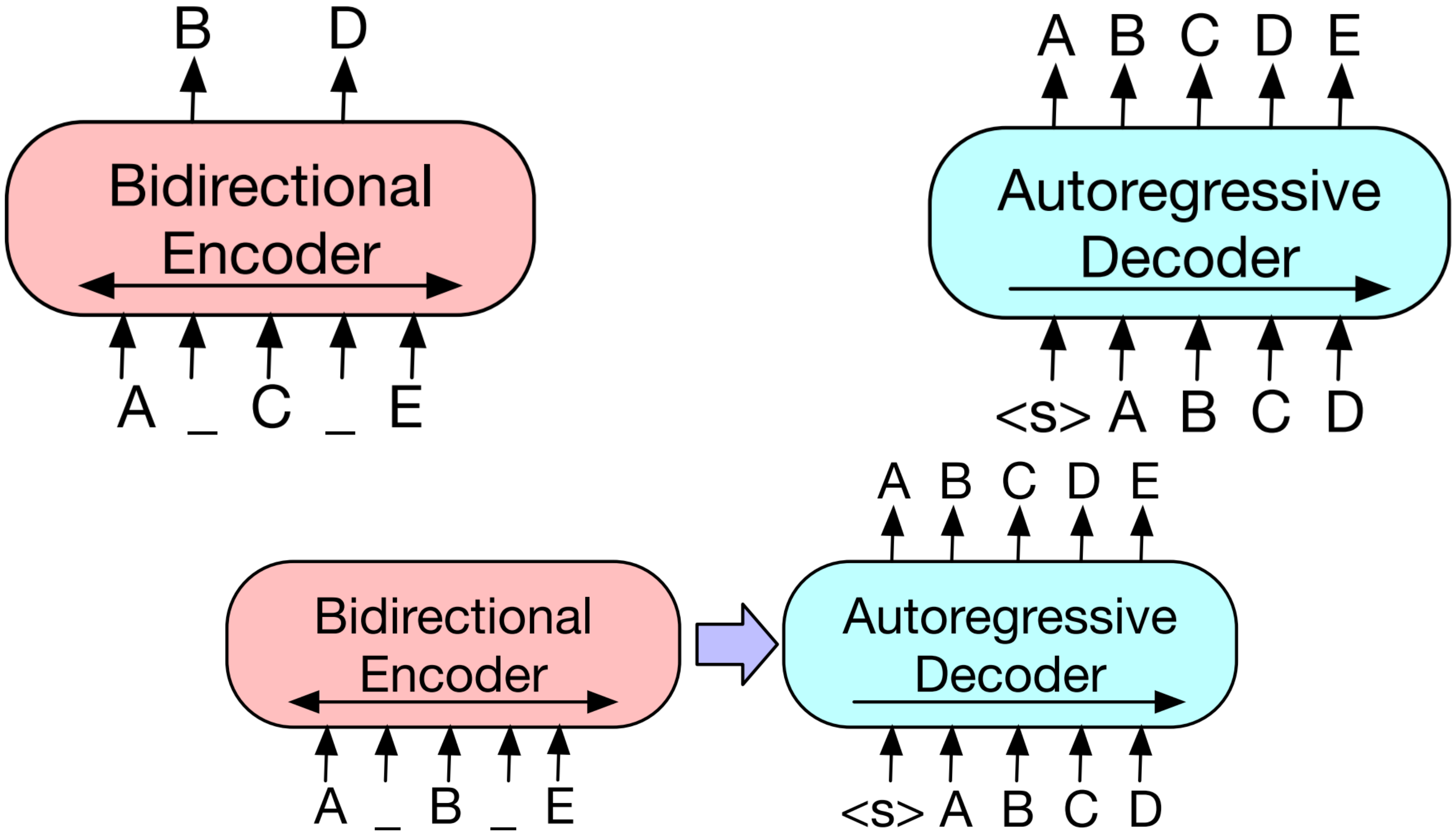
Decoder-only Modeling for Speech

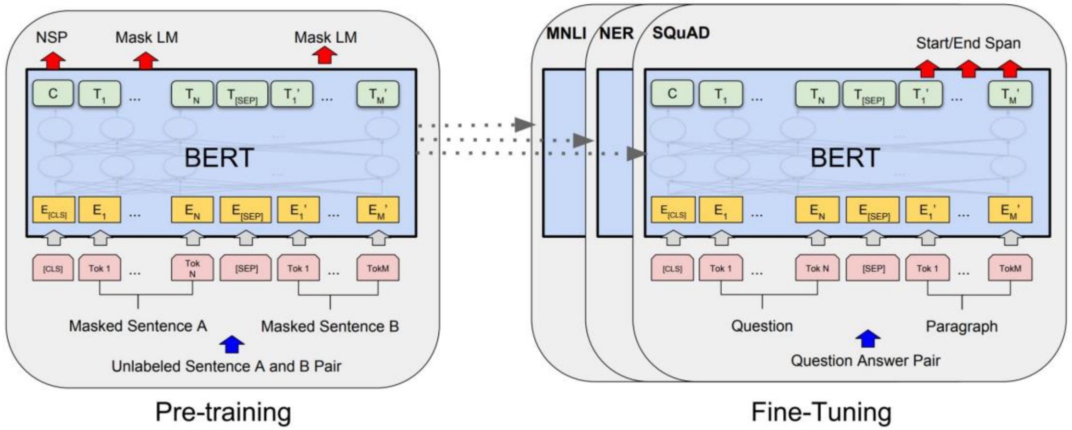
Foundation Models

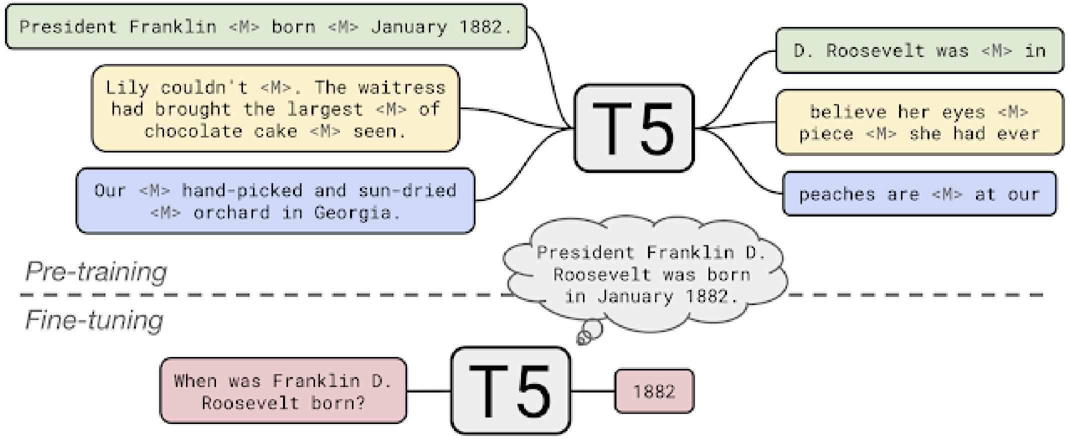


BART: Denoising Sequence-to-Sequence Pre-training for Natural

Why Decoder-only?

- Enc-only and Enc-Dec were very successful in the transfer-learning era.



