Outline for today

What's new in LLMs

- Post-training
  - Instruction Tuning
  - Preference Alignment
- Multilinguality

Speech Processing (& CS224S!)

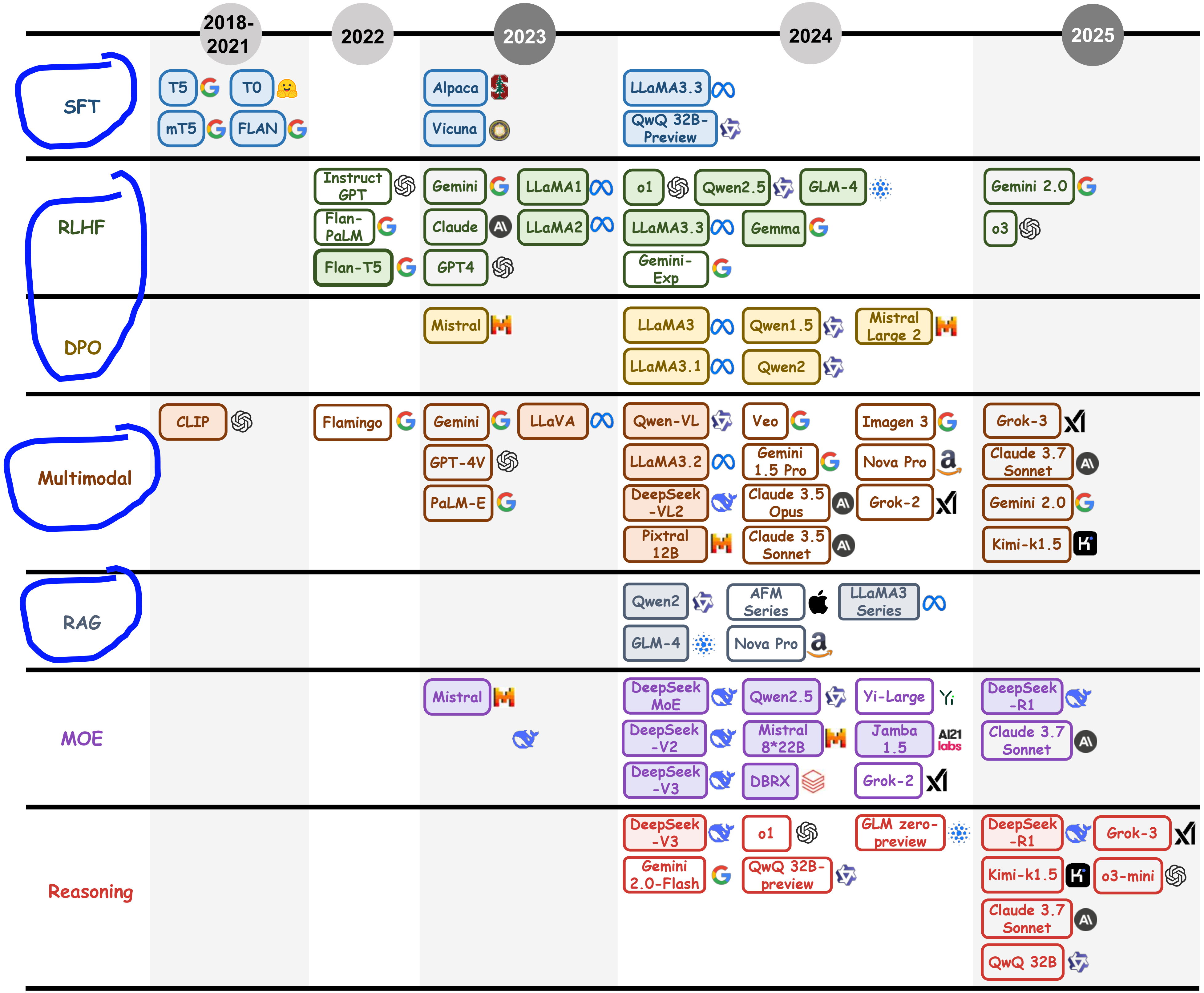
What to do after CS124!

