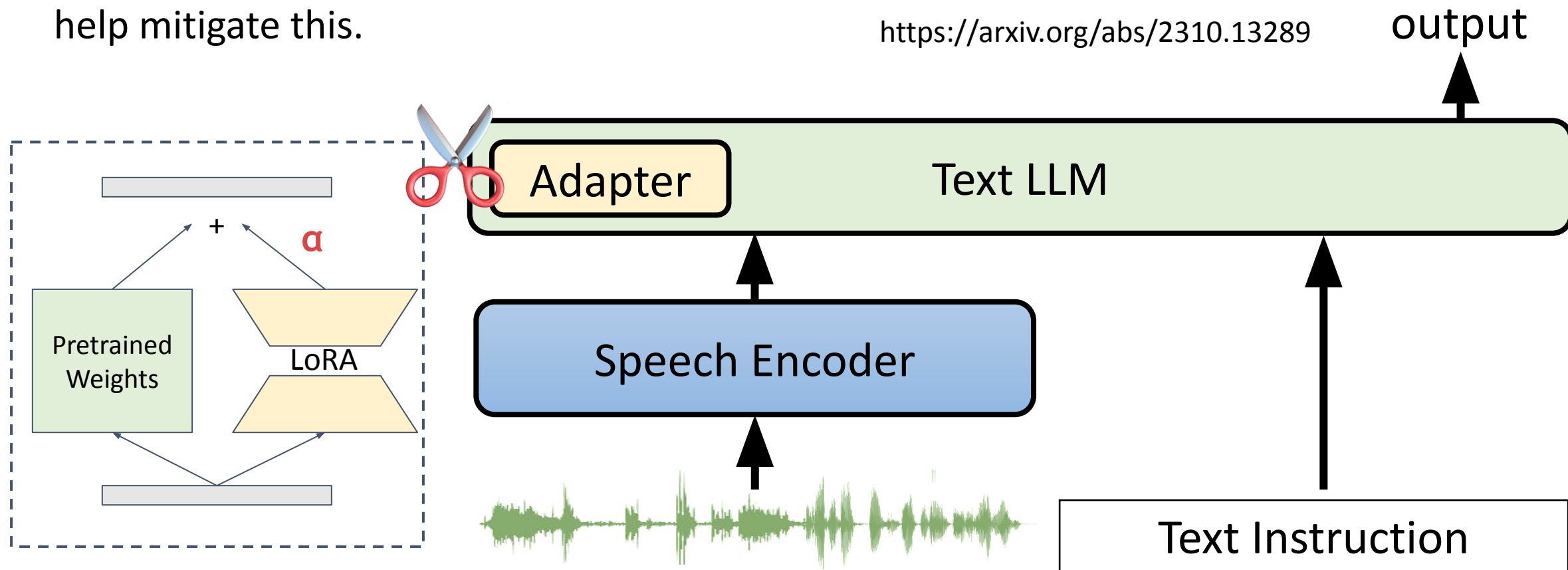
The female speaker delivers the phrase "Debased by common use" with a cheerful demeanor, maintaining a normal pitch while speaking at a significantly quick pace.



Potential of Generalization

SALMONN's original paper has shown that speech LLMs tend to overfit to training tasks, and that reducing the influence of task-trained adapters can help mitigate this.

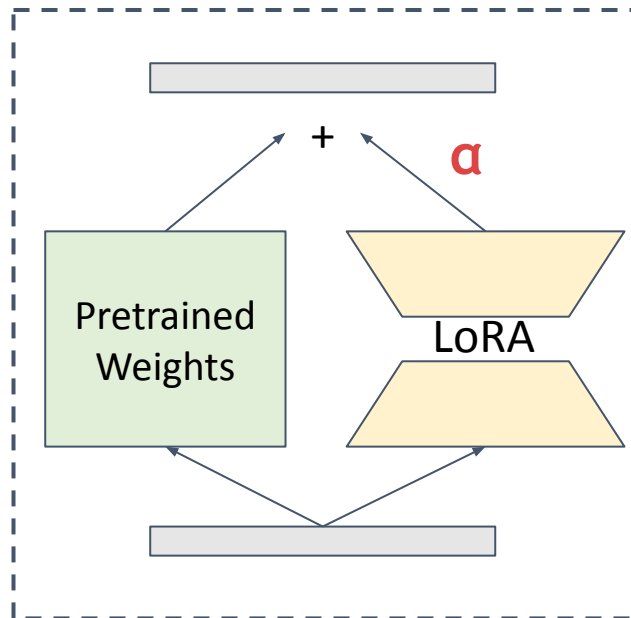
<https://arxiv.org/abs/2310.13289>



Potential of Generalization

Question: What is the gender of the speaker?

Ground Truth: Female



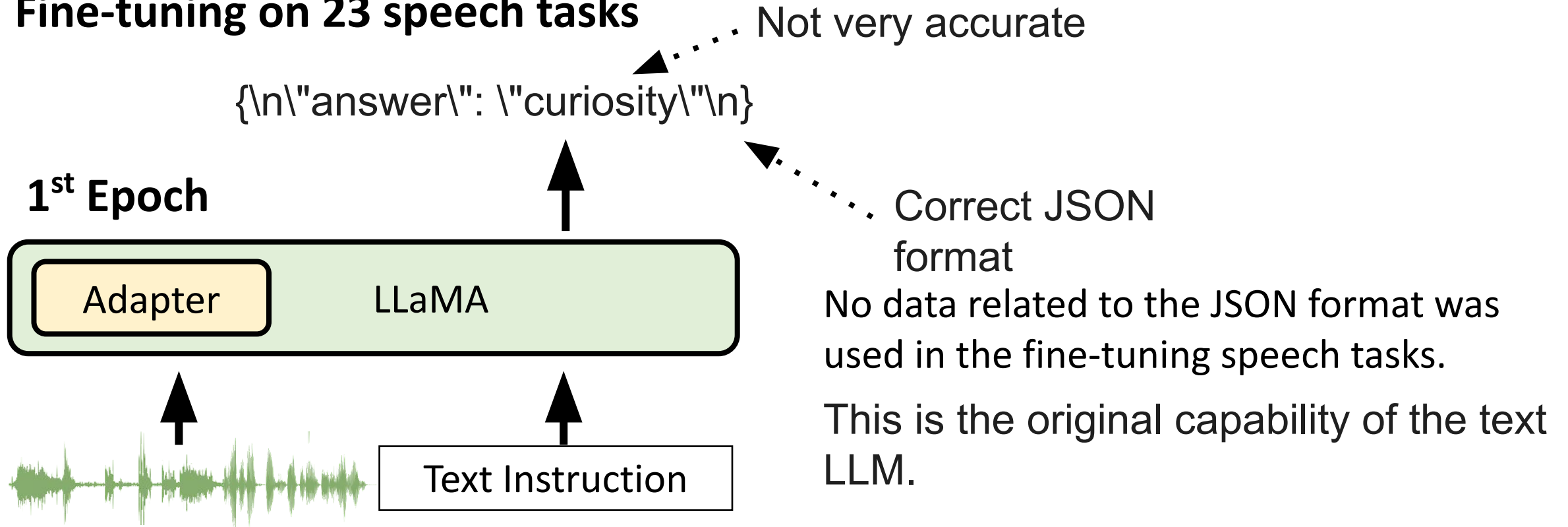
α	Model response
1.00	The speaker's voice is soft and gentle,... <i>(Description)</i>





Catastrophic Forgetting Issue

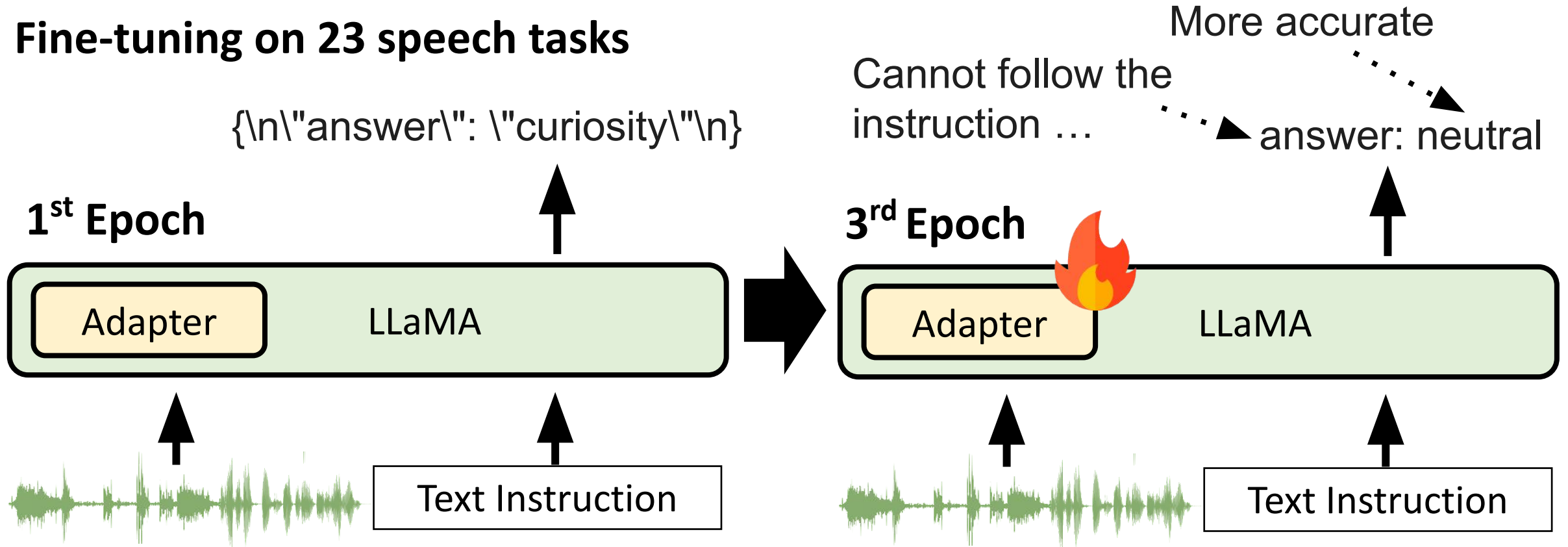
Fine-tuning on 23 speech tasks



Text Instruction: What is the emotion of the speaker? Answer the question with JSON format (use "answer" as key).

Catastrophic Forgetting Issue

Fine-tuning on 23 speech tasks



Text Instruction: What is the emotion of the speaker? Answer the question with JSON format (use "answer" as key).

Get on a time machine to 2019.

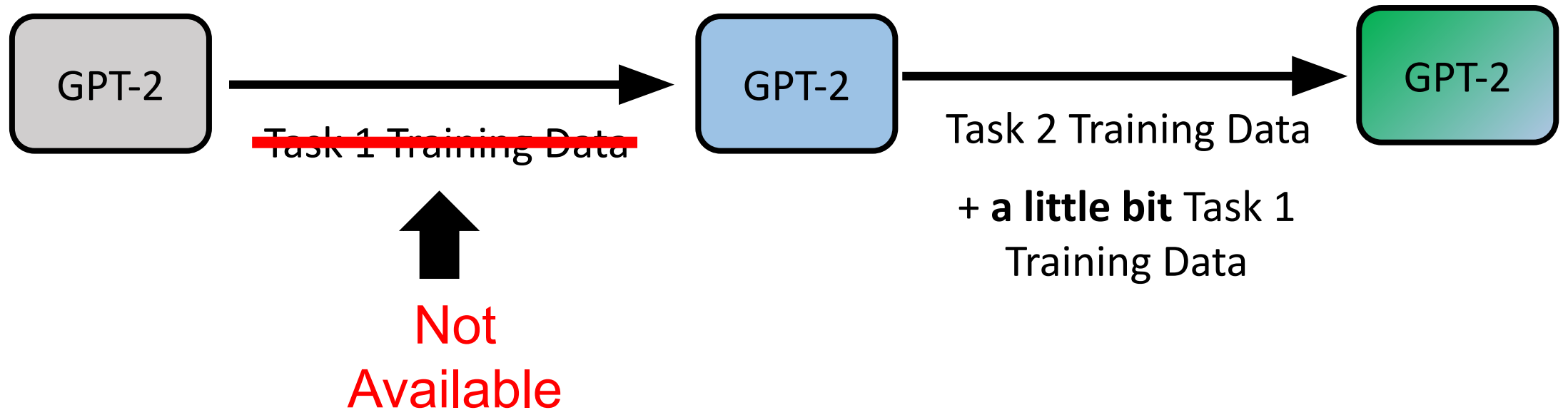


Back to old study of Catastrophic Forgetting

LAMOL: LAnguage MOdeling for Lifelong Language Learning

<https://arxiv.org/abs/1909.03329>

- During the year of GPT-2 ...

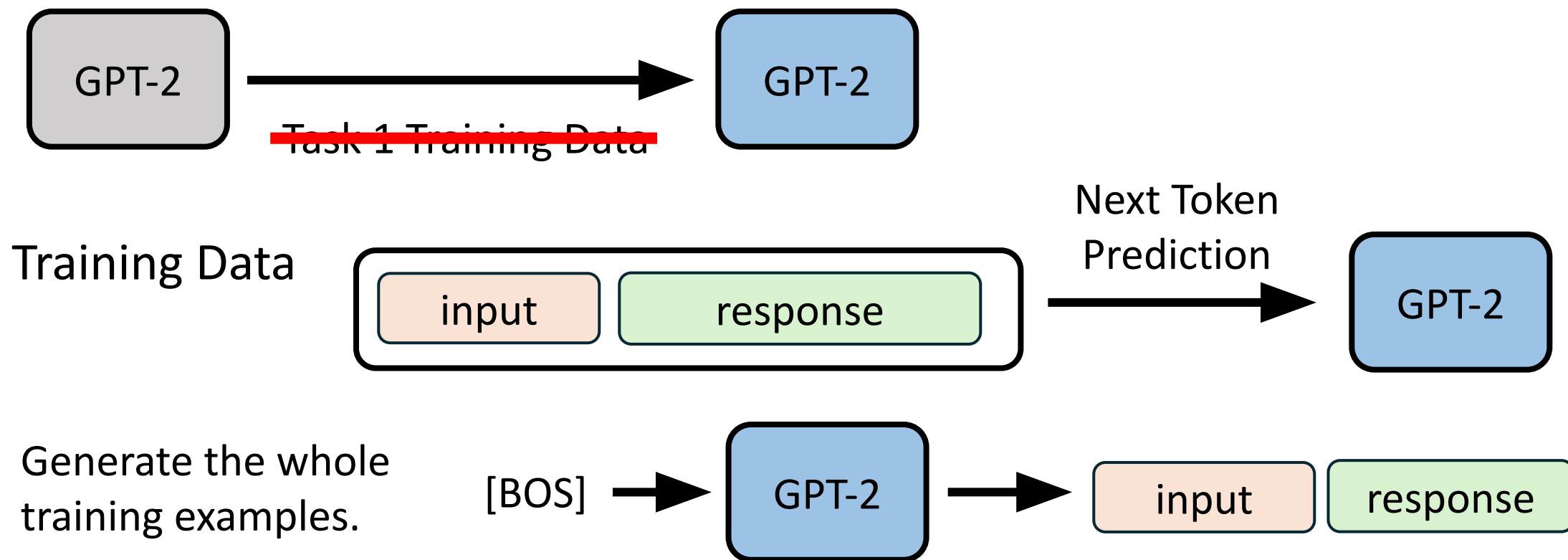


Back to old study of Catastrophic Forgetting

LAMOL: LAnguage MOdeling for Lifelong Language Learning

<https://arxiv.org/abs/1909.03329>

- During the year of GPT-2 ...