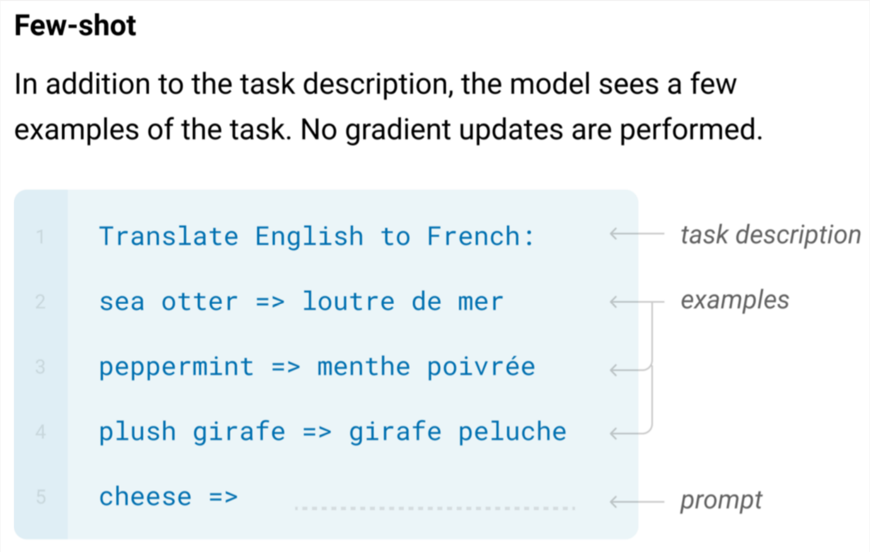


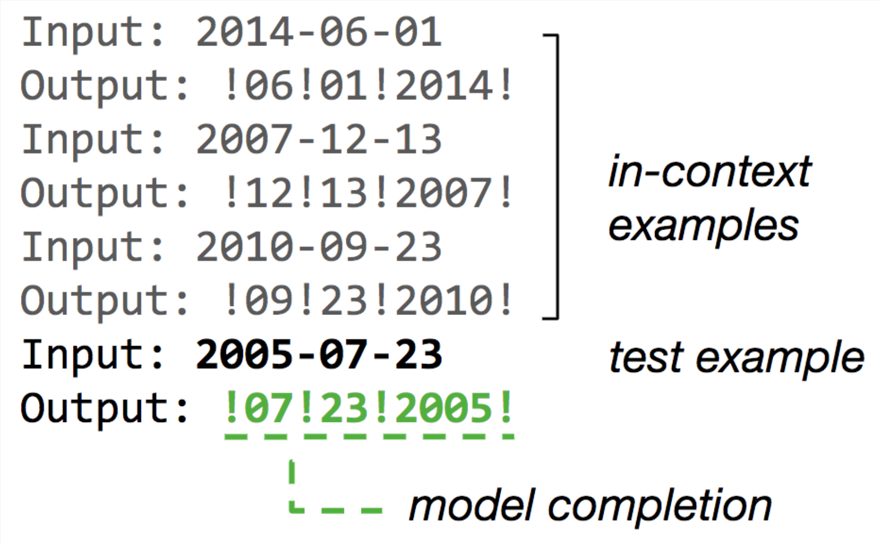
BERT: Pre-training of Deep Bidirectional Transformers & T5:
Text-To-Text Transfer Transformer

Why Decoder-only?

- Dec-only models exhibited zero-shot or few-shot generalization (in-context learning)

- A task could be anything that the prompt could describe.