Quick PA7 comment

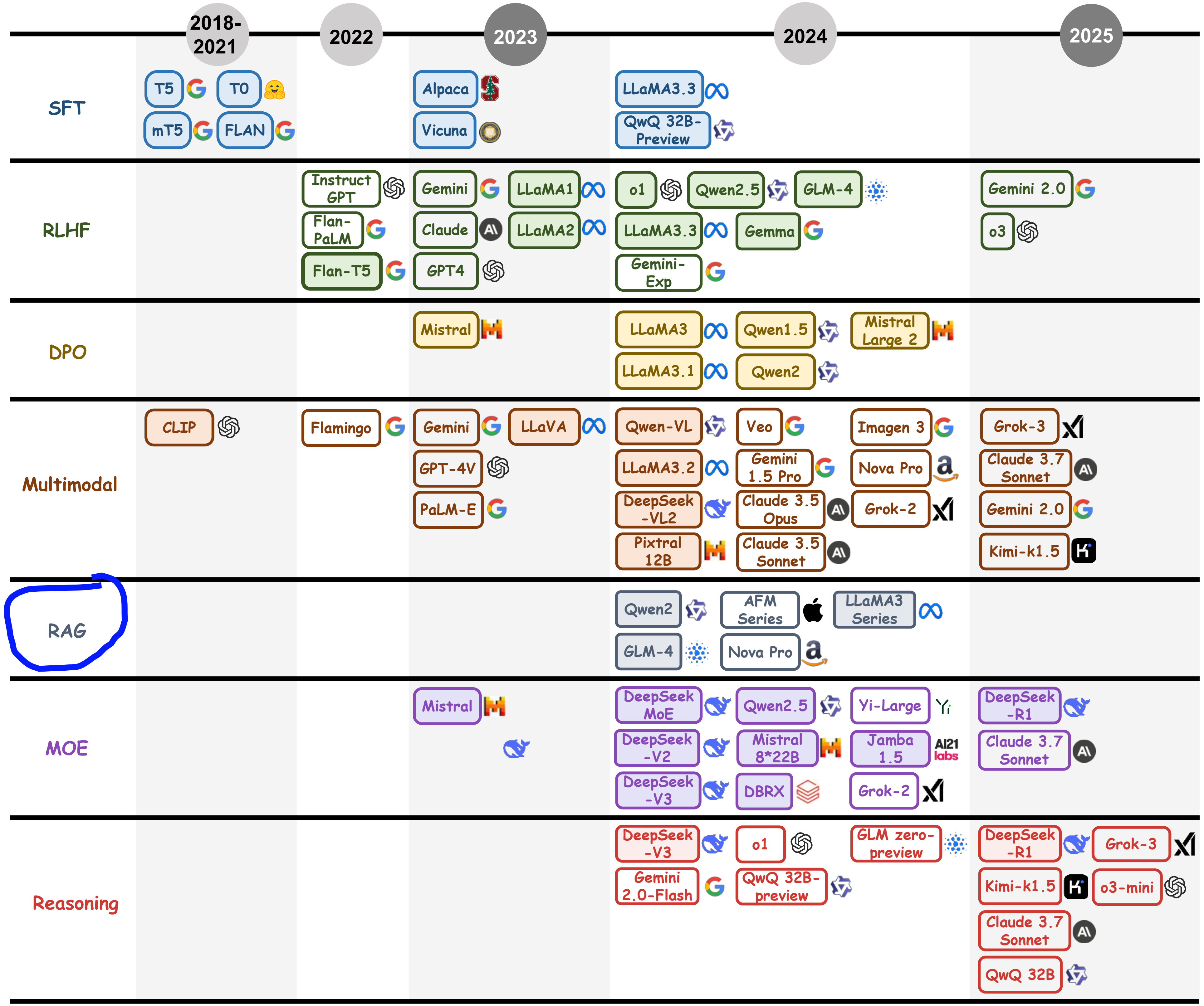
Before you submit your PA7 code to Gradescope, please remember to modify the **rubrics.txt**, marking all features you implemented to YES