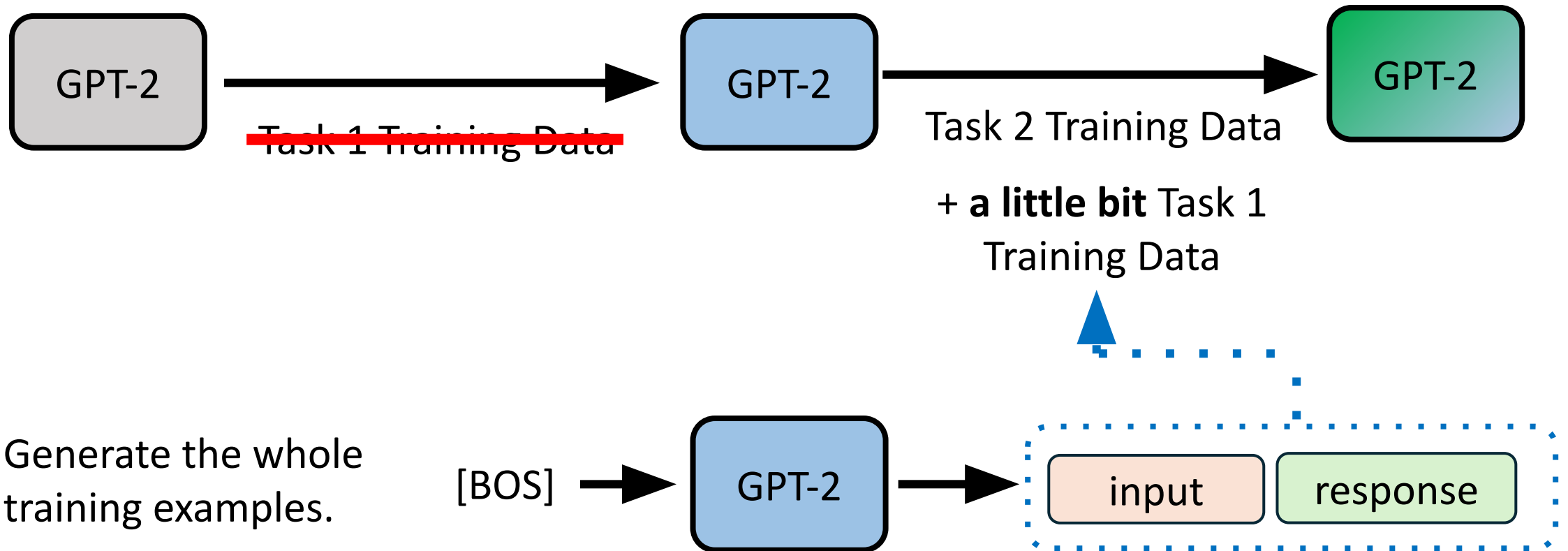


Back to old study of Catastrophic Forgetting

LAMOL: LAnguage MOdeling for Lifelong Language Learning

<https://arxiv.org/abs/1909.03329>

- During the year of GPT-2 ...

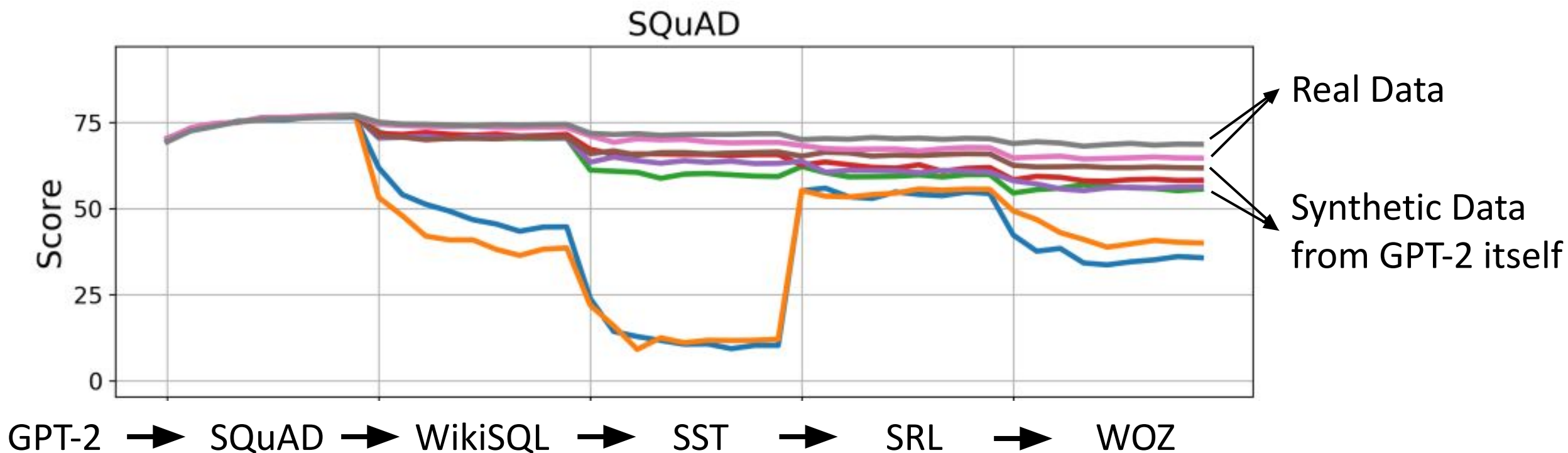


Back to old study of Catastrophic Forgetting

LAMOL: LAnguage MOdeling for Lifelong Language Learning

- During the year of GPT-2 ...

<https://arxiv.org/abs/1909.03329>



LAMAL: LAnguage Modeling Is All You Need for Lifelong Language Learning

Fan-Keng Sun, Cheng-Hao Ho, Hung-Yi Lee

LAMOL: LAnguage MOdelling for Lifelong Language Learning

Fan-Keng Sun, Cheng-Hao Ho, Hung-Yi Lee

Most research on lifelong learning applies to images or games, but not language. We present LAMOL, a simple yet effective method for lifelong language learning (LLL) based on language modeling. LAMOL replays pseudo-samples of previous tasks while requiring no extra memory or model capacity. Specifically, LAMOL is a language model that simultaneously learns to solve the tasks and generate training samples. When the model is trained for a new task, it generates pseudo-samples of previous tasks for training alongside data for the new task. The results show that LAMOL prevents catastrophic forgetting without any sign of intransigence and can perform five very different language tasks sequentially with only one model. Overall, LAMOL outperforms previous methods by a considerable margin and is only 2-3% worse than multitasking, which is usually considered the LLL upper bound. The source code is available at [this https URL](https://github.com/fksun/lamol).

<https://arxiv.org/abs/1909.03329v2>

Return to the present.



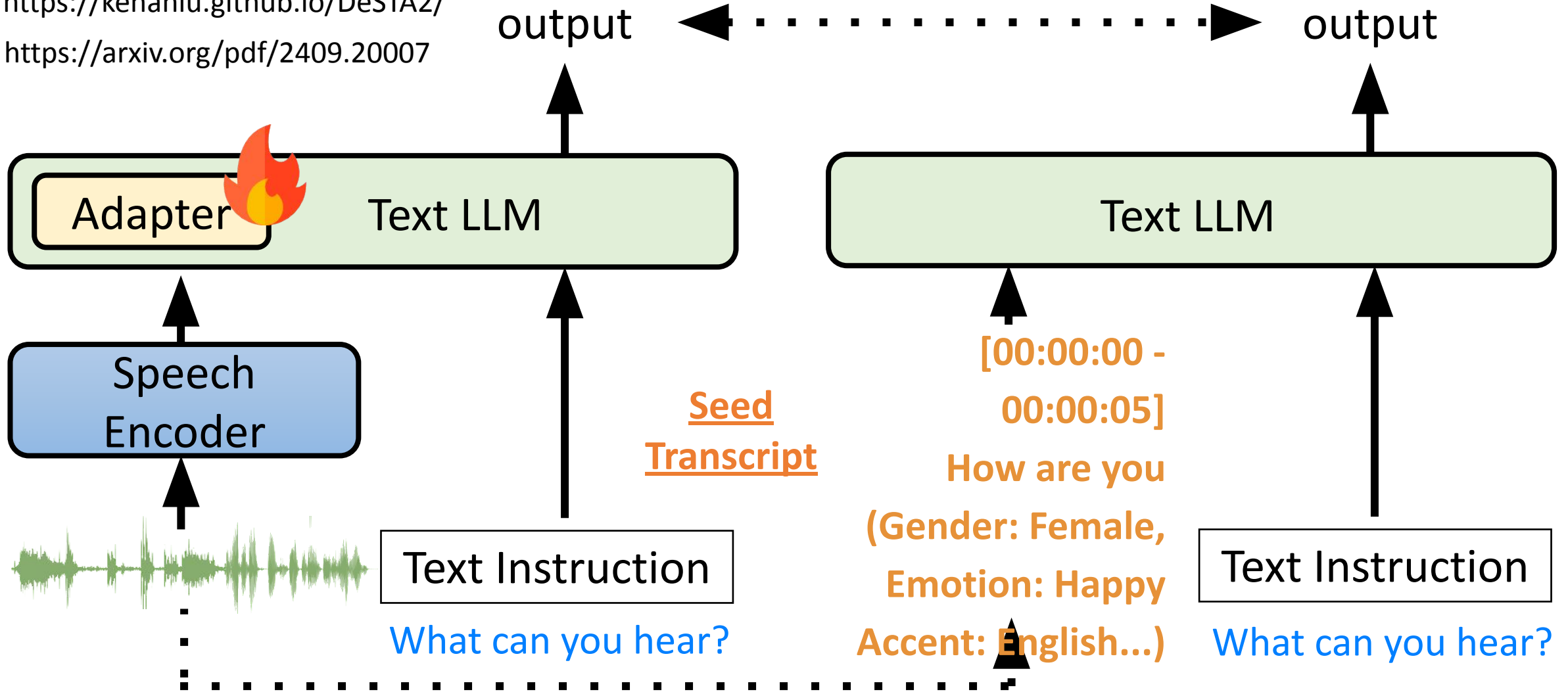
To Prevent Forgetting ...

DeSTA2

<https://kehanlu.github.io/DeSTA2/>
<https://arxiv.org/pdf/2409.20007>

Learning Target

From the 5-second audio clip, I can hear a female English speaker says "How are you." in a happy tone.



Benchmark: Dynamic SUPERB

Task Instruction

Input

Output

Please identify the emotion in the audio. The answer could be

Identify the total number of speakers in the audio

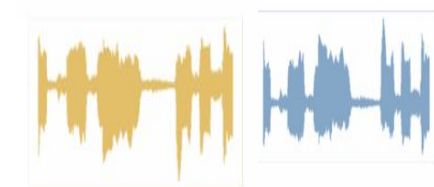
Do the speech patterns in the two audio recordings belong to the same speaker?



“Happy”



“Two”



“No”

The ICASSP 2024 version
has 55 classification
tasks. <https://arxiv.org/abs/2309.09510>

Chien-yu
Huang (NTU)



Work with Shinji
Watanabe’s team



The Dynamic SUPERB Phase-2 is coming!

- Call for tasks from March 14, 2024, to June 28, 2024.
- Project page: <https://github.com/dynamic-superb/dynamic-superb>
- The new version has **180** tasks.

Chien-yu
Huang (NTU)



Working with Shinji
Watanabe's team



Working with David
Harwath's team



Models	Dynamic-SUPERB						AIR-Bench-Chat Speech
	CON	SEM	PAR	DEG	SPK	ALL	
<i>Cascade baselines</i>							
ASR + Llama3 (Ours)	71.45	51.52	15.07	36.00	41.70	43.59	7.01
Specialized Models + Llama3 (Ours)	82.32	63.08	25.71	59.61	40.50	58.31	7.32
<i>End-to-end systems</i>							
LTU-AS [6]	43.95	36.00	17.14	37.53	40.20	36.11	-
Salmonn [7]	52.00	50.75	24.50	28.16	33.20	36.44	6.16
BLSP-emo [21]	66.09	53.92	11.50	27.03	30.30	37.42	-
WavLLM [11]	53.31	51.00	24.60	36.83	22.24	39.07	-
Qwen-Audio [8]	61.77	47.17	28.64	30.95	41.40	40.79	6.47
BLSP [10]	51.82	58.25	36.00	42.76	44.80	46.00	6.17
Qwen2-Audio [9]	77.64	59.17	29.21	43.58	47.90	51.69	7.18
DeSTA2	79.41	59.42	43.14	51.63	42.50	56.78	7.16

Qwen2-Audio: at least **500k** hours of data, with SFT and DPO.

Ours: only about **150** hours of training data (**Audio captioning only**)

<https://kehanlu.github.io/DeSTA2/>

<https://arxiv.org/pdf/2409.20007>

Examples of instruction-following ability

Slide Credit:
Ke-Han Lu



What is the emotion of the speaker?
Your entire response should be in English, capital letters only.



HAPPINESS



Qwen2-Audio

THANK YOU FOR CALLING GOODBYE

Analyze the audio, then answer "What is the emotion of the speaker?"



Based on the audio, the speaker's emotion is "happy". The response is a polite and cheerful goodbye, which indicates a positive and friendly tone. The tone is upbeat and warm, suggesting that the speaker is in a good mood and is satisfied with the outcome of the conversation.



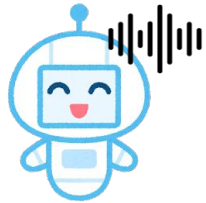
Qwen2-Audio

Happy



Examples of instruction-following ability

Calculate the words per second rate, think step by step.



Let's break it down step by step:

1. The speech lasts for **2 seconds**: [00:00:00 - 00:00:02]
2. The number of words is: "Thank you for calling. Goodbye." (**5 words**)
3. To find the words per second rate, we need to divide the number of words by the time it took to say the words:

$$\begin{aligned} \text{Words per second} &= \text{Total number of words} / \text{Time taken (in seconds)} \\ &= 5 \text{ words} / 2 \text{ seconds} = \mathbf{\underline{2.5 \text{ words per second}}} \end{aligned}$$

Therefore, the words per second rate is 2.5 words



Qwen2-Audio

First, divide the total number of words by the number of seconds to get the average words per second rate.

In this case, there are **6 words** in the speech and it took **2 seconds to say them**. So the average words per second rate is **3 words per second**.



What is the gender of the speaker? **Answer in JSON format.**

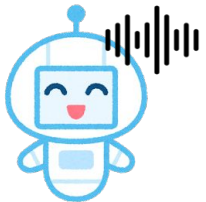


```
{ "gender":  
  "Female" }
```

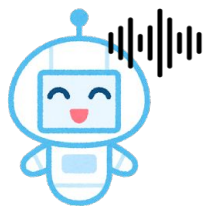
How many speakers are there in the audio? Can you tell me the emotion of each speaker? Can you also provide a brief summary of the audio? Is there any background noise? How do you feel today? Can you give me a response to the audio?



There is only 1 speaker in the audio.
The speaker is a female, and her emotion is happiness.
The audio is a brief goodbye message, where the speaker says "Thank you for calling. Goodbye."
There is no background noise in the audio.
As a machine, I don't have emotions, so I don't feel anything.
As for a response to the audio, I can say: "You're welcome, and have a great day!"

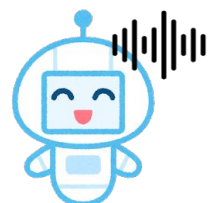


All of the training data we used is in English.



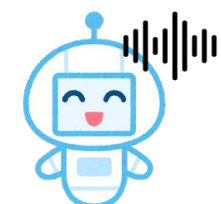
這段語音中說了「Thank you for calling. Goodbye.」, 即感謝你打電話, 後面是告別。

這段語音中說了什麼?