Large Language Model (LLM)-centric AI

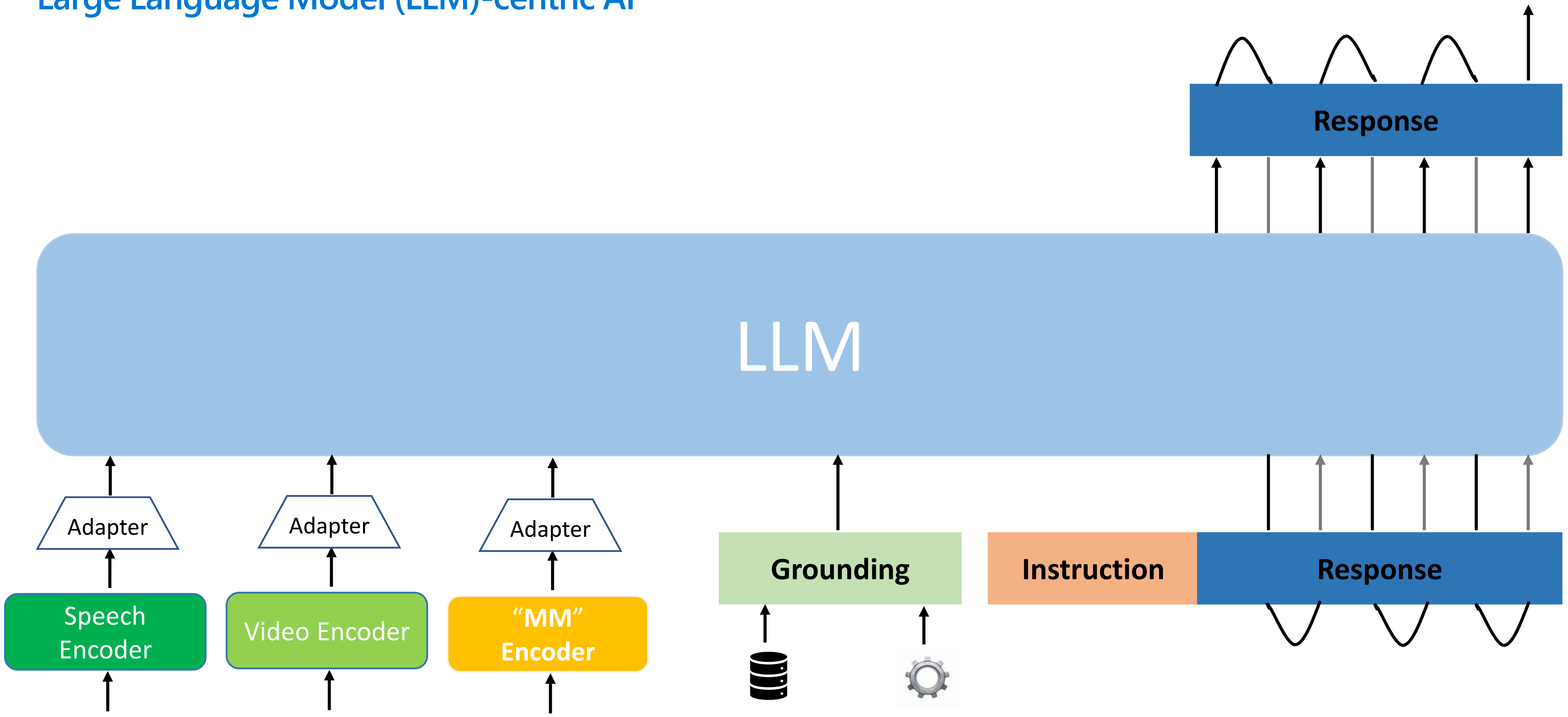


Image credit to XD Huang

What does GPT-V mean to computer vision?

TASK **Logo Recognition**

GPT-V The image shows a close-up of a person's foot wearing a white **Nike Air Force 1 sneaker**. The sneaker is a low-top style with a white rubber sole and white laces. …

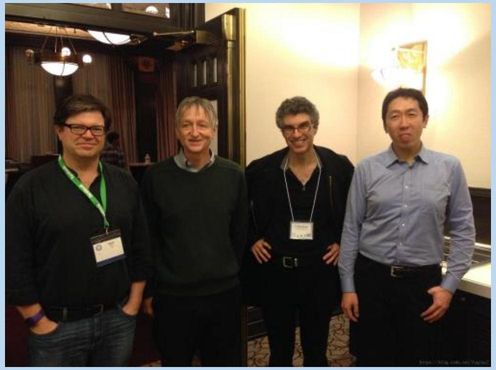
TASK **Object Counting**

GPT-V There are **7 apples** in the image.

TASK **Object Localization/Detection**

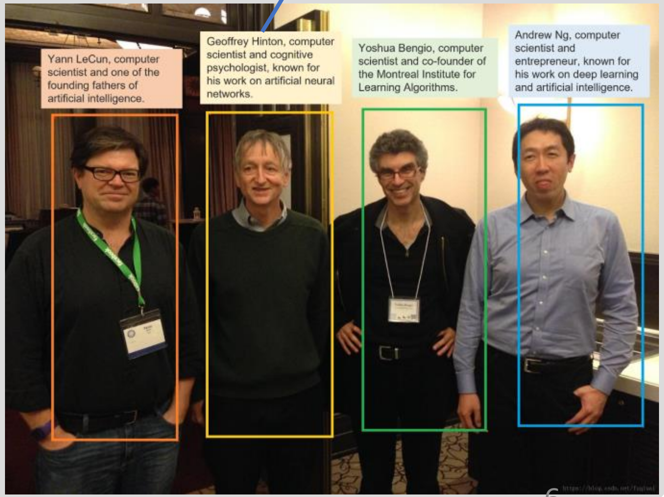
GPT-V There is only one person in the image, so the bounding box for this person would be:
- Top left: (254, 118)
- Bottom right: (731, 693)

TASK **Dense Captioning**



GPT-V

Geoffrey Hinton, computer scientist and cognitive psychologist, known for his work on artificial neural networks.



Yang, Z., et al. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). *arXiv:2309.17421*, 2023.

Image credit to Lijuan Wang

What does GPT-V mean to computer vision?

TASK Image Recognition

GPT-V The object in the image is a **traffic light**.

TASK Celebrity Recognition

GPT-V The person in the image is **Jensen Huang**, the CEO and co-founder of NVIDIA. He is holding and presenting a product from NVIDIA, likely a graphics processing unit (GPU).

TASK Landmark Recognition

GPT-V The landmark in the image is the **Space Needle**, an observation tower located in Seattle, Washington. It stands 605 feet tall and offers panoramic views of the city and surrounding areas. …

TASK Food Recognition

GPT-V This dish is called a **Cajun Sampler**. It is a sampler platter that consists of four small bowls of traditional Cajun dishes - gumbo, jambalaya, red beans and rice, and crawfish étouffée. …

Image credit to Lijuan Wang

What does GPT-V mean to computer vision?

TASK **Scene Text Recognition**

GPT-V The scene text in the image are:

- **Vacation** - **Beach** - **Smile**
- **Summer** - **Watermelon**
- **Flip flops** - **Lemonade**

TASK **Document Reasoning**

Which country has the highest profit margin?