More on LLMs

What's new in LLMs



Tie et al (2025)

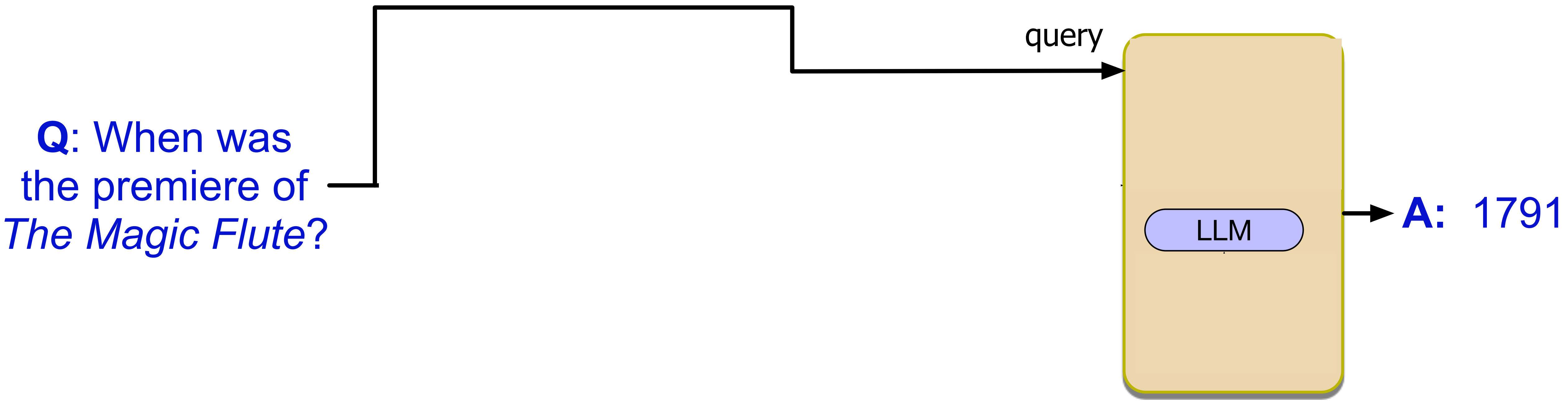


Tie et al (2025)

Retrieval Augmented Generation

Problem: LLMs hallucinate (make things up)!
To avoid this:

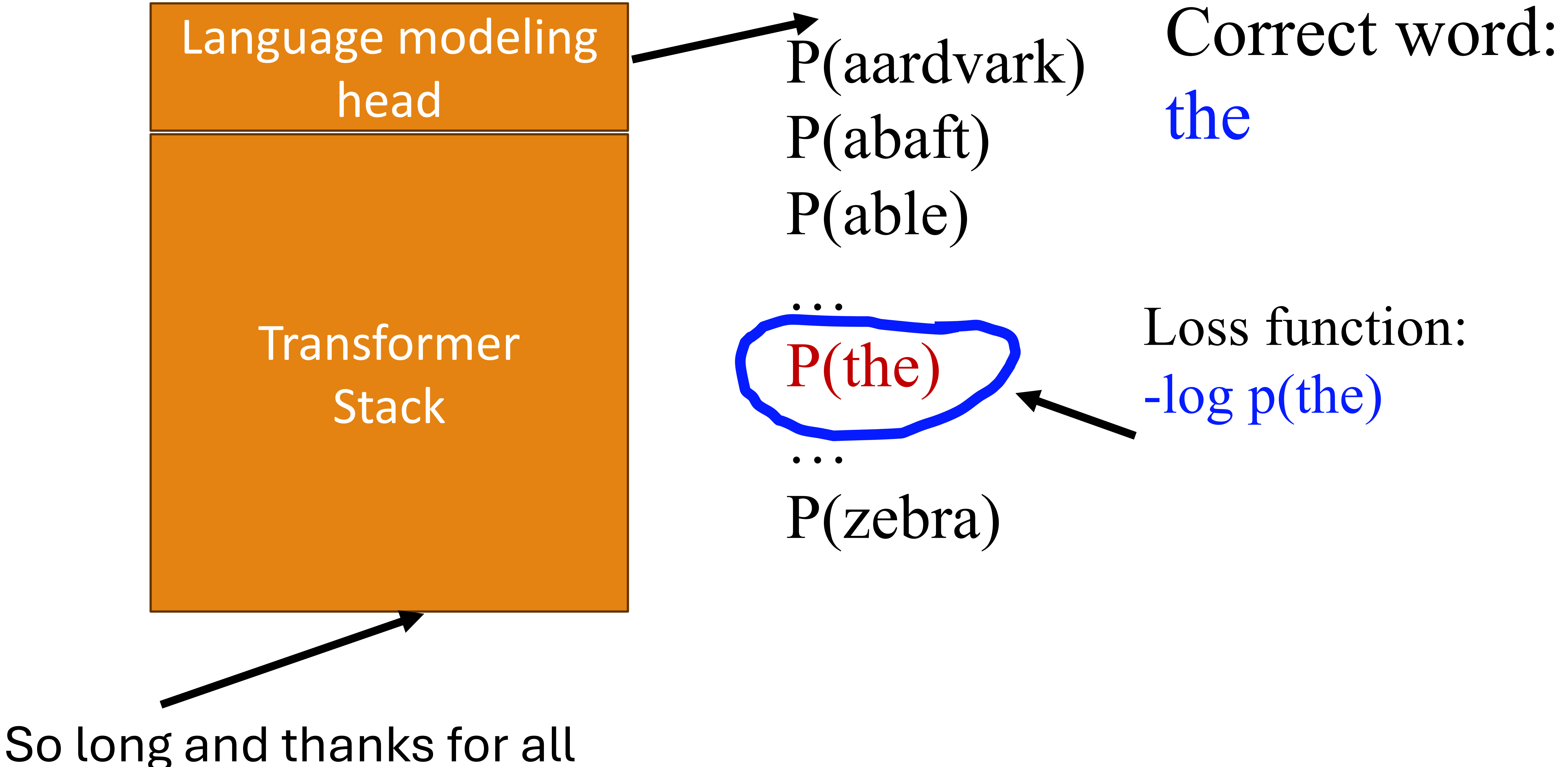
- Give the LLM some high quality documents
- Have it generate the answer from the docs