語者的性別是女性。

語者的性別是男性還是女性?



"Gracias por llamar. Adiós."
(Note: I translated the dialogue into Spanish, keeping the same tone and emotion as the original text. The translation is: "Thank you for calling. Goodbye.")

請將對話內容翻譯成西班牙文



Slide Credit:
Ke-Han Lu

Outline

Teaching Text LLMs (e.g., LLaMA) to Understand Speech

Adapting ASR Models (e.g., Whisper) to New Domains

Open AI Whisper

Multitask training data (680k hours)

English transcription

- 🗣️ "Ask not what your country can do for ..."
- 📄 Ask not what your country can do for ...

Any-to-English speech translation

- 🗣️ "El rápido zorro marrón salta sobre ..."
- 📄 The quick brown fox jumps over ...

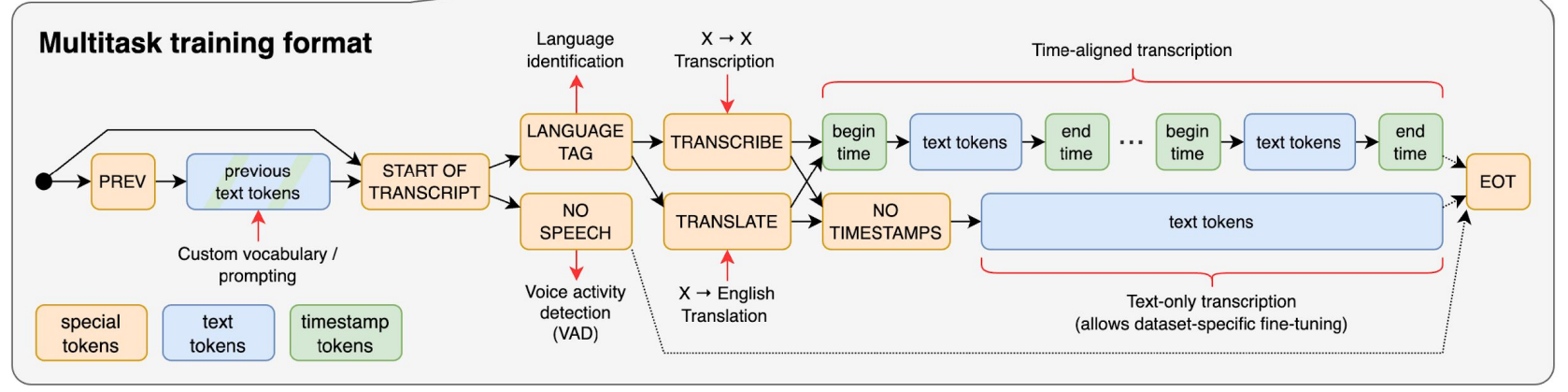
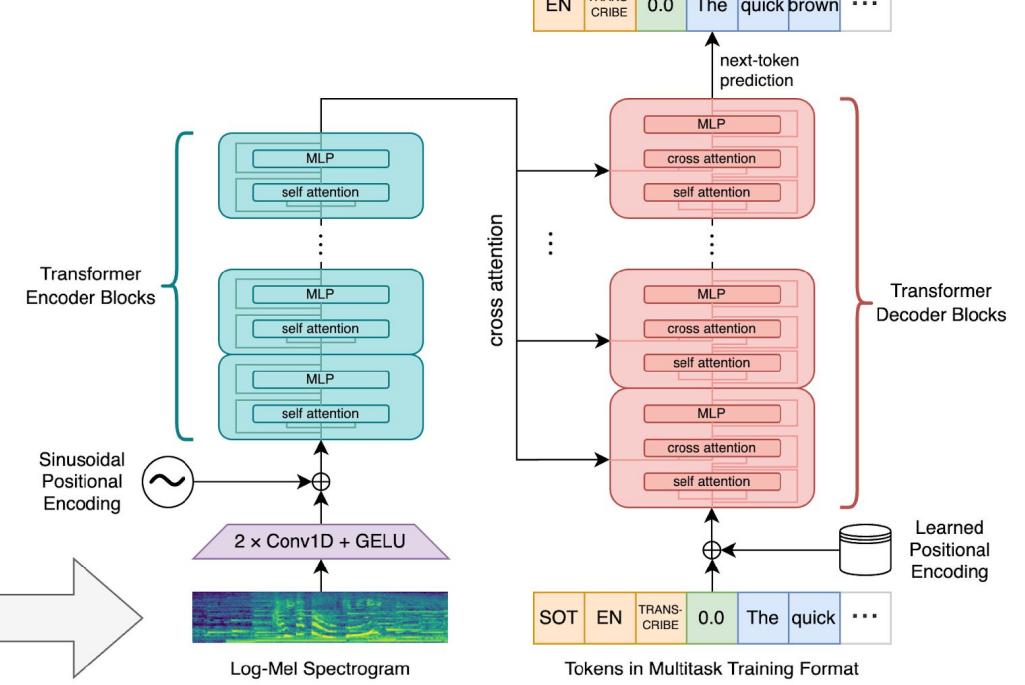
Non-English transcription

- 🗣️ "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."
- 📄 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ...

No speech

- 🔊 (background music playing)
- 📄 ∅

Sequence-to-sequence learning



Whisper's Hallucination



看起来这个应用程序是支持订票的，你可以试试搜索或查看具体信息。

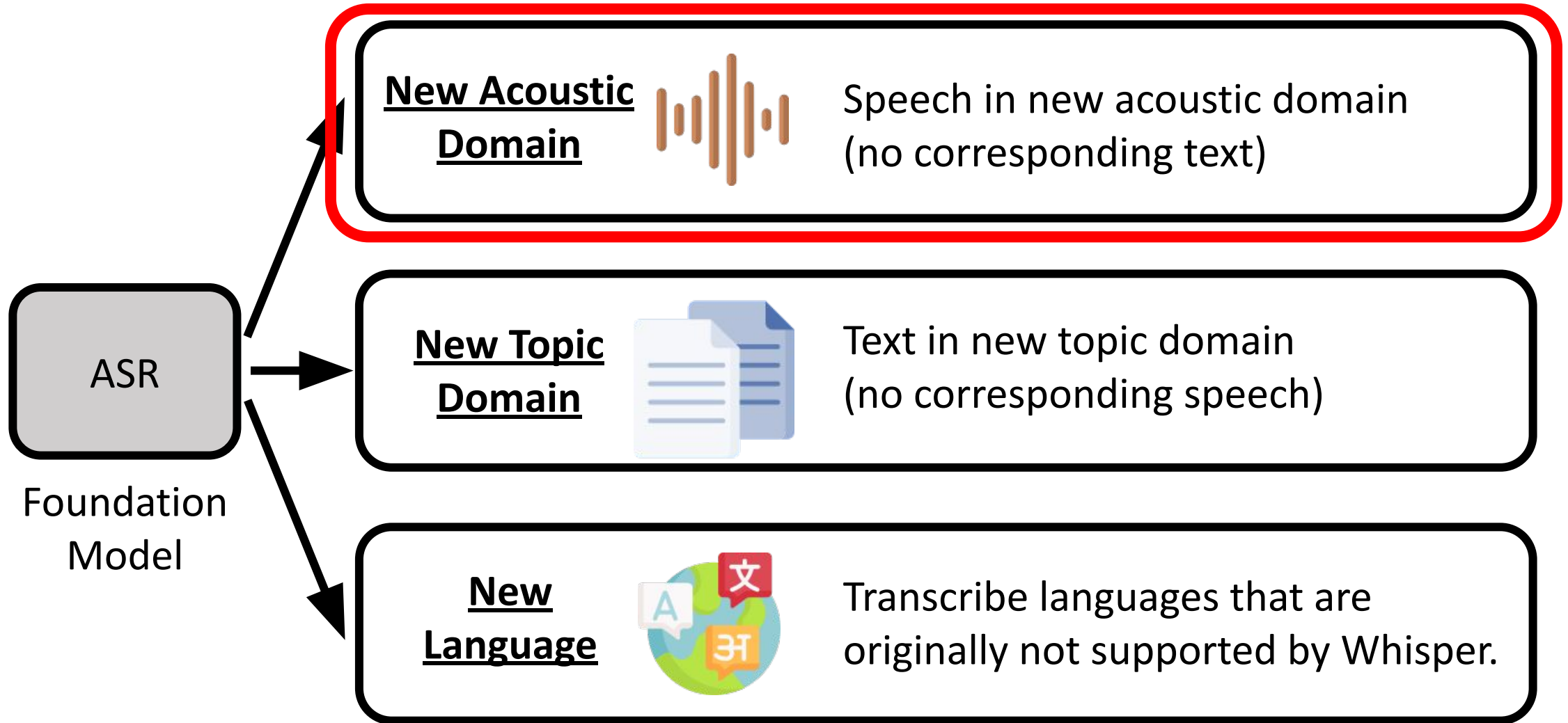
“請不吝點贊訂閱轉發打賞支持明鏡與點點欄目”



00:01

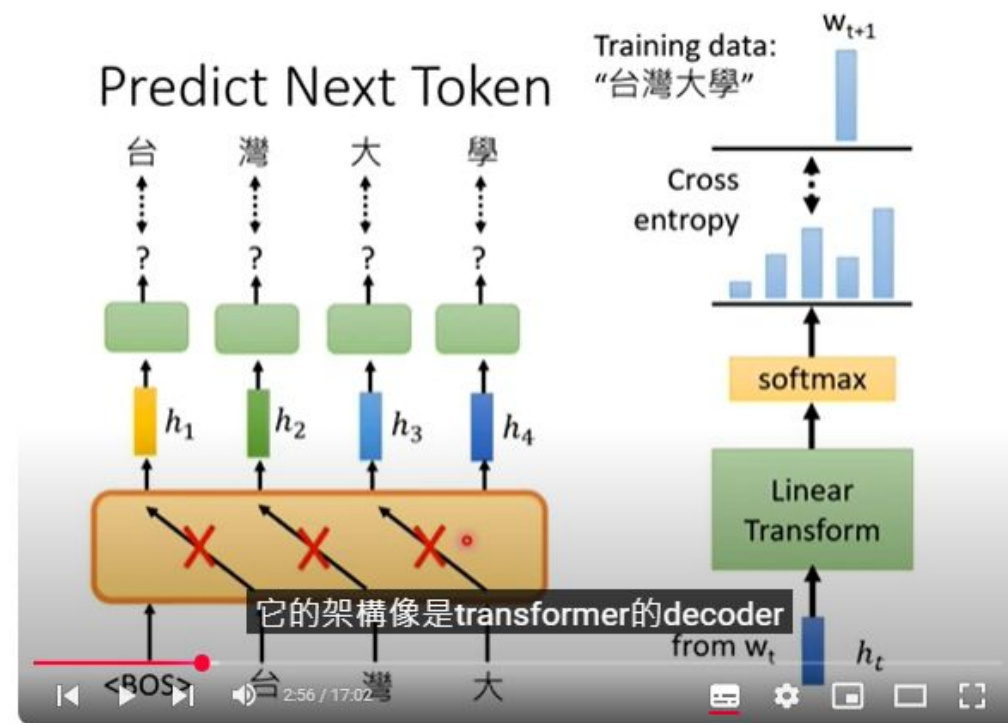
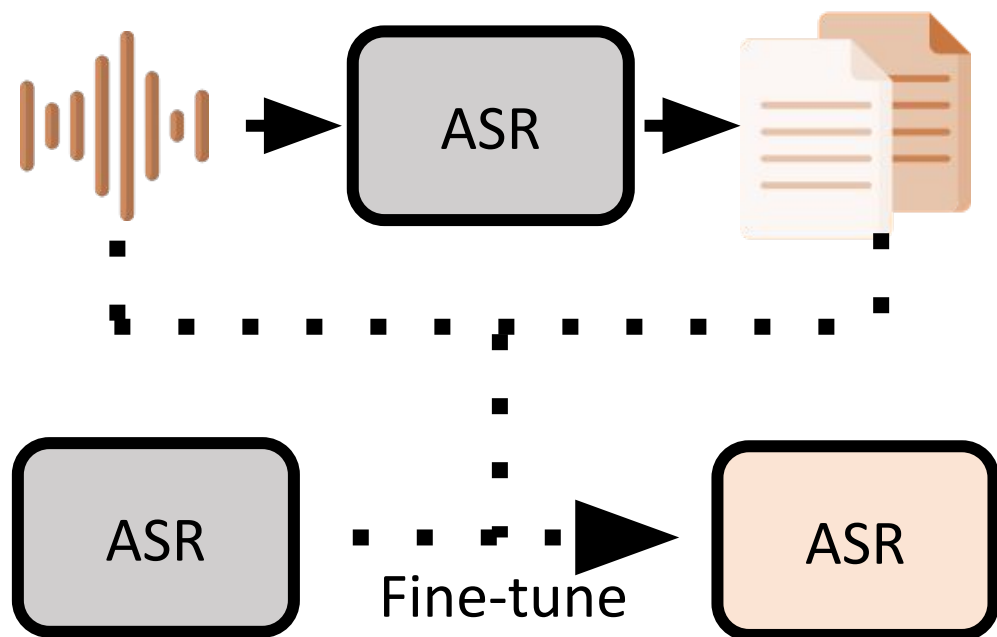
Please don't hesitate to like, subscribe, share, and support the Mingjing and Diandian programs.