Country	Sales Volume	Revenue	Profit	Profit Margin
USA	40.080	$15.971.880	$3.086.421	19,3%
China	35.070	$15.866.670	$3.032.162	19,1%
Australia	27.054	$14.812.566	$2.868.636	19,4%
India	23.046	$10.608.174	$1.853.710	17,5%
South Korea	16.032	$10.494.948	$1.975.844	18,8%
Total / Avg	141.282	$67.754.238	$12.816.772	18,8%

GPT-V Based on the data provided, the country with the highest profit margin is **Australia**, with a profit margin of 19,4%.

TASK **Multilingual Image Description**

GPT-V 这是一只小**柯基犬**，它在草地上走着，嘴巴张开，舌头伸出来，好像很开心的样子。它的身上有一条项圈，上面挂着一个**骨头形状的牌子**。

TASK **Multilingual Text Recognition**

GPT-V The text in the image is in Chinese and translates to "**Mapo Tofu**." Mapo tofu is a popular Chinese dish made with tofu and minced meat in a spicy sauce.

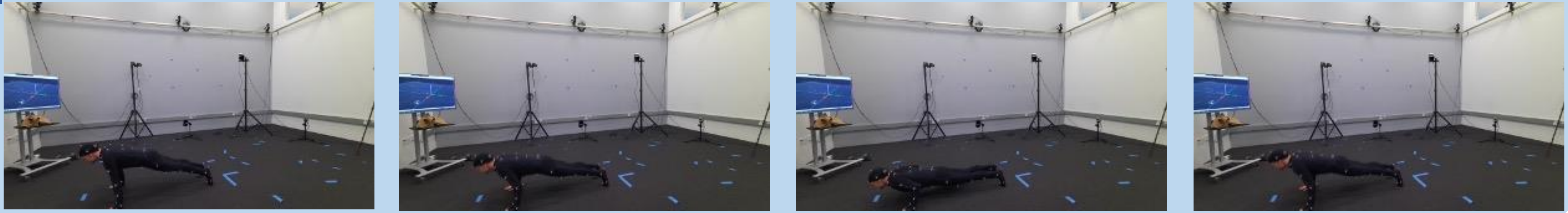
Image credit to Lijuan Wang

What does GPT-V mean to computer vision?

TASK **Video Activity Anticipation**

GPT-V The images show the process of making sushi rolls. The first image shows rice being mixed with vinegar in a bowl. The second image shows the rice being spread onto a sheet of seaweed. The third image shows the addition of salmon and cucumber to the roll. The next steps would be to **roll the sushi using a bamboo mat, cut the roll into smaller pieces, and serve with soy sauce and wasabi**.

TASK **Video Action Recognition**



GPT-V The images show a person in a motion capture suit in a studio with multiple cameras and sensors. The person is **performing a push-up exercise**. In the first image, the person is in the starting position with their arms extended and body in a straight line. In the second image, the person is lowering their body towards the ground. In the third image, the person is at the bottom of the push-up with their chest close to the ground. In the fourth image, the person is pushing their body back up to the starting position.

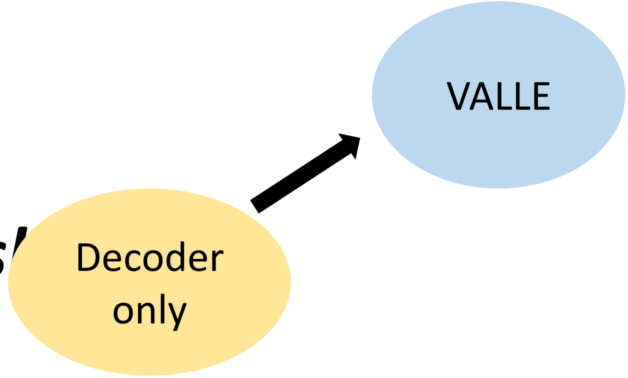
Image credit to Lijuan Wang

Why Decoder-only *for speech*?

- We are starting to see a variety of tasks, across modalities, converge to Decoder-only architecture. They use similar
  - Architectures
  - training & pre-training objectives
  - Inference methods
- Might better leverage off-the-shelf pretrained LLMs
  - Which are also decoder-only..
- Put us in a better place for doing more effective in-context learning
- Runtime is not complicated
  - Can leverage runtime developed for serving (L)LMs.