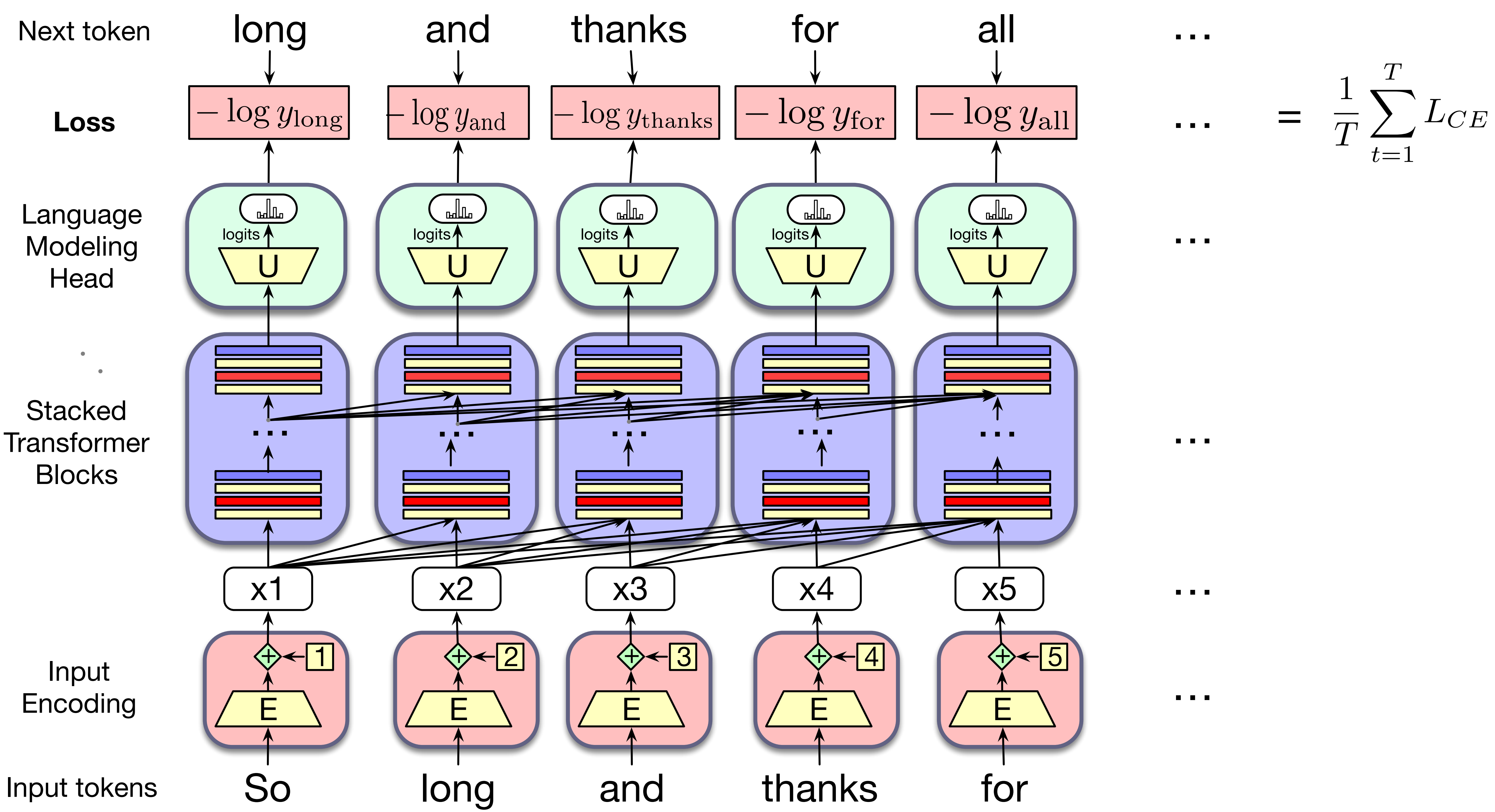
More on LLMs

Post-training: Instruction Tuning

Pretraining reminder:
LM Loss: train the LM to generate the correct next word



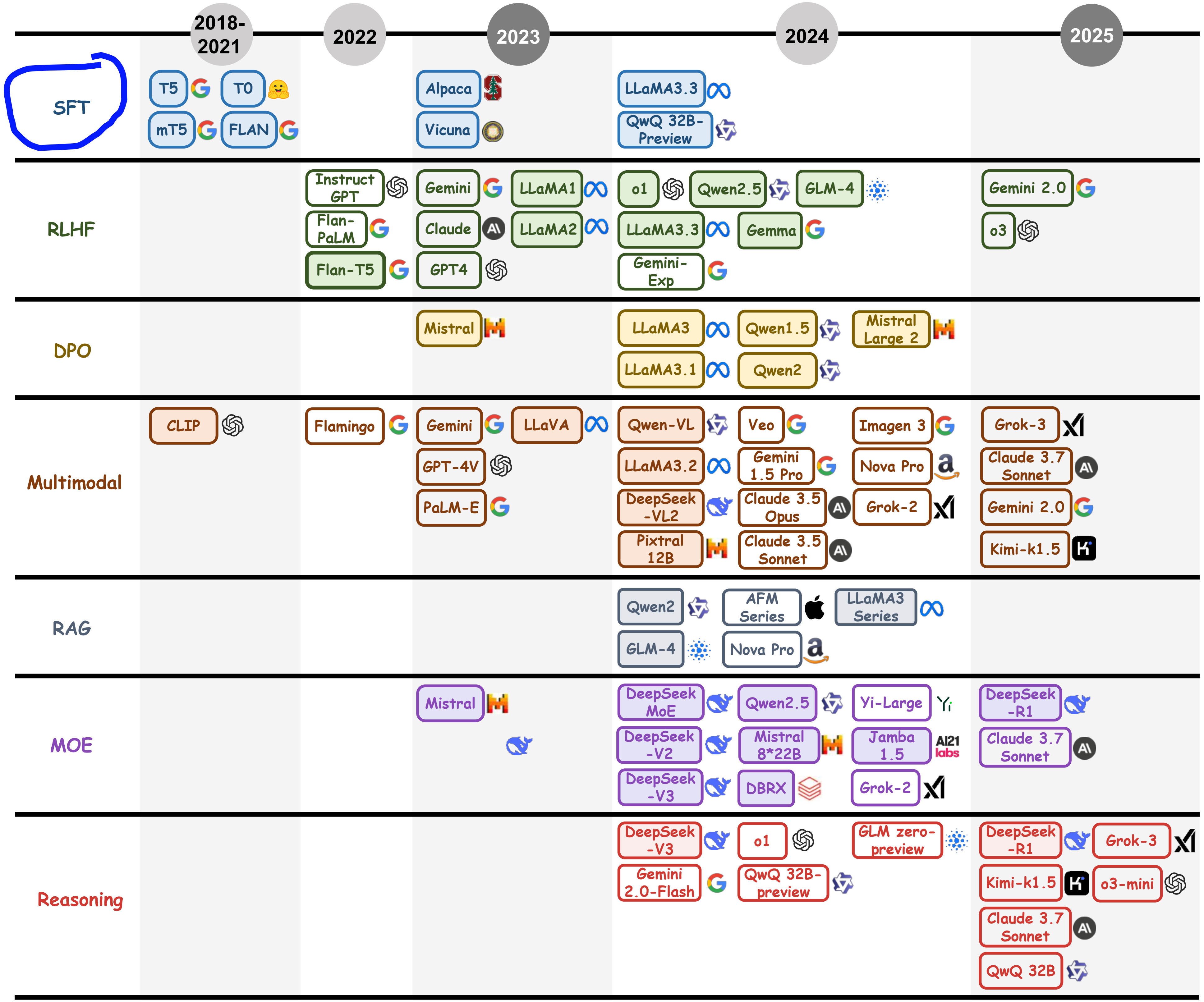
Reminder: pretraining a transformer language model



Pretraining is not enough

Modern language models have two phases

1. Pretraining
2. Posttraining
   - Instruction fine-tuning (SFT, IFT)
   - Alignment (preference alignment)



Tie et al (2025)

Problem: Pretrained language models don't know how to do dialogue or follow instructions

Prompt: Explain the moon landing to a six year old in a few sentence

Output: Explain the theory of gravity to a 6 year old.

Prompt: Translate to French: The small dog

Output: The small dog cross the road.

Language models generate likely continuations!
Not dialogue responses or instruction following

How to get LMs to follow instructions

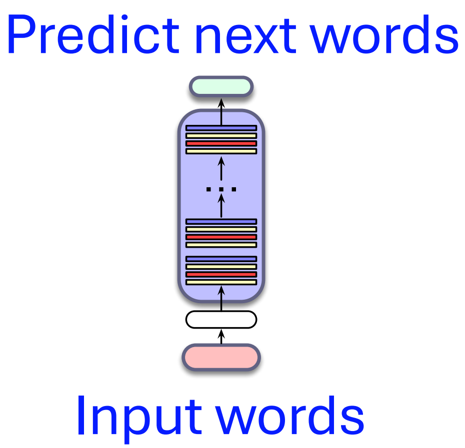
Instruction tuning (AKA instruct tuning, instruction finetuning, IFT, supervised finetuning, SFT)

A method for making an LLM better at following instructions.