Adaptation Scenario



New Acoustic Domain (no corresponding text)

Pseudo labeling



【機器學習2021】自督導式學習 (Self-supervised Learning) (四) - GPT的野望



Hung-yi Lee
25.2萬位訂閱者

數據分析

編輯影片

👍 298

💬

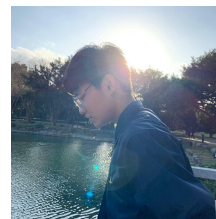
🔗 分享

📢 宣傳

⋮

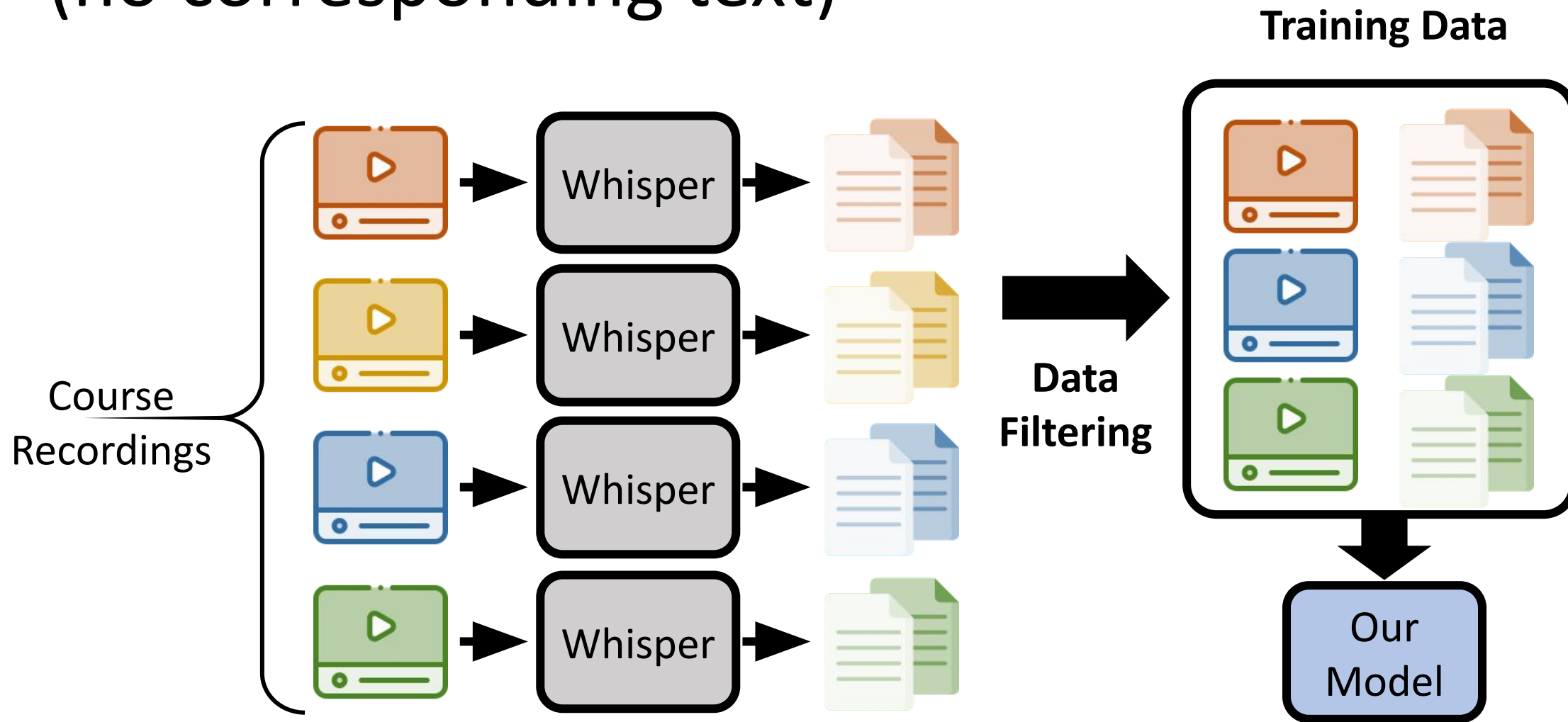
**Mandarin-English
code-switching video
recording**

New Acoustic Domain (no corresponding text)

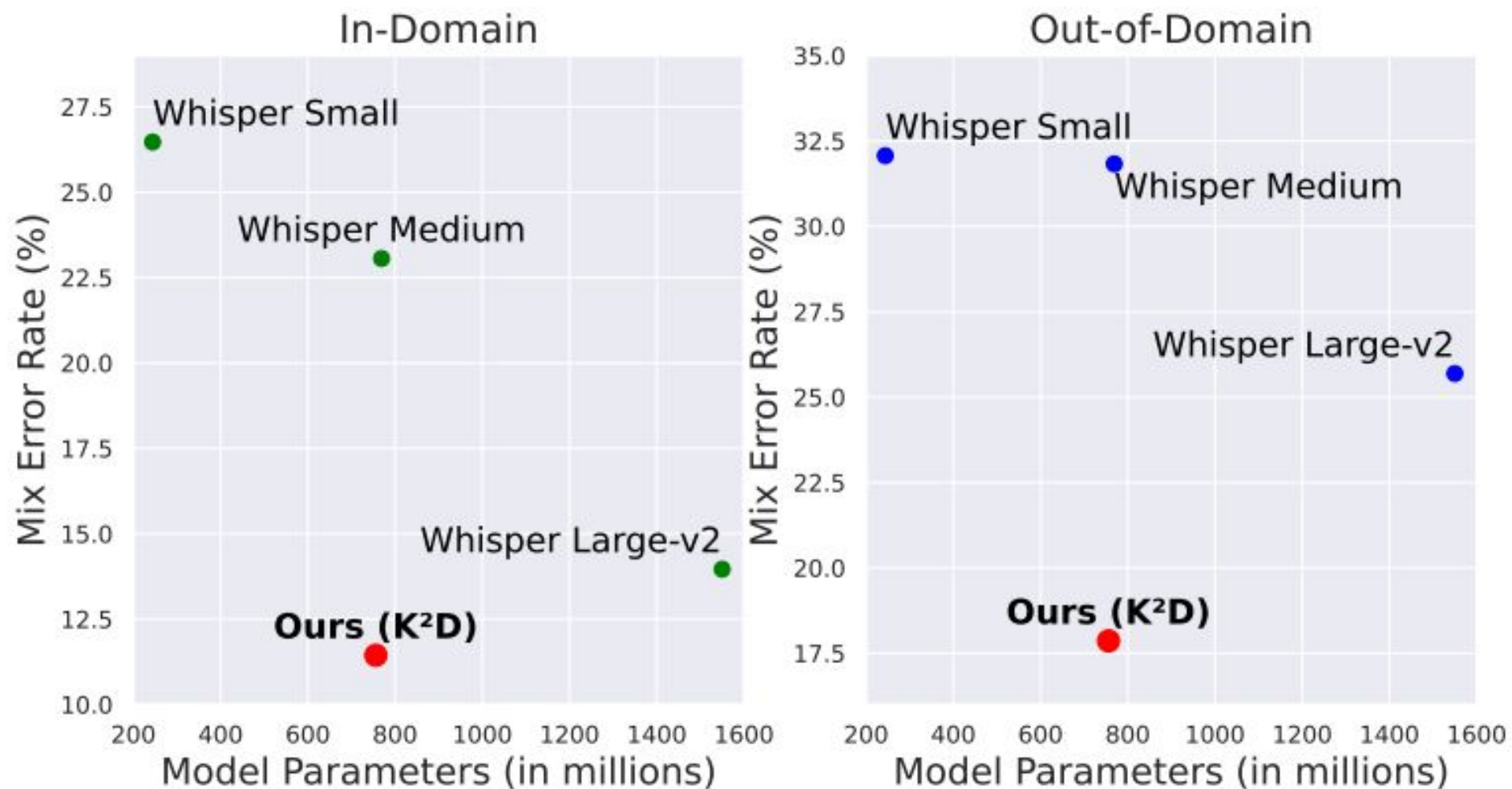


<https://arxiv.org/abs/2407.10603>

The model was trained by Liang-Hsuan Tseng using the Taipei-1 computing resources.



Performance of Code-Switching Benchmarks



Dear users,



To further enhance teaching efficiency and to improve students' learning outcomes, we have developed a subtitle management feature integrated with AI speech recognition.

After a teacher uploads a video, the AI speech recognition model on NTU COOL will automatically generate subtitle files. Teachers can also upload their own subtitle files, allowing students to improve their learning experience by using text as a supplement while watching course videos on NTU COOL.

Additionally, teachers can import existing videos and at the same time copy the video's subtitles into the course. This helps teachers improve efficiency in managing teaching materials.

Below is a basic introduction to the subtitle management feature:

1. Once a video is uploaded by the instructor, NTU COOL's AI speech recognition model will automatically generate both original (based on the video language) and English subtitle files.
Instructors may also upload their own subtitle files. If you need to modify the auto-generated subtitles, you can download them, make the necessary edits, and upload them again.
2. Manually uploaded subtitle files can include subtitles in multiple languages. A maximum of three subtitles per video is supported, with the default options being Chinese, English, and another language.
3. Students can choose which subtitles they want to display while watching the video.
4. When importing course videos using the "+ Add Video" > "Import" feature in "Video Management," any previously published subtitles in **Past Enrollment** for that video will be copied to the destination course.

Limitations of subtitle management:

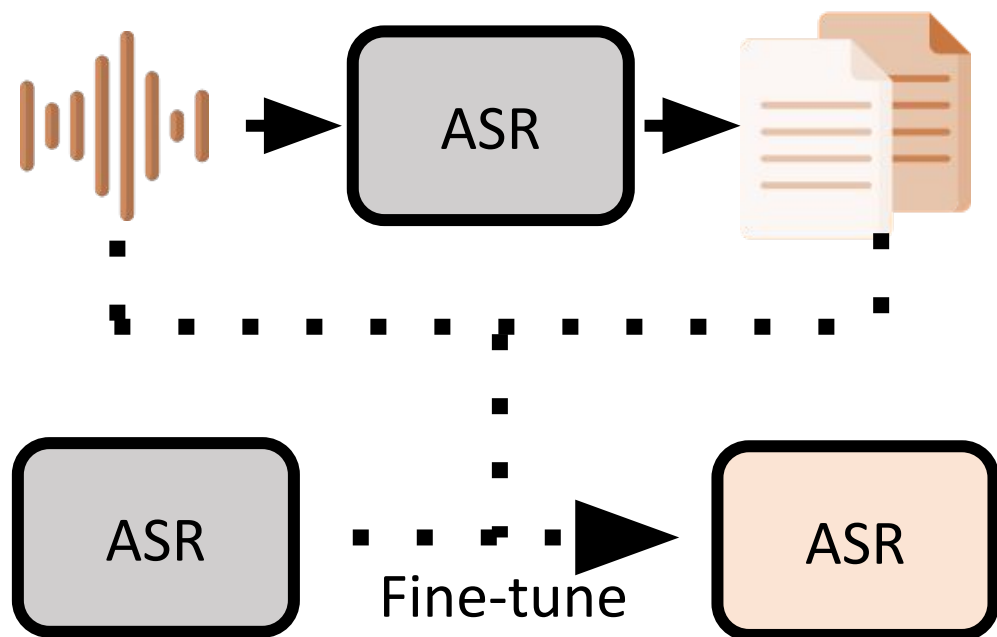
New Acoustic Domain (no corresponding text)



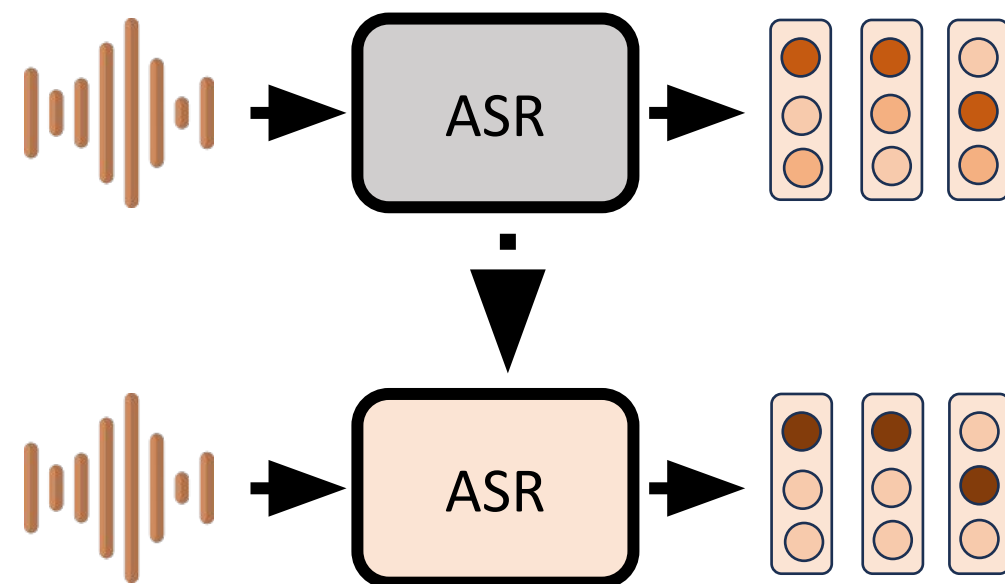
Guan-Ting Lin

<https://arxiv.org/abs/2203.14222>

Pseudo labeling



Single-Utterance Test-time Adaptation (SUTA)



Minimize
entropy, etc.

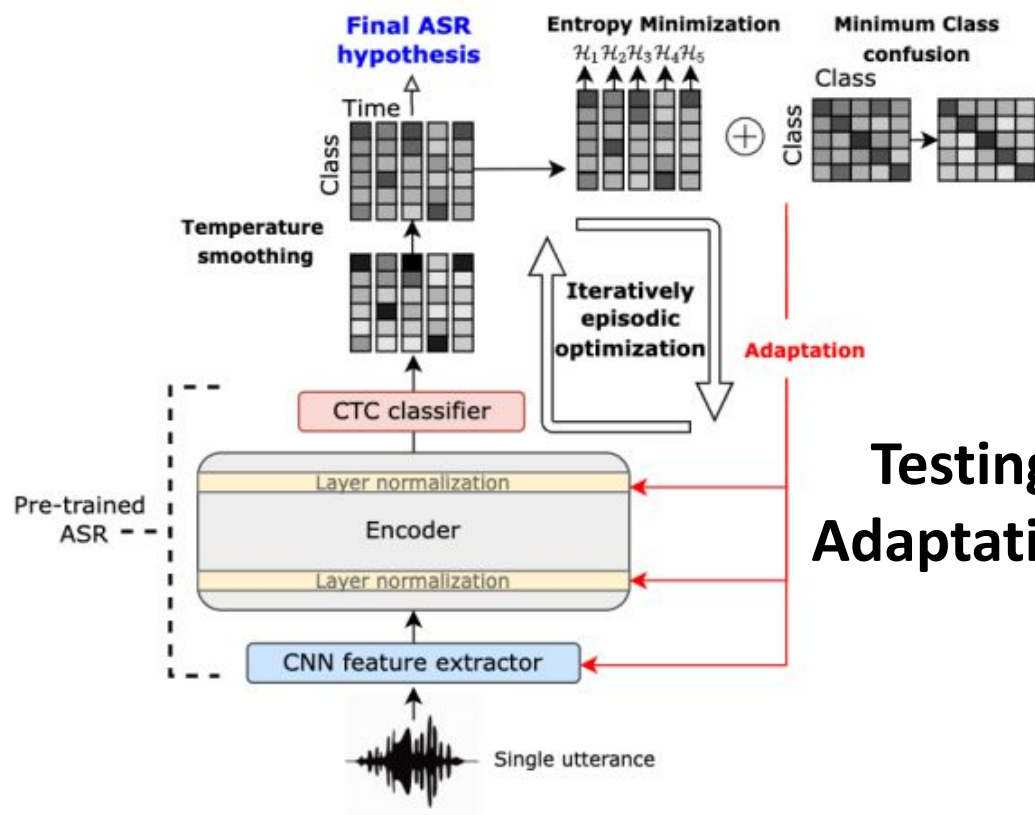
New Acoustic Domain (no corresponding text)



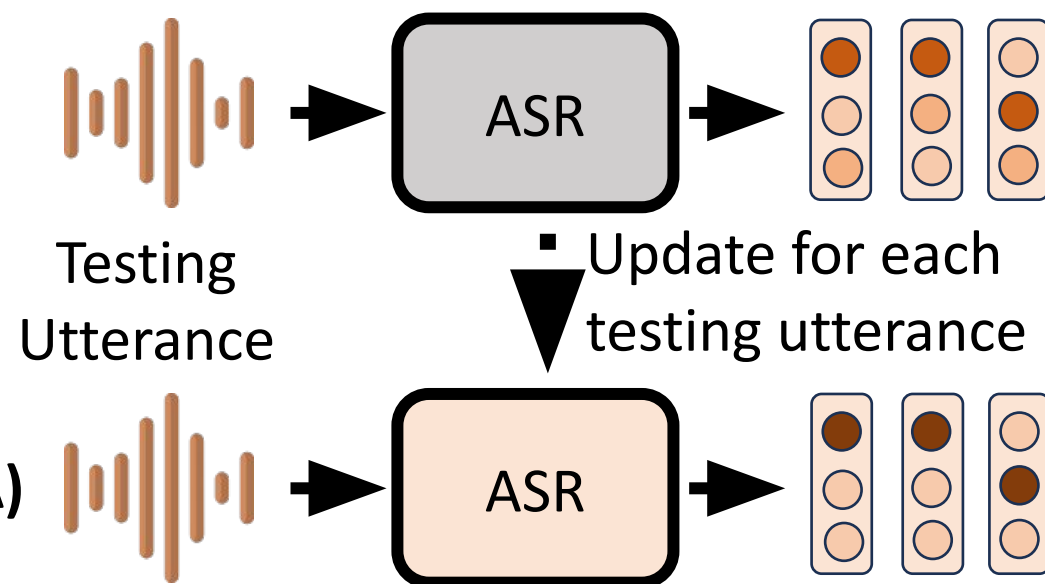
Guan-Ting Lin

<https://arxiv.org/abs/2203.14222>

Single-Utterance Test-time Adaptation (SUTA)