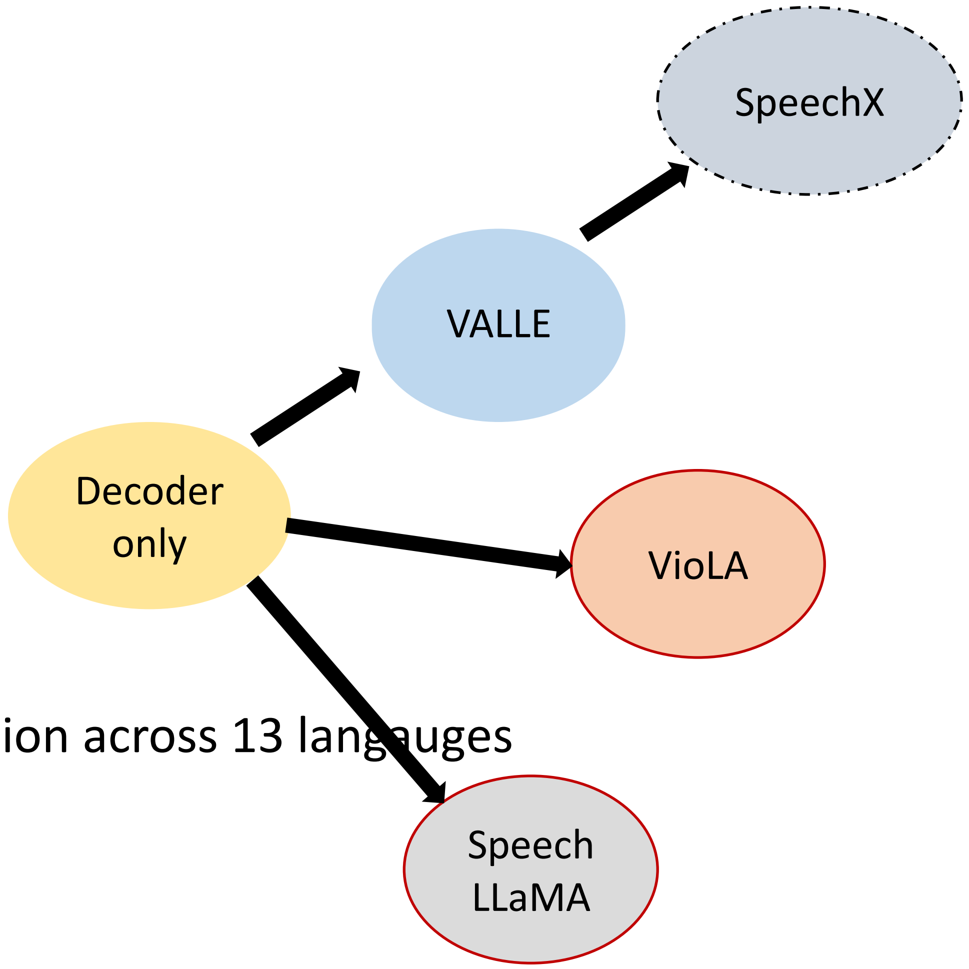
Decoder-only for speech modeling

- VALL-E:
  - Neural codec LM
  - TTS
  - Strong in-context abilities
- Decoder-only arch ***for all S & L tas***
  - Param efficiency?
  - Data efficiency?
  - Learn across tasks?
  - Fusion with off-the-shelf LLM?



Decoder-only for speech modeling

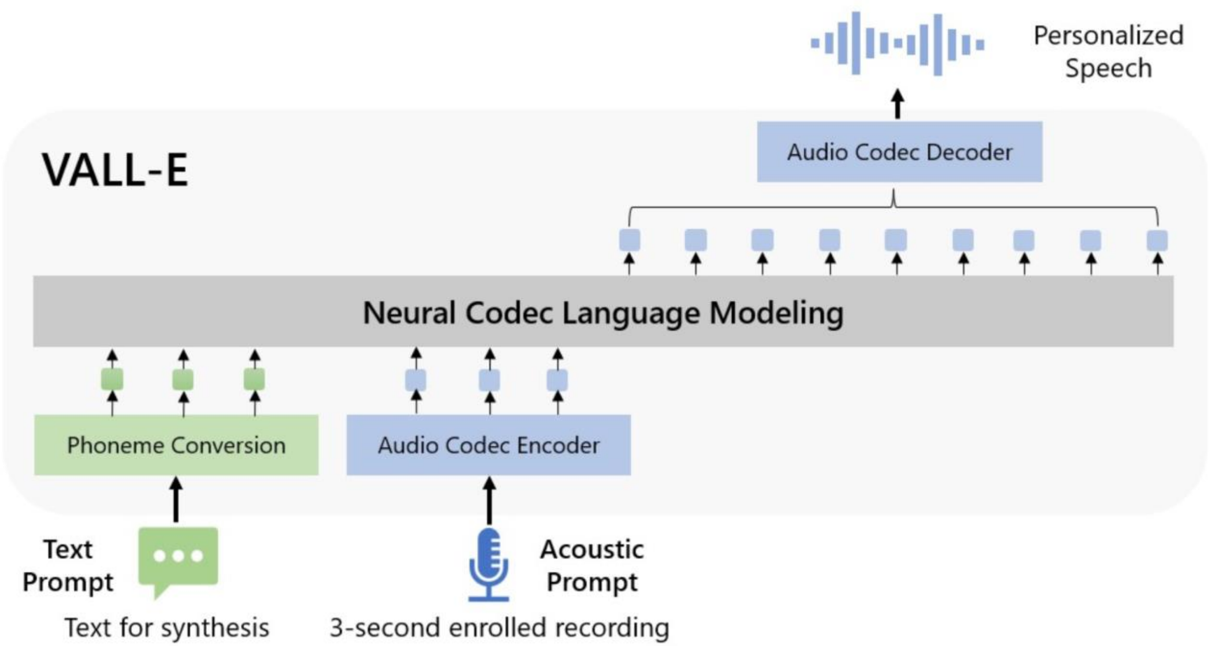
- 2 new decoder-only models:
  - VioLA:
    - Multi-task decoder-only model
    - ASR, TTS, MT, ST, S2ST
    - Benefits from cross task learning
    - Better parameter efficiency
  - SpeechLLaMA
    - Seamless Integration with LLM
    - Better performance on speech translation across 13 languages



VALL-E: Neural codec language model

- High quality Zero shot TTS: In context learning through prompts
    - “Steal voice from 3 second's prompt”