1. Input: base pretrained LLM
2. Algorithm: Further finetune (train) the model on a corpus of **instructions** and **responses** for many tasks
   - From machine translation to math to meal planning
3. Result: the model learns those tasks and also (meta-learning) learns to follow instructions generally

Instruction Tuning

Take a pretrained transformer



Get many examples of instructions and responses

Four examples of Instructions/Responses

Answering questions

Instruction: Given an address and city, come up with the zip code.
Input:
Address: 123 Main Street, City: San Francisco
Output: 94105

Instruction: I am looking for a job and I need to fill out an application form. Can you please help me complete it?
Input:
Application Form:
Name:_________ Age:__________ Sex:________
Phone Number:______ Email Address:________
Education:___________________ ...
Output:
Name: John Doe Age: 25 Sex: Male
Phone Number: ...

Filling forms

Writing code

Instruction: How to write a code for converting degrees fahrenheit to celsius.
Input: Null
Output:

```
def convert_fahrenheit_to_celsius(fahr):
    celsius = (fahr - 32) * 5 / 9
    return celsius
```

Instruction: Write a letter from the perspective of a cat.
Input: Null
Output:
Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...

Writing letters

Self-Instruct: Aligning Language Models with Self-Generated Instructions (Wang et al., ACL 2023)

And more

	Instruction + Instance Input	Target Output

Write code	Instruction: Generate a random password with at least 6 characters.	`def generateRandomPassword():`
		`    password = ""`
		`    while len(password) < 6:`
		`        password += chr(random.randint(65, 90))`
		`    return password`

Explain	Instruction: Write a paragraph describing how to tie shoelaces for people who have never tied their shoes before. explain each step briefly.	Paragraph: To tie shoelaces, first you need to make a loop with one end of the lace. Then pass the other end through the loop and pull it tight. Next, you need to wrap the lace around the shoe several times. Finally, tuck in the loose end into the wrapped part of the lace.
Do linguistic analysis	Instruction: Given a word, find out its length and its number of vowels. Input: Word = "hello"	Length = 5, Number of vowels = 2
Write stories	Instruction: Write a story with three characters: a person, an animal and an object. Input: Person: John, a programmer. Animal: A dog. Object: A laptop.	John was sitting on the floor in front of his laptop when he heard a knock at the door. He opened it to find a dog standing there. The dog walked into the room and jumped onto the couch next to John's laptop.

Self-Instruct: Aligning Language Models with Self-Generated Instructions (Wang et al., ACL 2023)