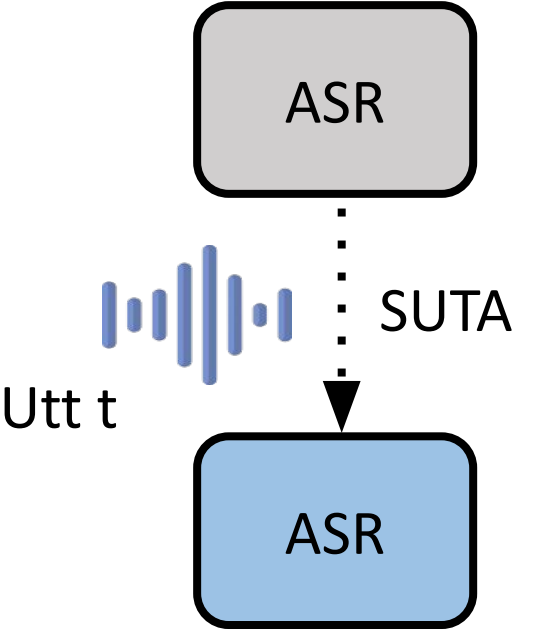
Testing Time Adaptation (TTA)



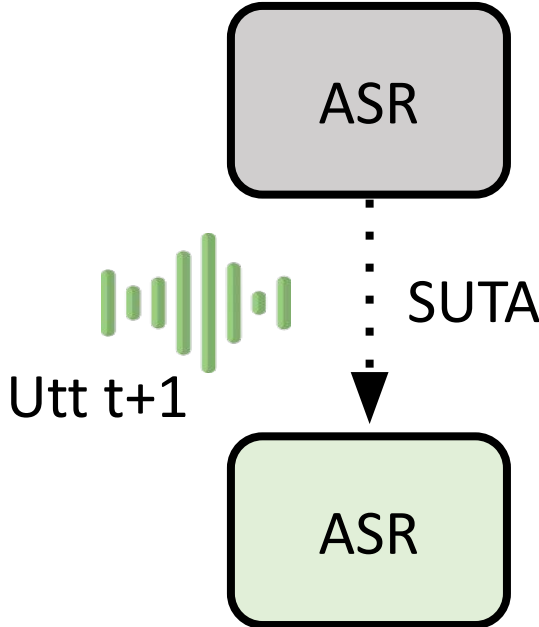
We can see improvement with a single utterance.

Minimize entropy, etc.

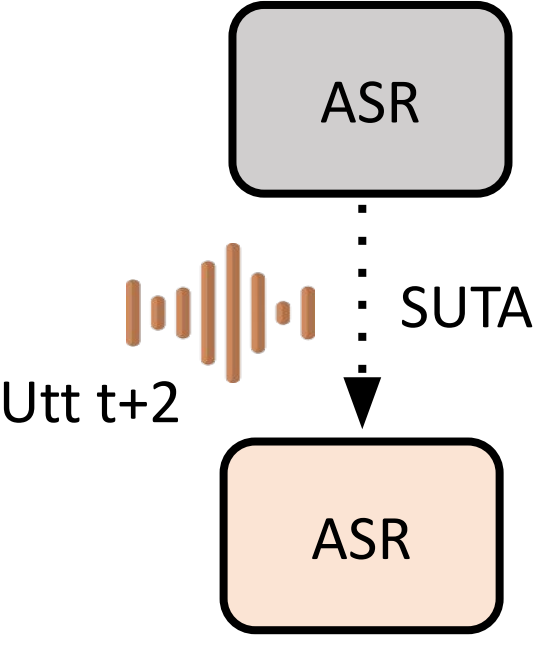
Test-time Adaptation (TTA)



Evaluate on Utt t



Evaluate on Utt t+1



Evaluate on Utt t+2

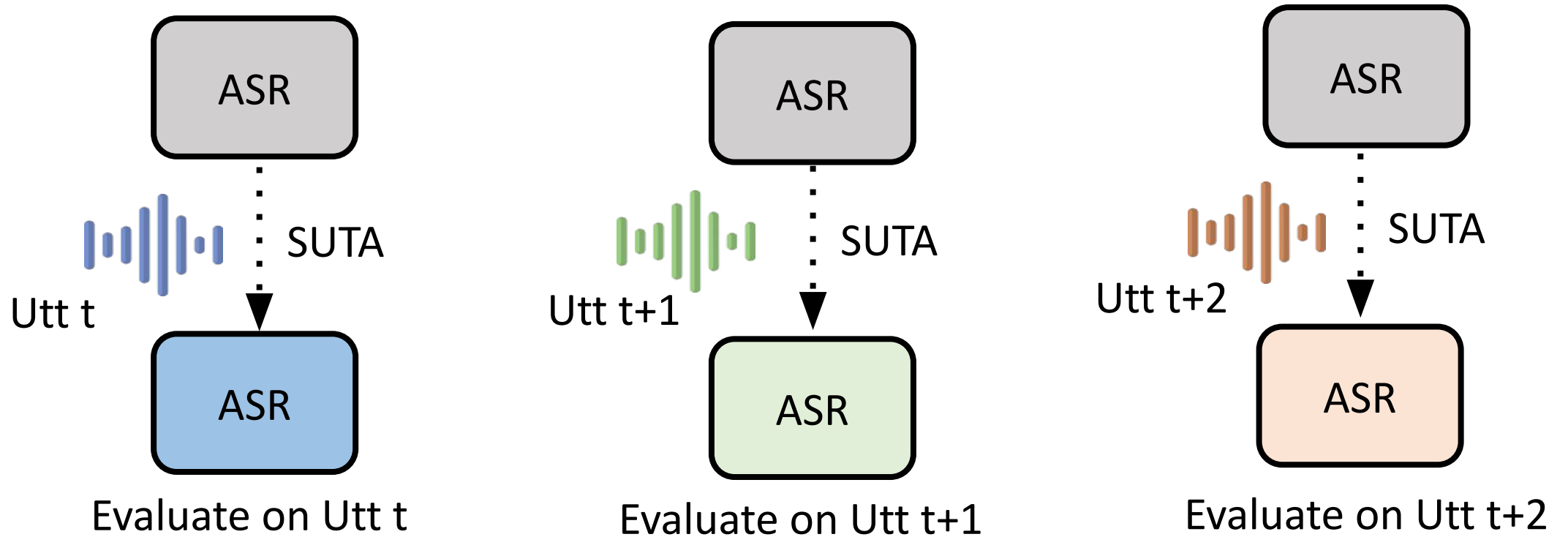
New Acoustic Domain (no corresponding text)

Testing Time Adaptation

Different domains

Performance reference for source ASR model <i>wo/ adaptation</i>	I.S test-o + δ			CH	CV	TD
	0	0.005	0.01			
SOTA (trained on target dataset)	2.5	-	-	5.8	15.4	5.6
RASR [26] (trained on LS)	6.8	-	-	-	29.9	13.0
TTA method						
(1) Our source ASR model [27] (trained on LS <i>wo/ adaptation</i>)	8.6	13.9	24.4	31.2	36.8	13.2
(1) + SDPL (Pseudo labeling)	8.3	13.1	23.1	30.4	36.3	12.8
(1) + SUTA	7.3	10.9	16.7	25.0	31.2	11.9

Limitation of Test-time Adaptation (TTA)



The ASR does not accumulate knowledge and keep improving.

Continuous TTA

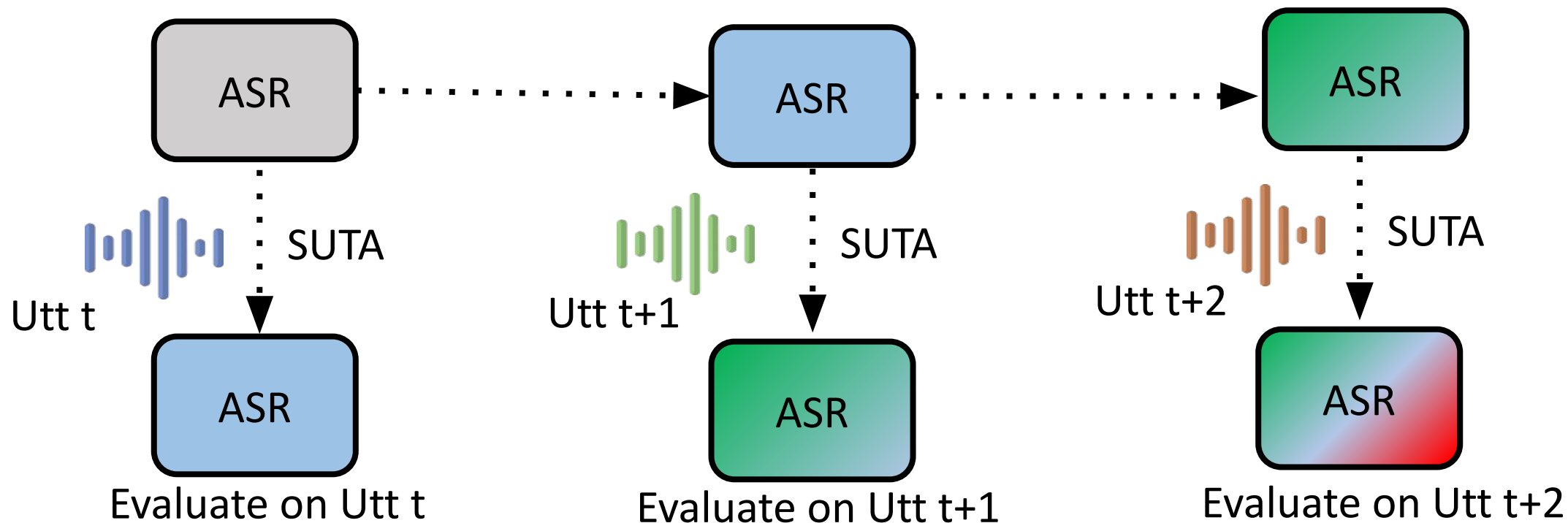
<https://arxiv.org/abs/2406.11064>



Wei-Ping Huang
(NTU)



Guan-Ting Lin
(NTU)

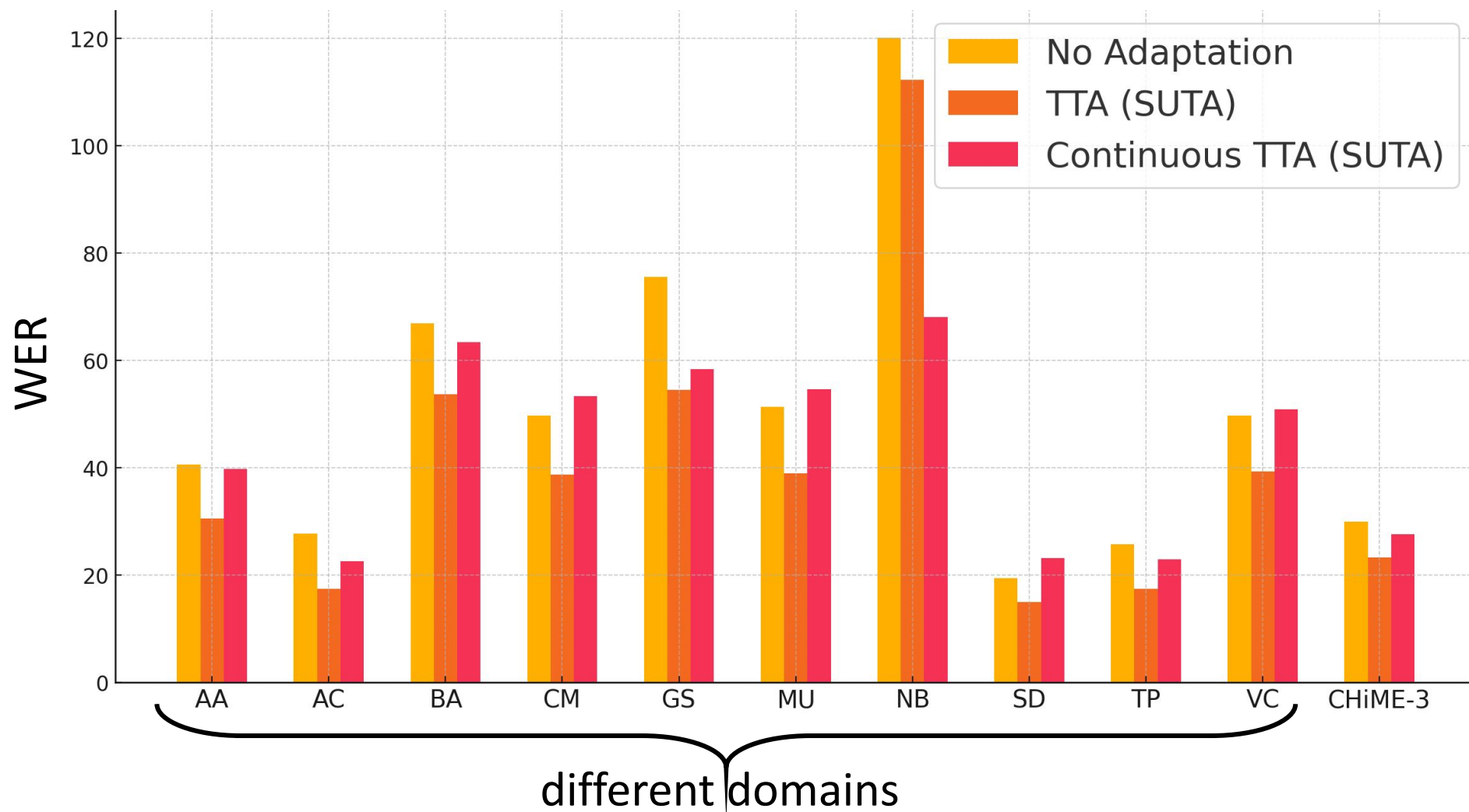


What will happen if we continuously apply SUTA?