Model Overview



Prompt		Output
	I like hamburger but I love noodles much more	

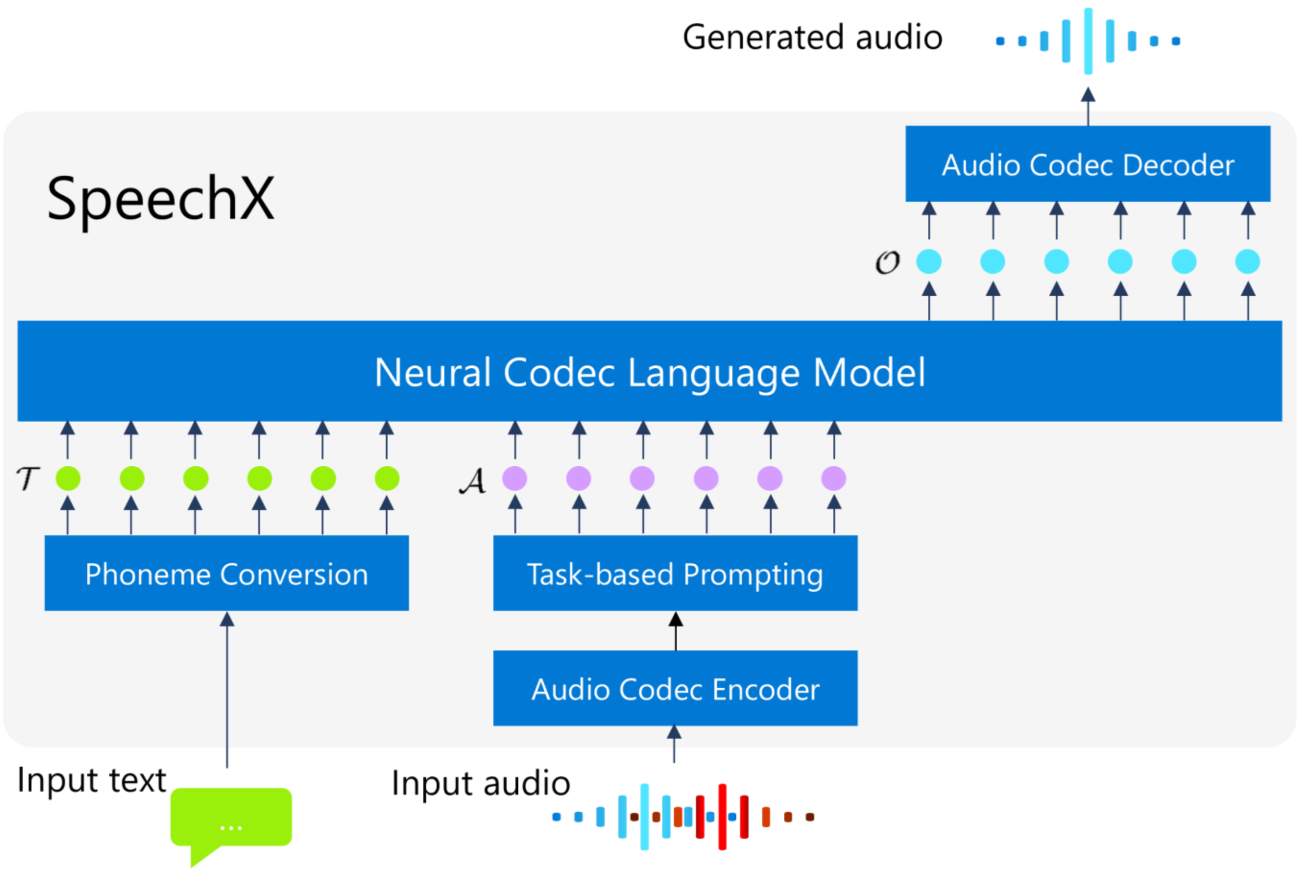
Wang, C., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv:2301.02111, 2023*.

SpeechX – A versatile speech generation model

Versatility: able to handle a wide range of tasks from audio and text inputs.

Robustness: applicable in various acoustic distortions, especially in real-world scenarios where background sounds are prevalent.

Extensibility: flexible architectures, allowing for seamless extensions of task support.



Task	Input text	Input audio	Output audio
Noise suppression	Transcription (optional)	Noisy speech	Clean speech
Speech removal	Transcription (optional)	Noisy speech	Noise
Target speaker extraction	Transcription (optional)	Speech mixture, Enrollment speech	Clean speech of target speaker
Zero-short TTS	Text for synthesis	Enrollment speech	Synthesized speech mimicking target speaker
Clean speech editing	Edited transcription	Clean speech	Edited speech
Noisy speech editing	Edited transcription	Noisy speech	Edited speech with original background noise

[More demo samples: SpeechX - Microsoft Research]

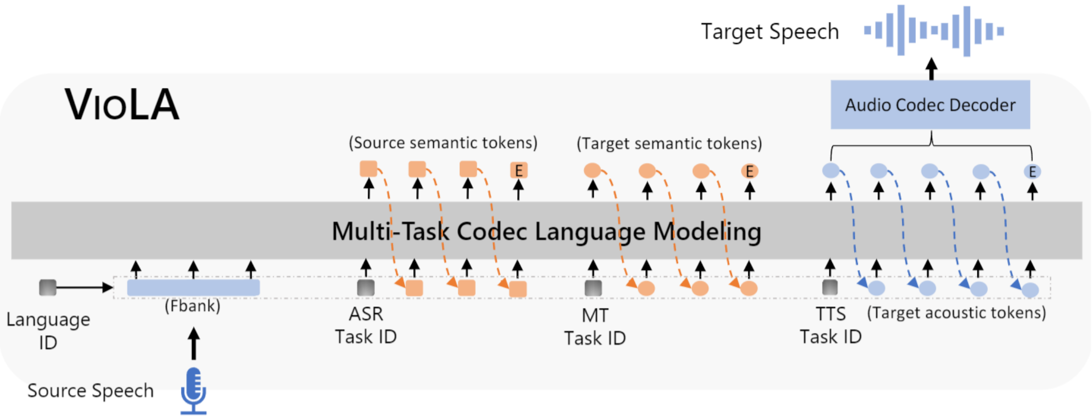
Slides credit: Xiaofei Wang

Wang, X., et al. Speechx: Neural codec language model as a versatile speech transformer. *arXiv:2308.06873,* 2023.

VioLA: A multi-modal model with discrete audio inputs

Speech and text can freely serve as input and output

- An extension to audio codec language model
- Naturally merge speech-language tasks
  - Speech recognition
  - Machine translation
  - Speech generation



Attempt to answer questions:

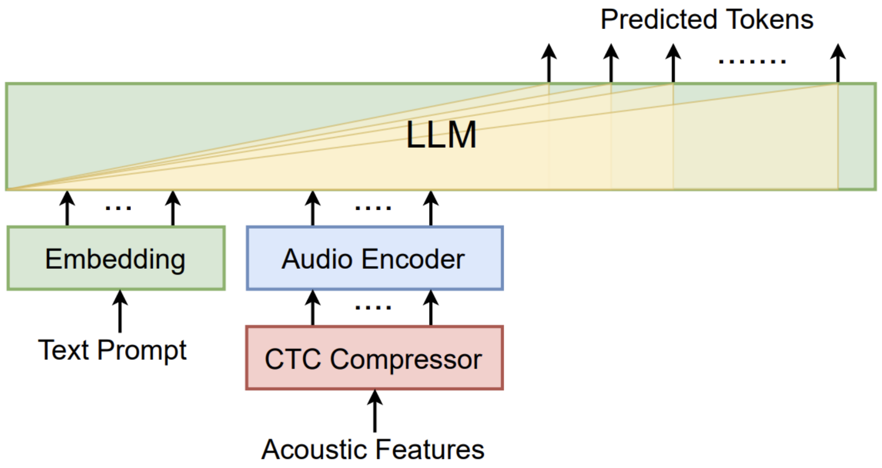
- Enc-dec or decoder?
- Codec or Fbank?
- Expert or General?

Input	Output	Typical Tasks
Speech	Text	ASR, ST
Text	Text	MT, LM
Text	Speech	multilingual TTS

Wang, T., et al., VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation. *arXiv:2305.16107, 2023*.

Speech-LLaMA: From decoder to LLM

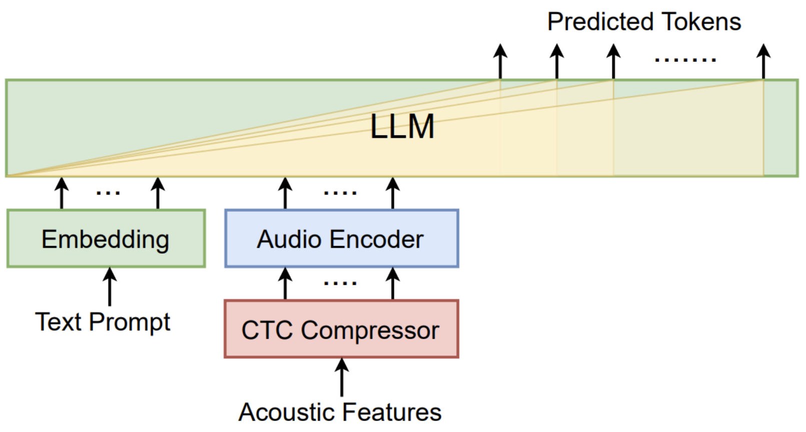
- Intuition: Decoder and LLM shares similar architecture, can we combine them?
  - ➢Leveraging the generalization capabilities of LLM
  - ➢Tight integration between the speech and LLM



Wu, J., et al. On decoder-only architecture for speech-to-text and large language model integration. *In Proc. ASRU*, 2023.

Speech-LLaMA: Overview

- Task
  - ➢X to EN speech translation
- LLM
  - ➢Pretrained and keep frozen
  - ➢LLaMA-7B
- CTC Compressor
  - ➢Reduce the acoustic feature length
  - ➢Pretrained on 14 language ASR/AST task
  - ➢Remove blank frames or average frames within same unit
- Audio Encoder
  - ➢Few transformer layers to further process the CTC compressor output
  - ➢Learn the shared representations within the space of the LLaMA embeddings



Speech-LLaMA: Training

- Instruct learning
  - ➢ Sample text prompt from a list for each training example, with or without locale information
  - ➢ Use fixed one during evaluation (“translate the audio into English”)
- LoRA finetuning
  - ➢ Only introduce small amount of training parameters for efficient LLM adaptation
- Training scheme
  - ➢ Two-stage: training the audio encoder and then with LoRA finetuning
  - ➢ One-stage: training audio encoder together with LoRA parameters
- Training and evaluation data
  - ➢ 14 languages (each 1K hours) with ST transcriptions
  - ➢ 13 languages on CoVOST 2 test set

Speech-LLaMA: Experiments

- Baselines
  - Seq2seq ST model with Whisper structure (240M, 12-layer encoder + 12-layer decoder)
  - Training with CE loss on the decoder and CTC loss on the encoder
  - Use LLaMA scores for the re-ranking

System	AR	DE	ZH	ES	FR	IT	NL	JA	RU	PT	ET	SV	SL	AVG
Enc-dec	22.8	22.6	7.0	23.7	21.8	20.7	34.6	15.3	26.4	28.9	9.4	24.4	13.3	20.8
+ LLaMA Rescore	24.9	23.6	7.2	24.9	22.7	21.6	36.0	15.7	27.7	30.2	9.4	25.6	12.7	**21.7**

Speech-LLaMA: Experiments

- Whether CTC compressor helps
  - ➢4-layer Transformer(15.8M) and pretrained on 14K ASR corpus
  - ➢For comparison, we use a convolution 1D on the top of the audio encoder for a subsampling rate of 32
  - ➢Using CTC compressor is better (18.2 to 22.5/24.0)
  - ➢Frame averaging is better than simply removing the blank frames

CTC-Comp.	AR	DE	ZH	ES	FR	IT	NL	JA	RU	PT	ET	SV	SL	AVG
w/o	16.9	16.9	3.4	19.6	15.4	16.7	28.3	10.3	22.8	22.8	15.4	26.3	22.2	**18.2**
Remove blank	24.6	22.6	9.6	23.5	20.9	21.4	32.4	17.5	31.0	26.8	17.0	25.3	20.3	**22.5**
Average frames	24.6	**24.3**	**10.1**	**25.4**	**22.6**	**23.7**	**34.1**	**17.7**	**33.3**	**29.2**	**17.2**	**26.7**	**22.8**	**24.0**

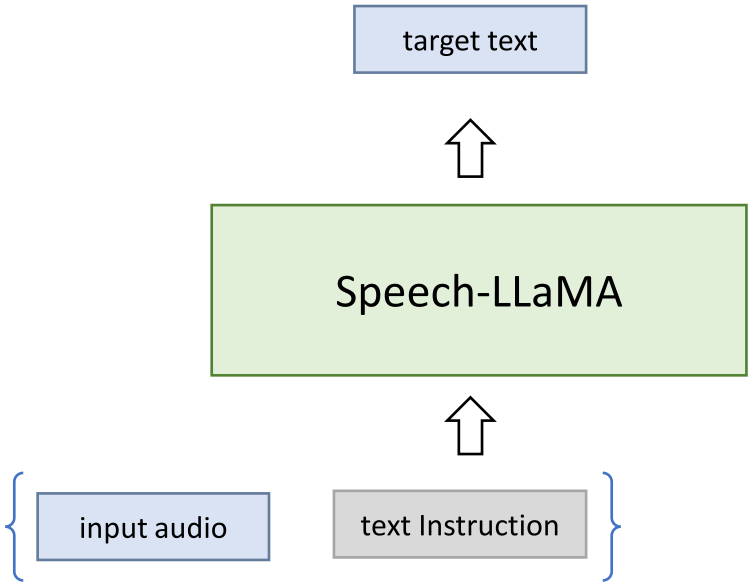
Speech-LLaMA: Experiments

- What kind of CTC compressor helps
  - We compared the CTC compressor on 14K training data with ASR and AST labels
  - Pretraining CTC compressor with ASR labels is significantly better than AST labels