Continuous TTA

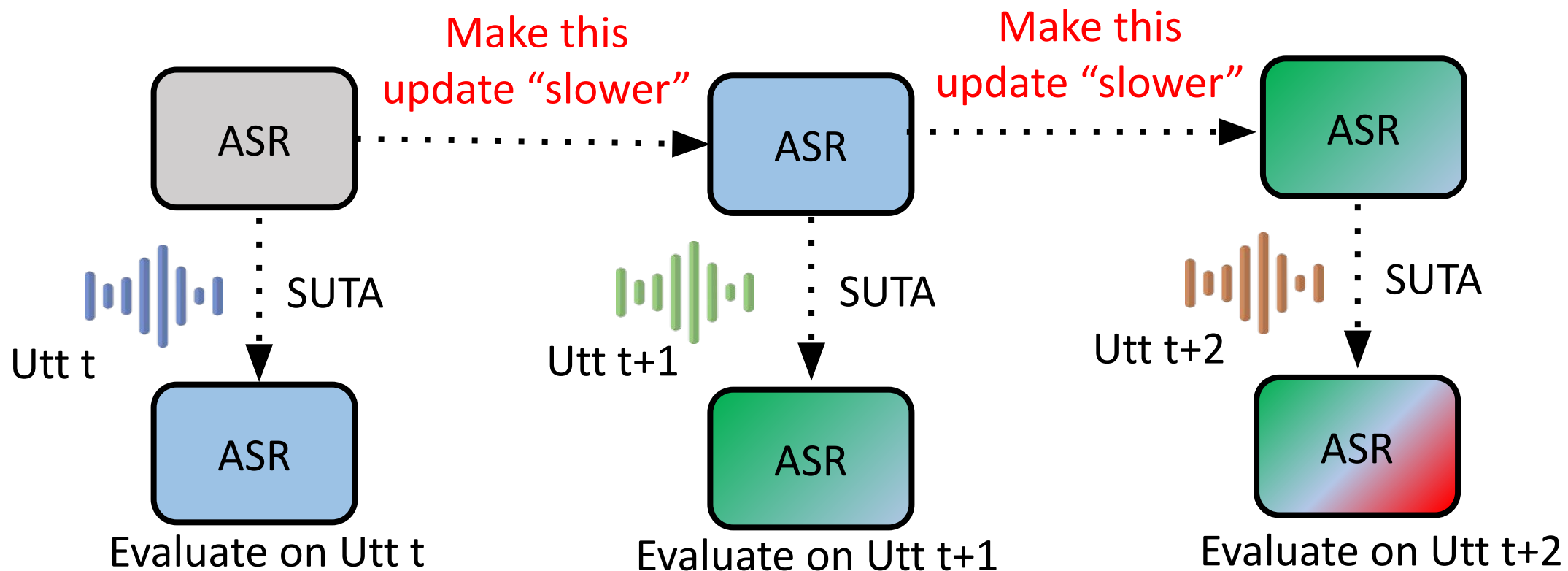
Start from: Pretrained wav2vec2.0 ASR

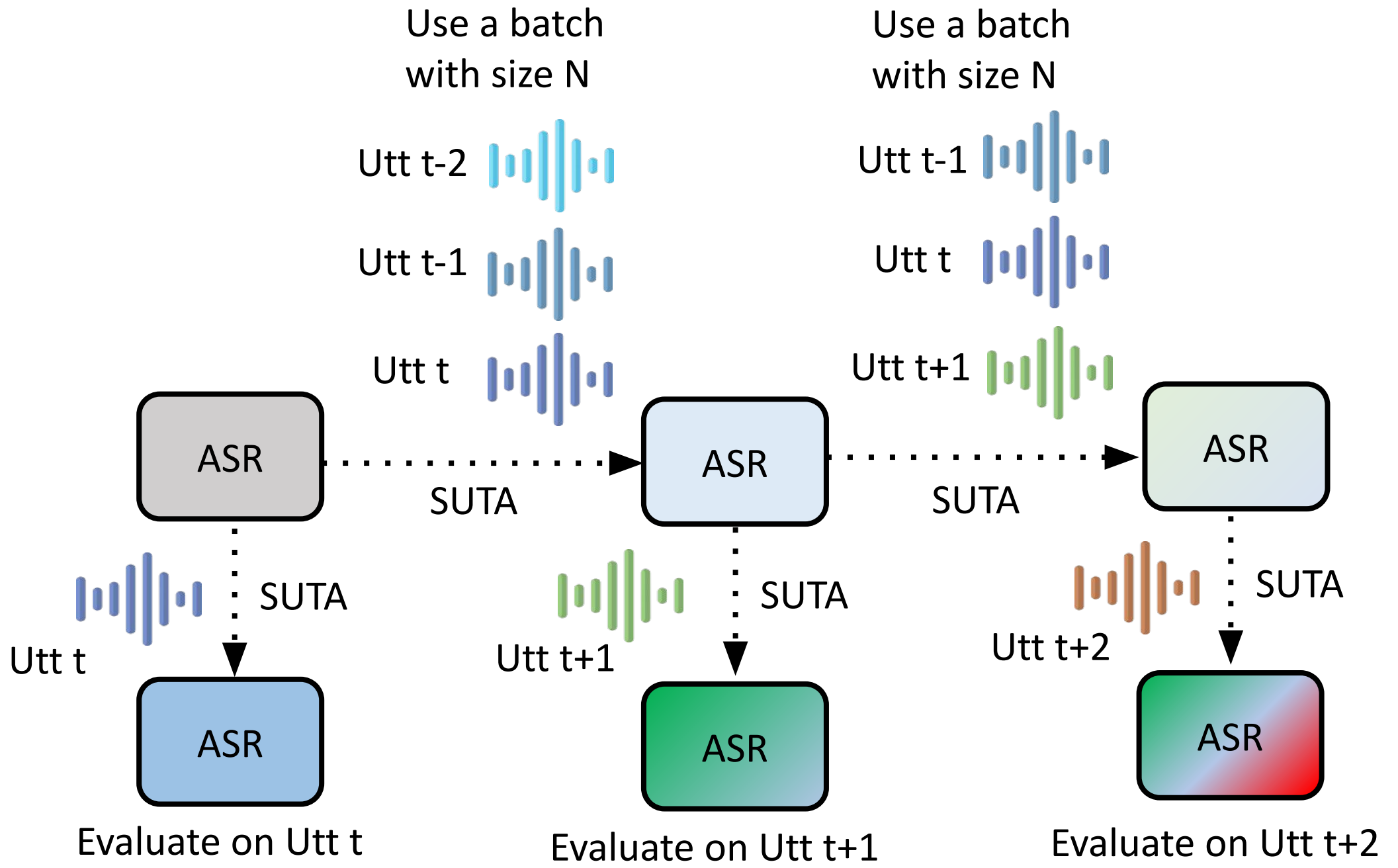
<https://arxiv.org/abs/2406.11064>



Continuous TTA

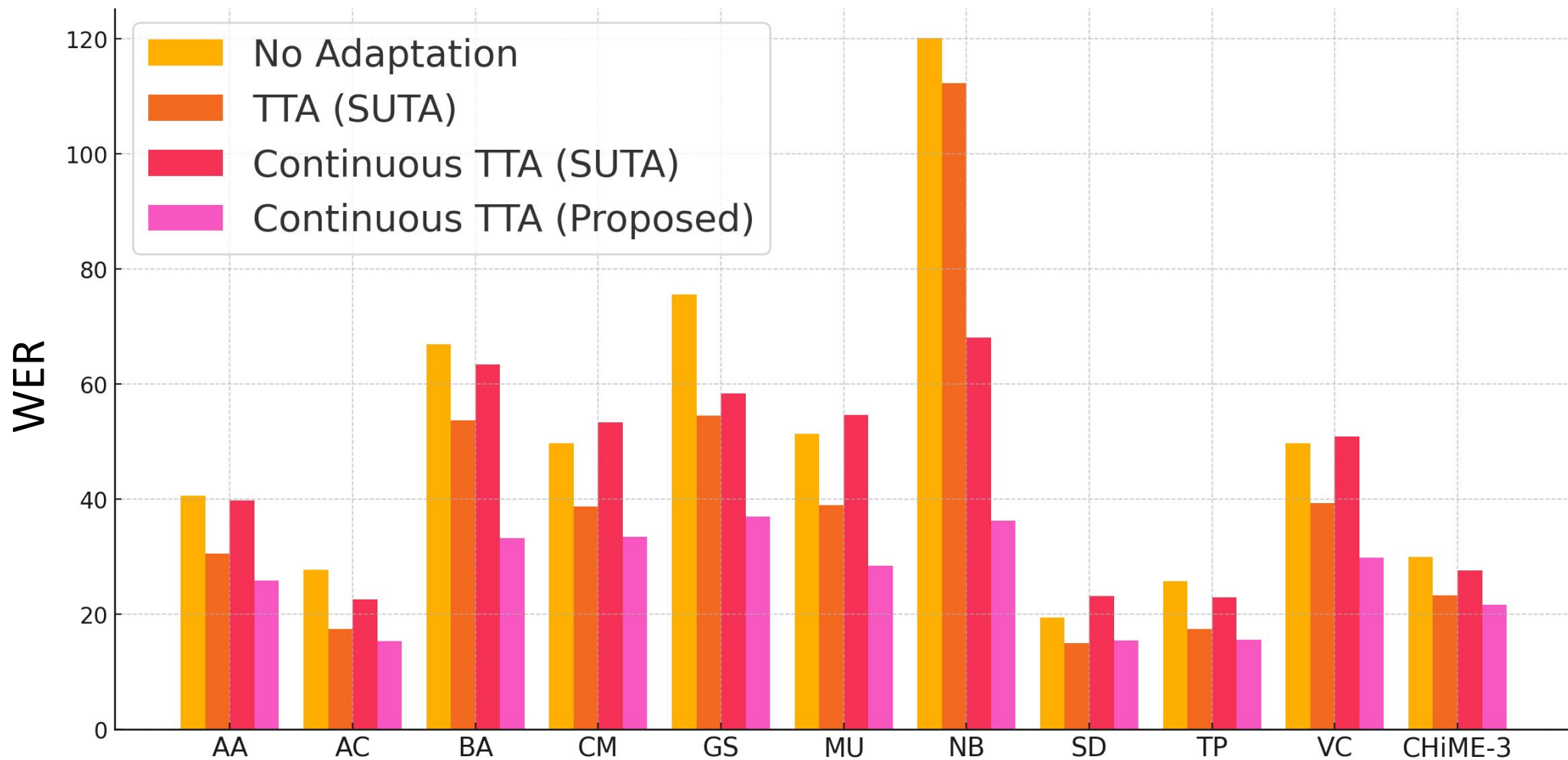
<https://arxiv.org/abs/2406.11064>

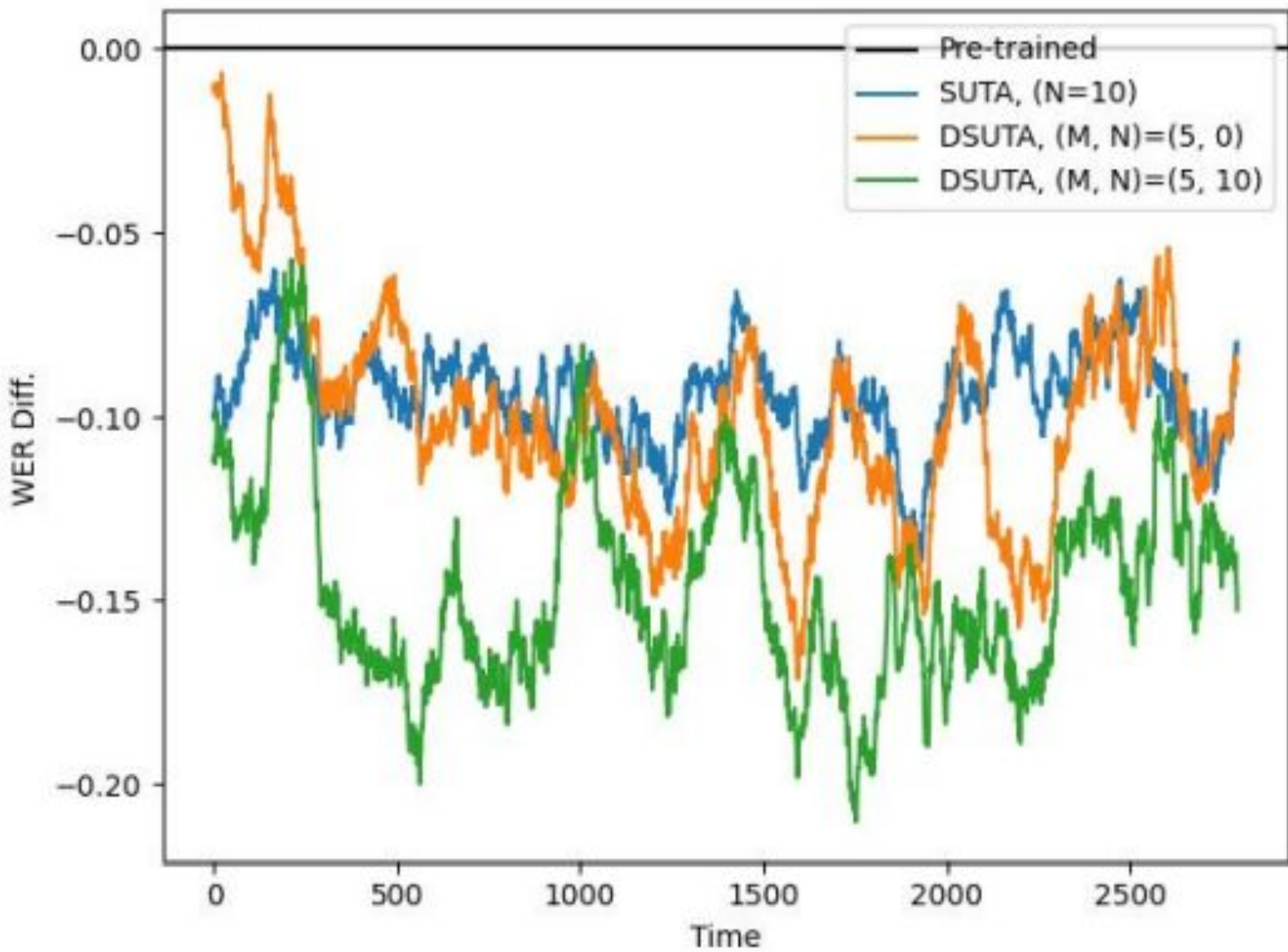




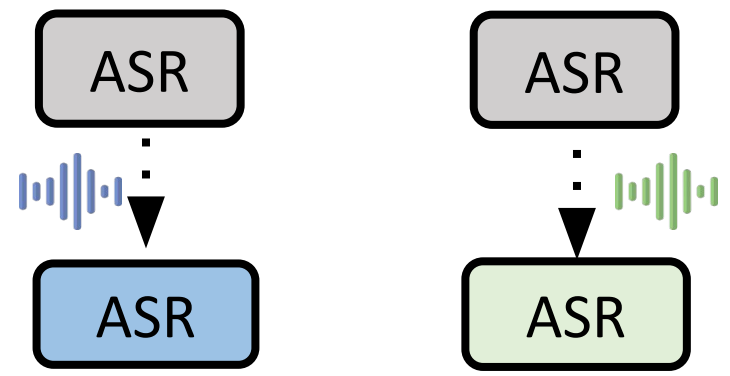
Continuous TTA

<https://arxiv.org/abs/2406.11064>

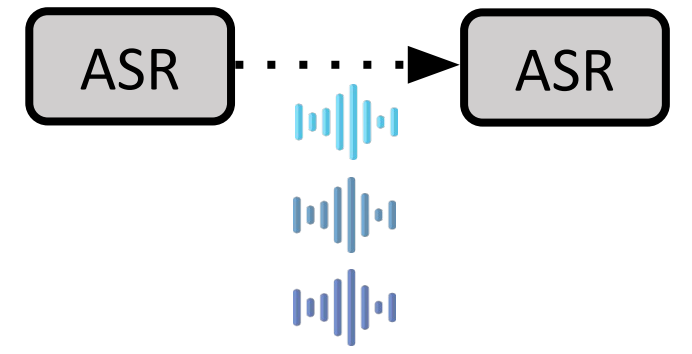




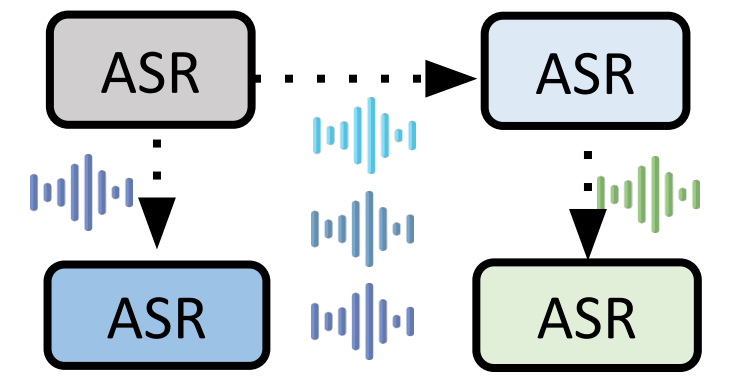
Blue Curve



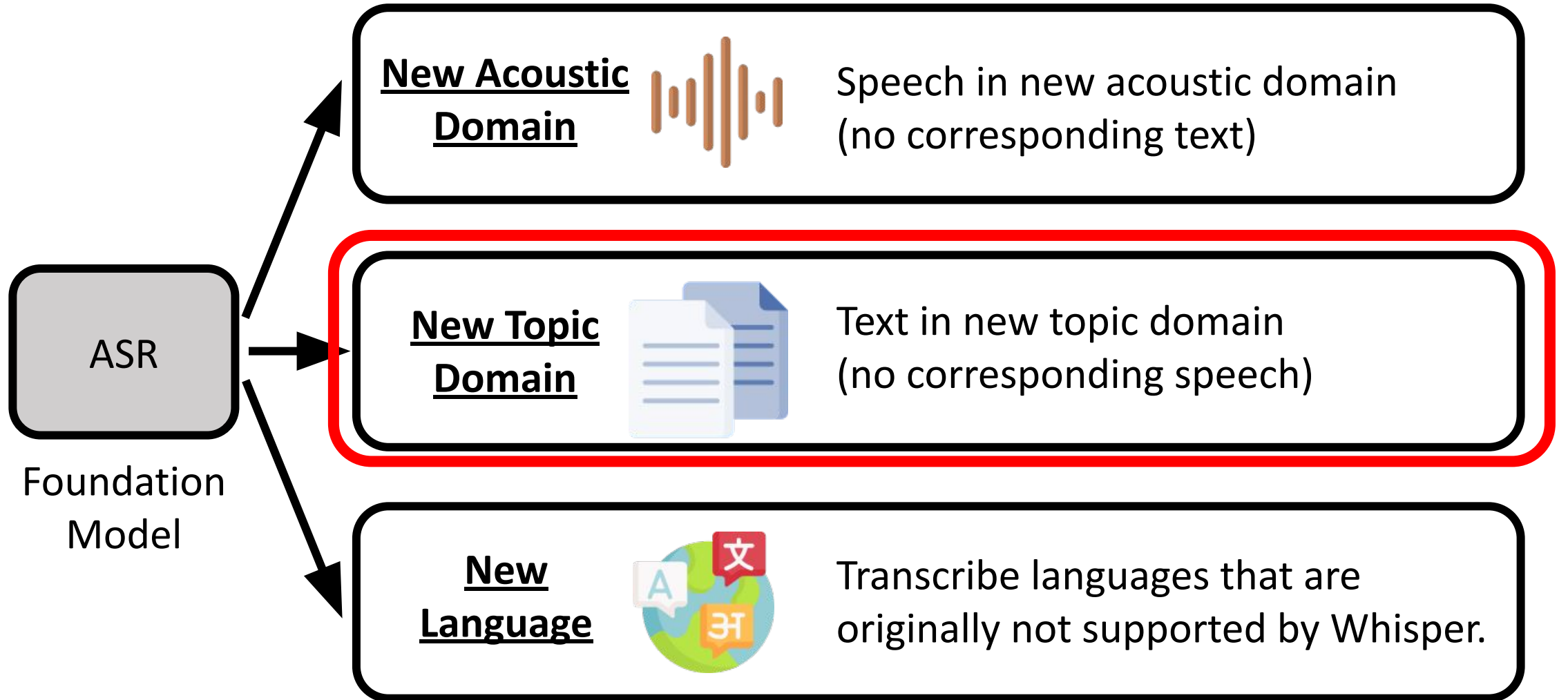
Orange Curve



Green Curve

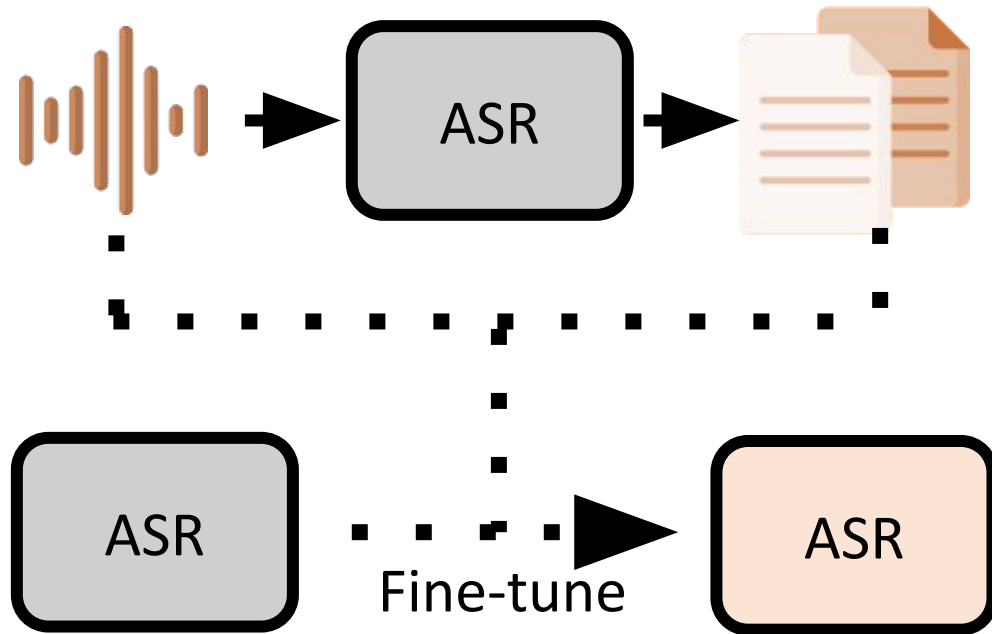


Adaptation Scenario

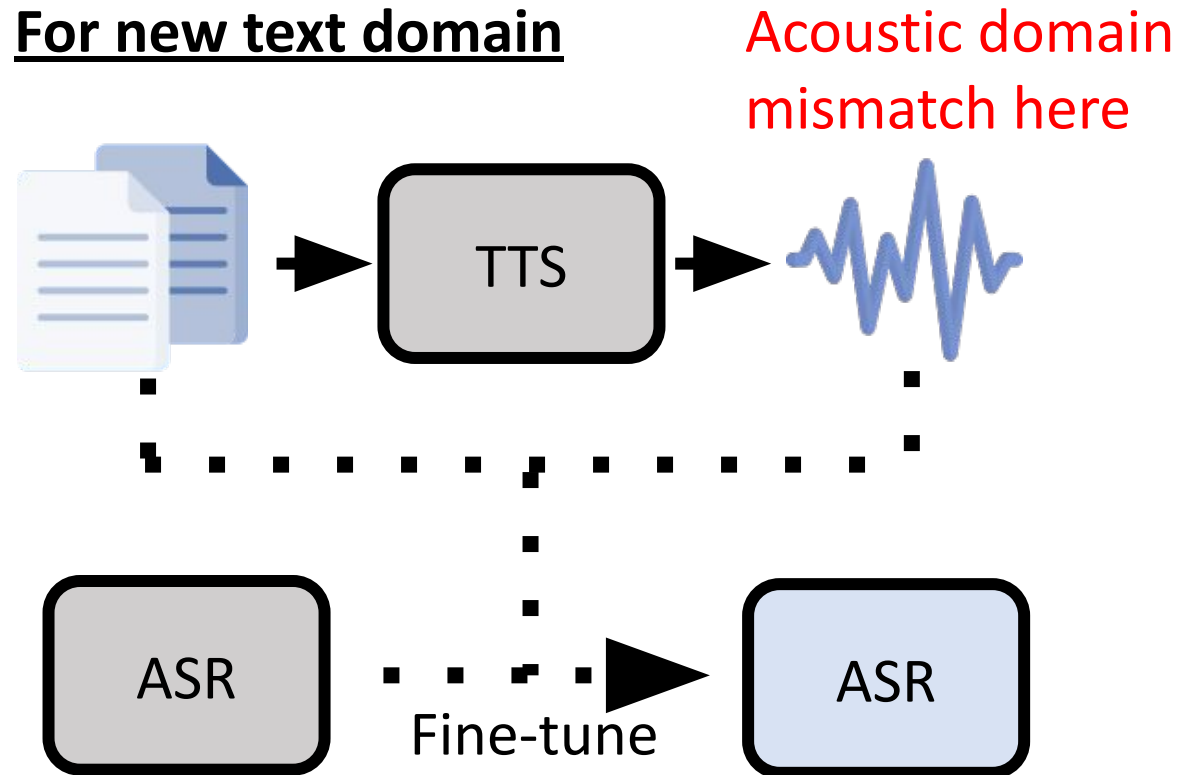