System	AR	DE	ZH	ES	FR	IT	NL	JA	RU	PT	ET	SV	SL	AVG
CTC-C-AST	14.2	14.4	2.9	14.7	11.9	13.0	29.7	4.4	22.1	15.9	9.9	21.2	13.3	14.43
CTC-C-ASR	26.6	24.9	11.0	26.4	23.2	23.8	36.6	19.3	33.7	30.1	18.6	29.0	22.6	**25.1**

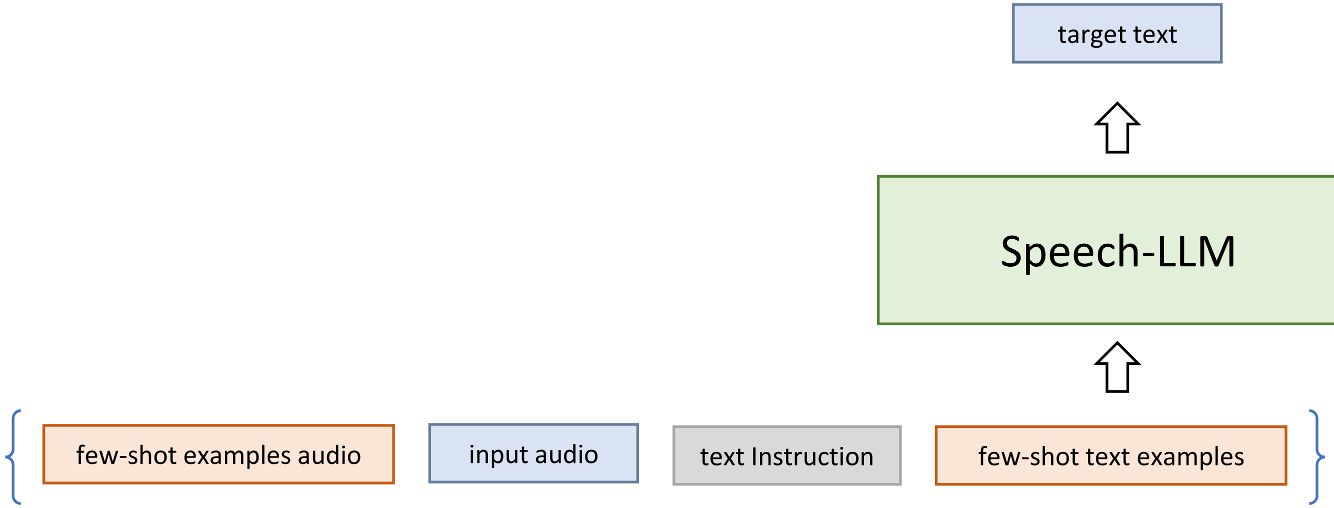
Speech-LLaMA: Instruction-following

- We train the Speech-LLaMA to do text generation according to instruction.
- Instruction could be task indicator, or the question targeting at the speech content.
- Training diagram:



Advancing Speech-LLM For In-context Learning

- Trained tasks (EN only)
  - ASR
  - Speech-based Question Answering
- Emergent Capable tasks
  - 0-shot and 1-shot En->X ST
  - 1-shot domain adaptation
  - Instruction-followed ASR



Pan, J., et al. COSMIC: Data Efficient Instruction-tuning For Speech In-Context Learning. *arXiv preprint*, 2023.

Instruction-following example: Contextual Biasing

- Speech Reference:
  - *…you know until we're tested we don't know what we're made of maybe that's what **adversity** gives … give ourselves a gift we can re imagine **adversity** as something…*
- Regular ASR instruction:
  - Instruction: *Transcribe the audio to text.*
  - Output: *…perhaps until tested we don't know where we're made of maybe that **diversity** gives … give ourselves the gift we can reimagine **diversity** as something…*
- ASR instruction with contextual biasing
  - Instruction: *Transcribe the audio to text. As context, the speaker talks about what **hardship** can bring us.*
  - Output: *…perhaps until we're tested we don't know where we're made of maybe that **adversity** gives … give ourselves the gift we can reimagine **adversity** as something…*

In-context learning example

- With the cross-lingual capabilities of LLM and in-context learning capability of Speech-LLaMA, it is able to achieve EN->X translation despite only being trained with EN data.
- We randomly pick 1 utterance from train set and provide it as 1-shot example along with its corresponding translation.*

Table 2. In-domain EN→X S2TT on TED-LIUM 3 test sets

Model	#Example	EN→X Target ES	FR	DE
Cascaded-7B	0-shot	26.07	22.61	15.53
COSMIC-ASR-7B	0-shot	2.53	2.32	2.77
	1-shot	2.04	2.32	4.78
COSMIC-7B	0-shot	17.13	20.88	15.45
	1-shot	28.89	26.45	19.18
COSMIC-13B	0-shot	8.59	13.24	10.12
	1-shot	30.57	28.41	21.36

** The model is trained and tested on Tedlium3 dataset. The ES/FR/DE translation reference during testing is generated by GPT-3.5 based on the English transcripts.*

In-context learning example

Table 3. Cross-domain EN→X S2TT on FLEURS test sets

Model	#Example	EN→X Target ES	FR	DE	ZH
Cascaded-7B	0-shot	6.55	9.07	4.68	11.11
COSMIC-7B	0-shot	6.47	13.45	7.45	2.71
	1-shot	12.39	16.92	11.23	12.41
COSMIC-13B	0-shot	3.62	7.82	4.98	5.88
	1-shot	12.56	19.41	12.00	15.78