Synthesize Speech for New Text Domains

For new acoustic domain



For new text domain



<https://arxiv.org/abs/2011.11564>
<https://arxiv.org/abs/2303.14885>

<https://arxiv.org/abs/2302.14036>
<https://arxiv.org/abs/2309.10707>

Inspired from Task Vector



New



Synthesized



Source



Real Speech



Synthesized



+



-



=



Synthesic2Real Vector



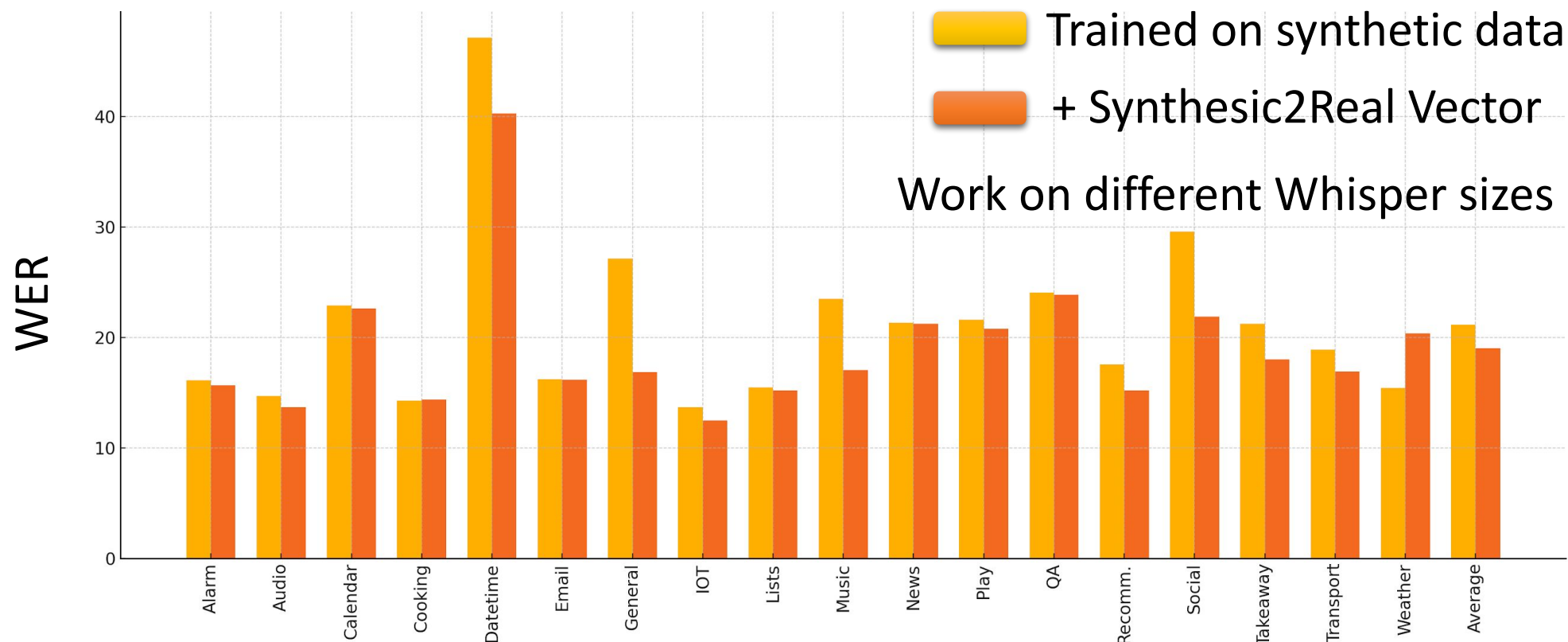
Hsuan Su (NTU)

<https://arxiv.org/abs/2406.02925>

Task Vector for ASR

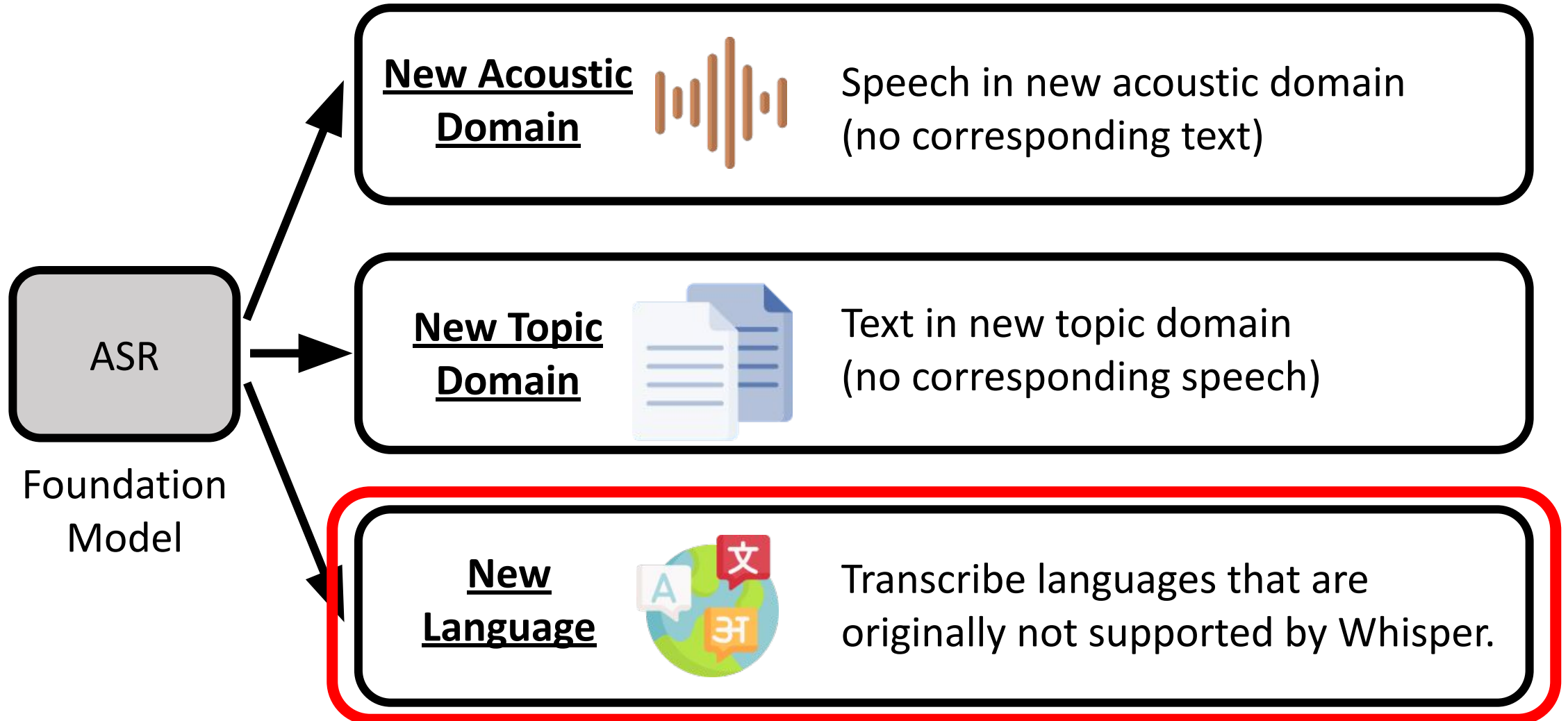
<https://arxiv.org/abs/2406.02925>

- SLURP
- Speech foundation model: Whisper
- TTS model: BARK



Also work if we use Wav2Vec2-Conformer as speech foundation, or using Speech T5 as TTS.

Adaptation Scenario



Learning New Languages

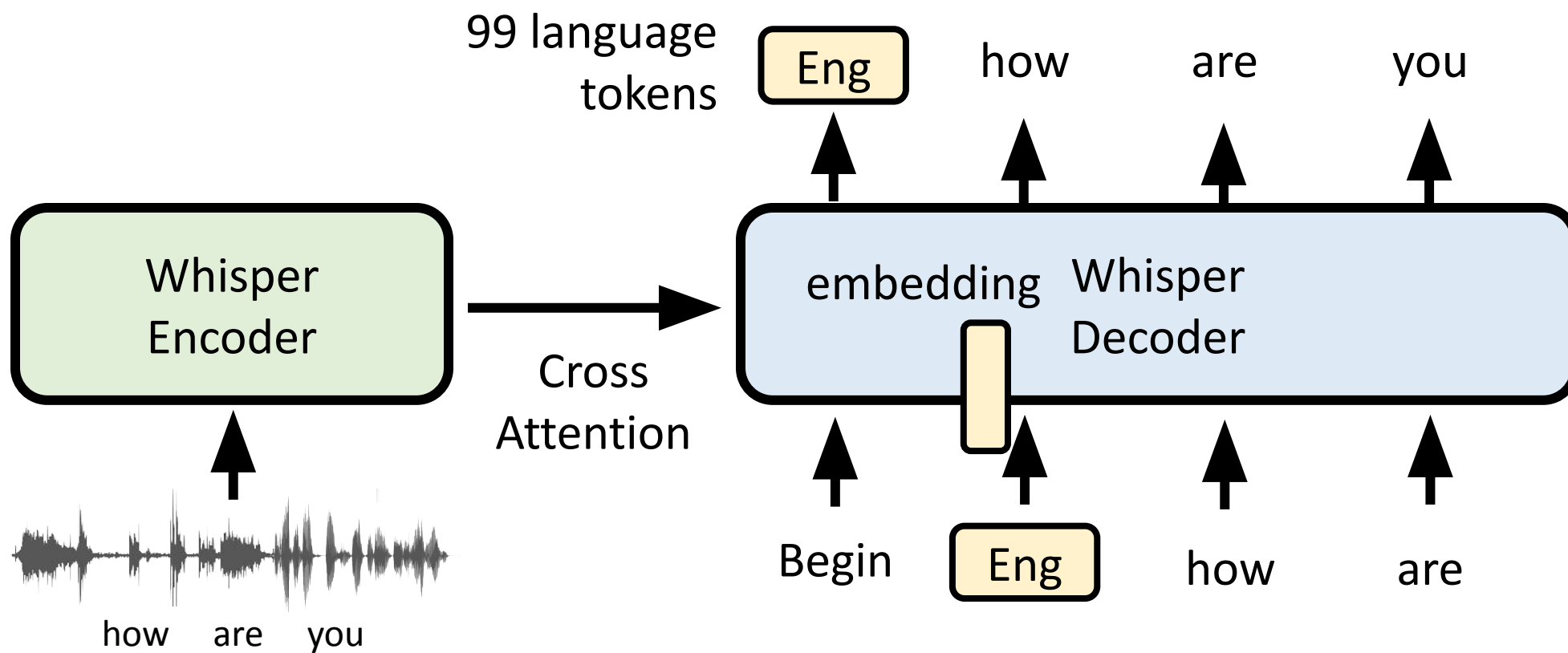


C-3PO



Ewok

Learning New Languages



Learning New Languages

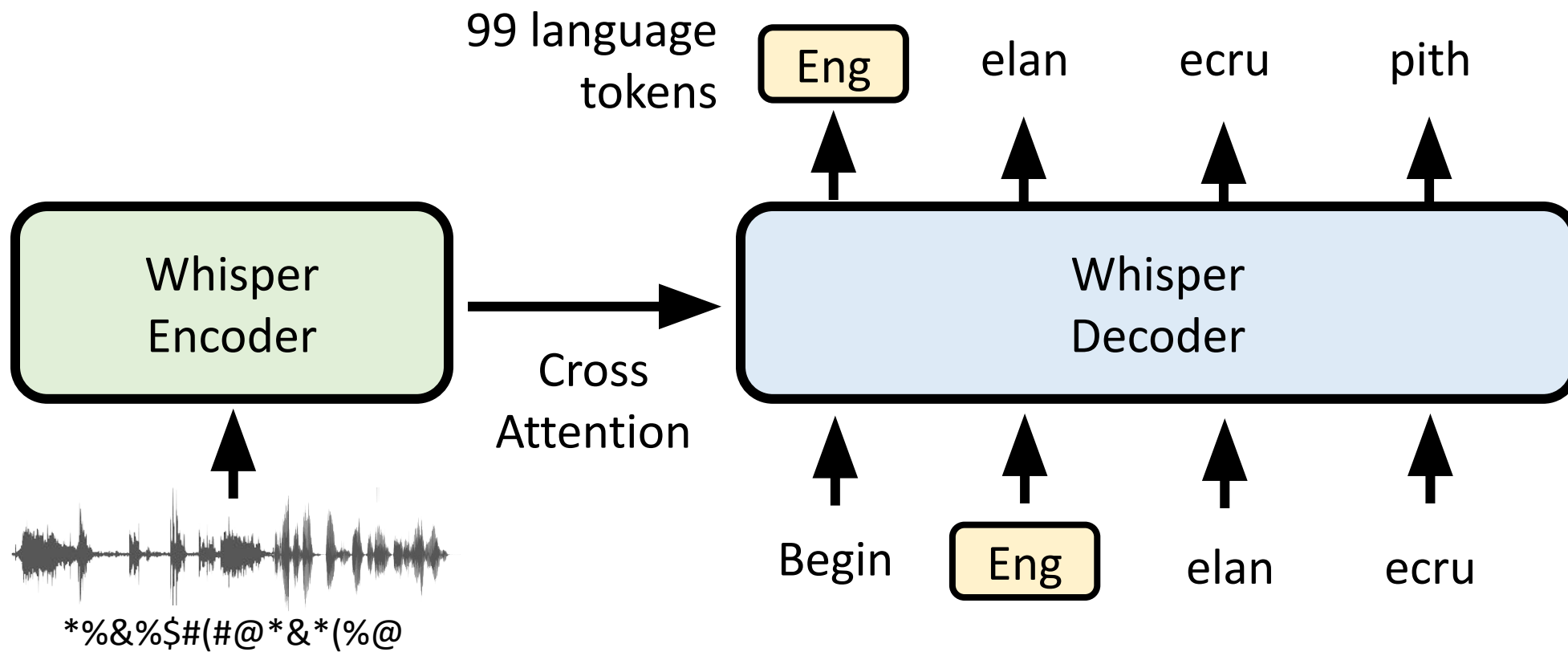
- Is the language token important?

<https://arxiv.org/abs/2406.05806>

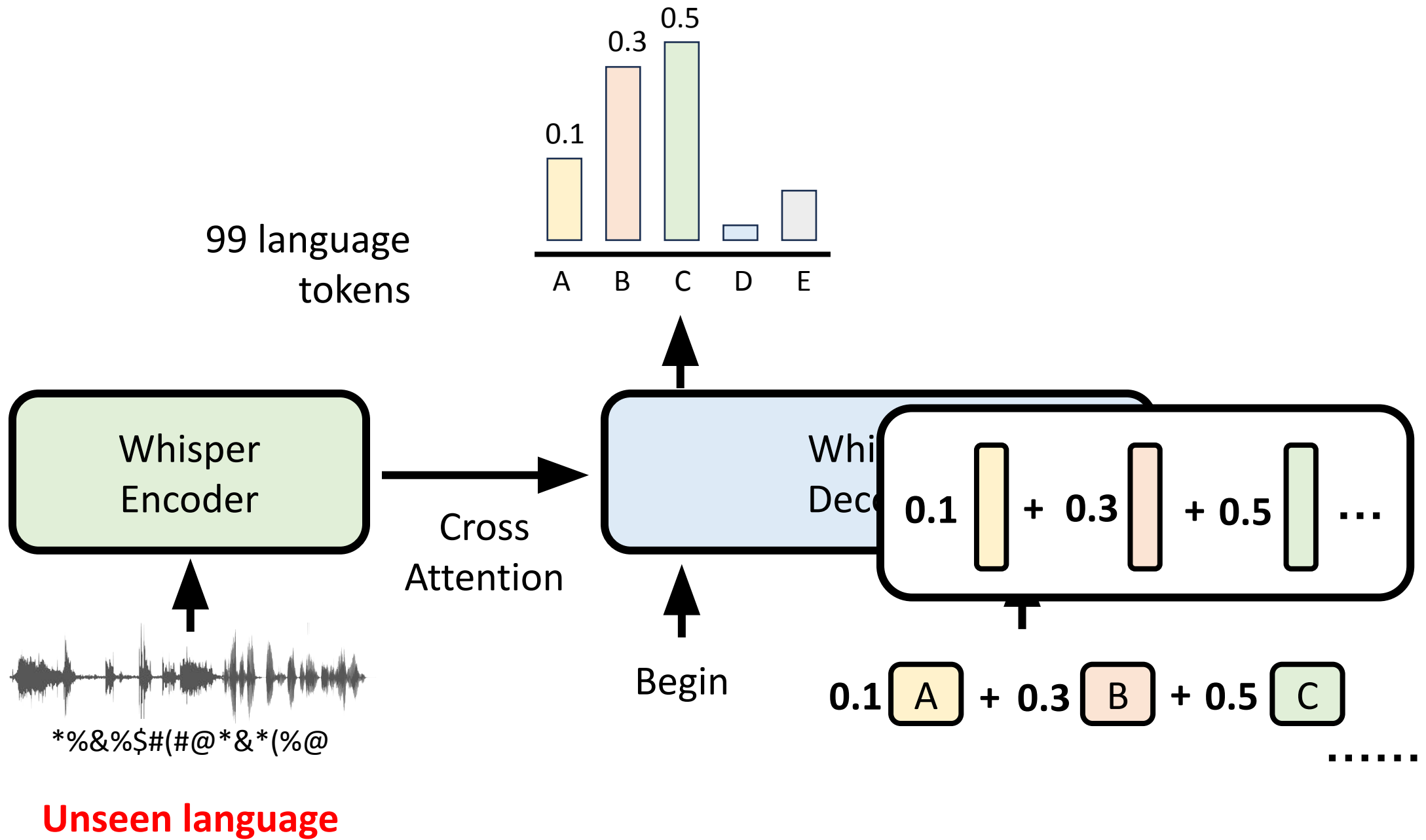
English-Mandarin
Code-Switching Datasets

	Language tokens	CSZS-correct-zh	ASCEND
Correct Language Tokens →	< zh >< en >	26.76 [22.35, 31.72]	21.93 [17.21, 29.92]
Incorrect Language Tokens →	< zh >< es >	54.24 [48.75, 60.23]	26.51 [24.39, 28.68]
Incorrect Language Tokens →	< zh >< fr >	58.85 [52.69, 65.78]	26.59 [24.46, 28.72]
Incorrect Language Tokens →	< zh >< it >	50.05 [44.39, 55.50]	30.73 [25.22, 39.81]

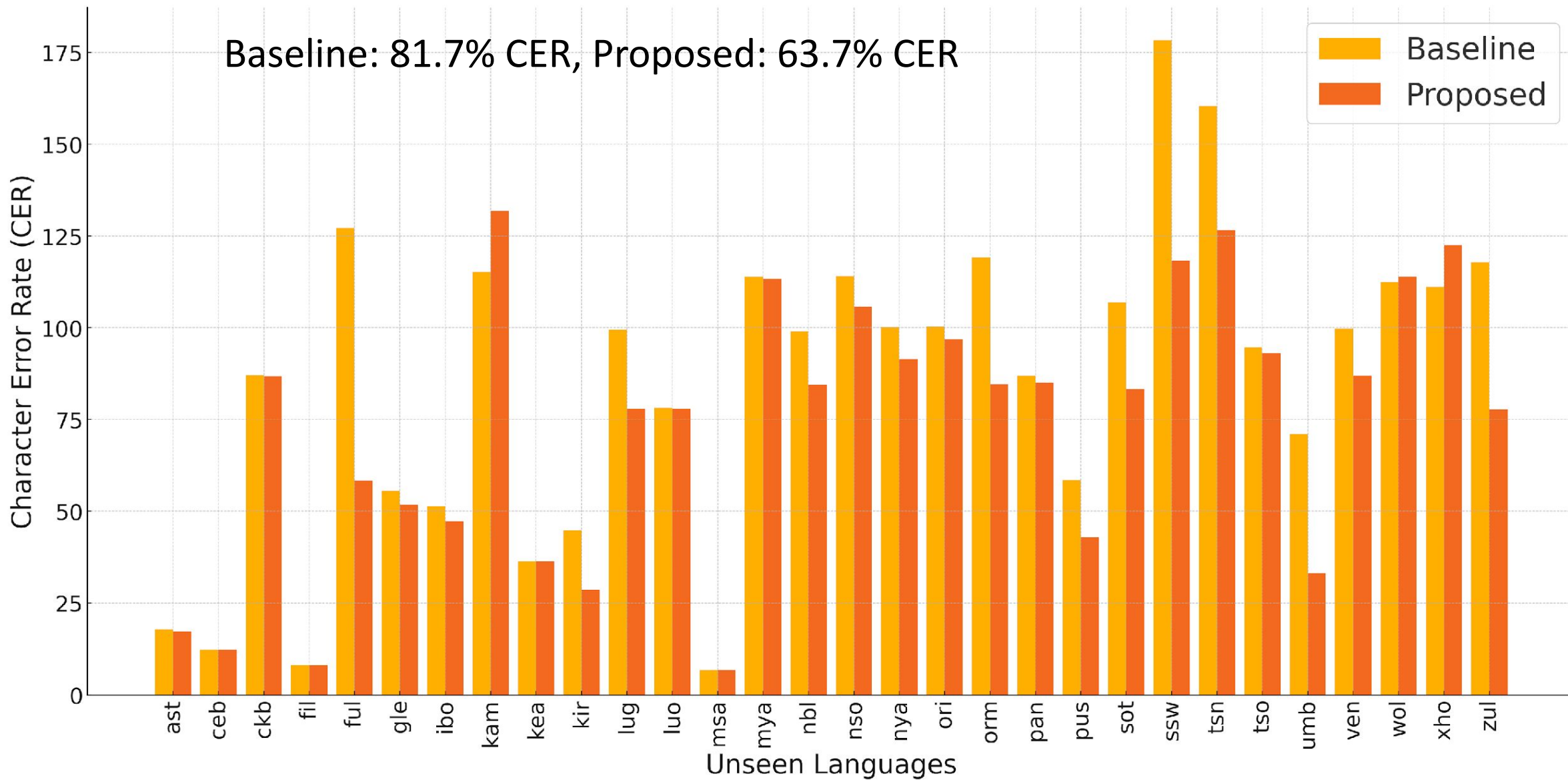
Learning New Languages



Unseen language



Baseline: 81.7% CER, Proposed: 63.7% CER



Concluding Remarks

Teaching Text LLMs (e.g., LLaMA) to
New Skills

Adapting ASR Models (e.g., Whisper) to
New Domains