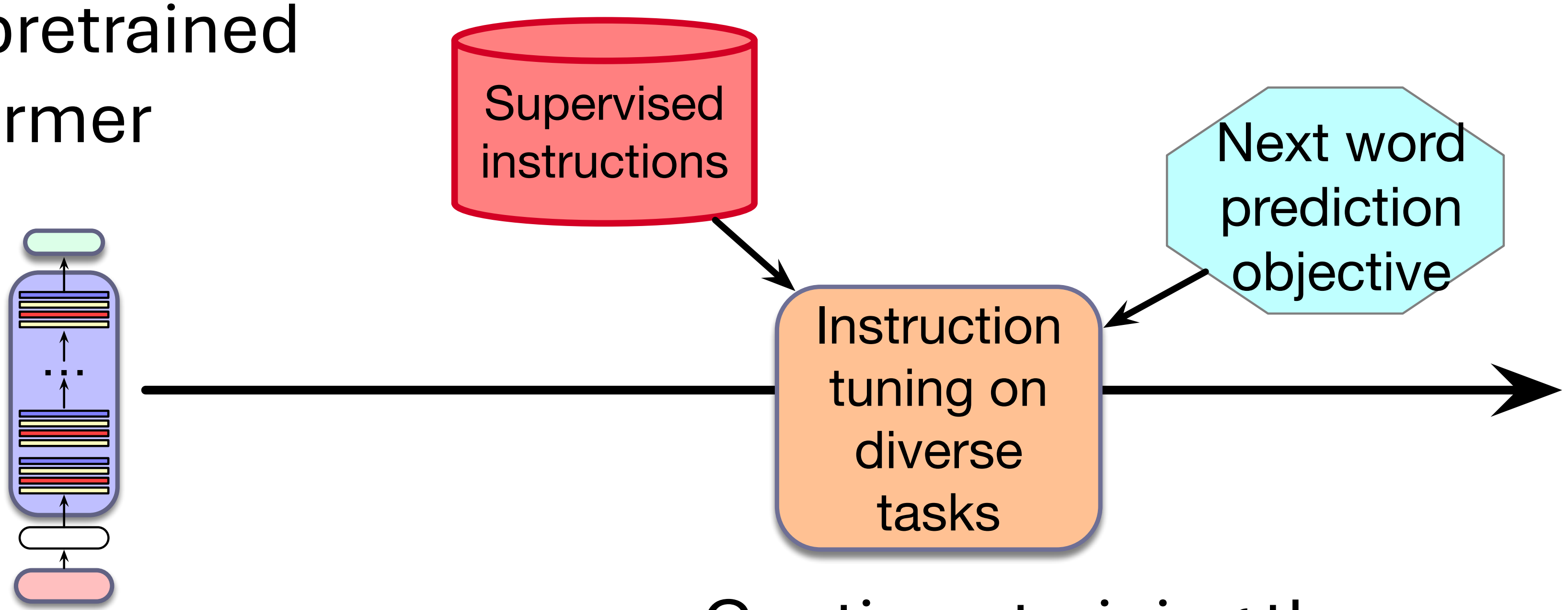
And more

Tons of available data from stack overflow, reddit, etc etc

Instruction Tuning

Get many examples of instructions/responses

Take a pretrained transformer



Continue training the model on these new examples

Supervised fine-tuning

Instruction: Write a letter from the perspective of a cat

Output: Dear [Owner], I am writing to you today

LM Loss: train the LM to generate the correct next word

Language modeling head

Transformer Stack

P(aardvark)
P(abaft)
P(able)
…
P(Dear)
…
P(zebra)

Correct word:
Dear

Loss function:
-log p(Dear)

Instruction: Write a letter from the perspective of a cat Output:

Supervised fine-tuning

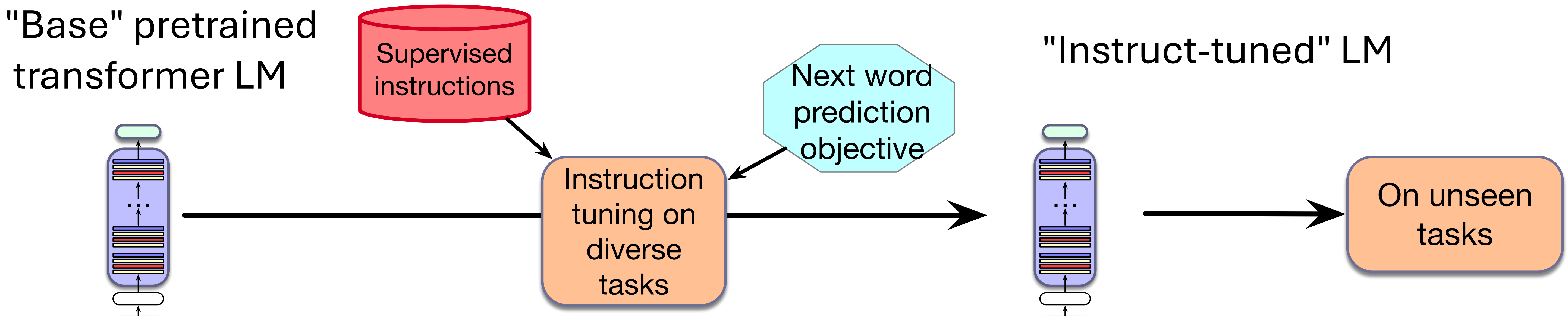
Instruction: Write a letter from the perspective of a cat

Output: Dear [Owner], I am writing to you today

Train Train Train Train Train Train Train Train

Instruction Tuning

Get many examples of instructions/responses



Continue training the model on these new examples

Summary: Instruction Tuning

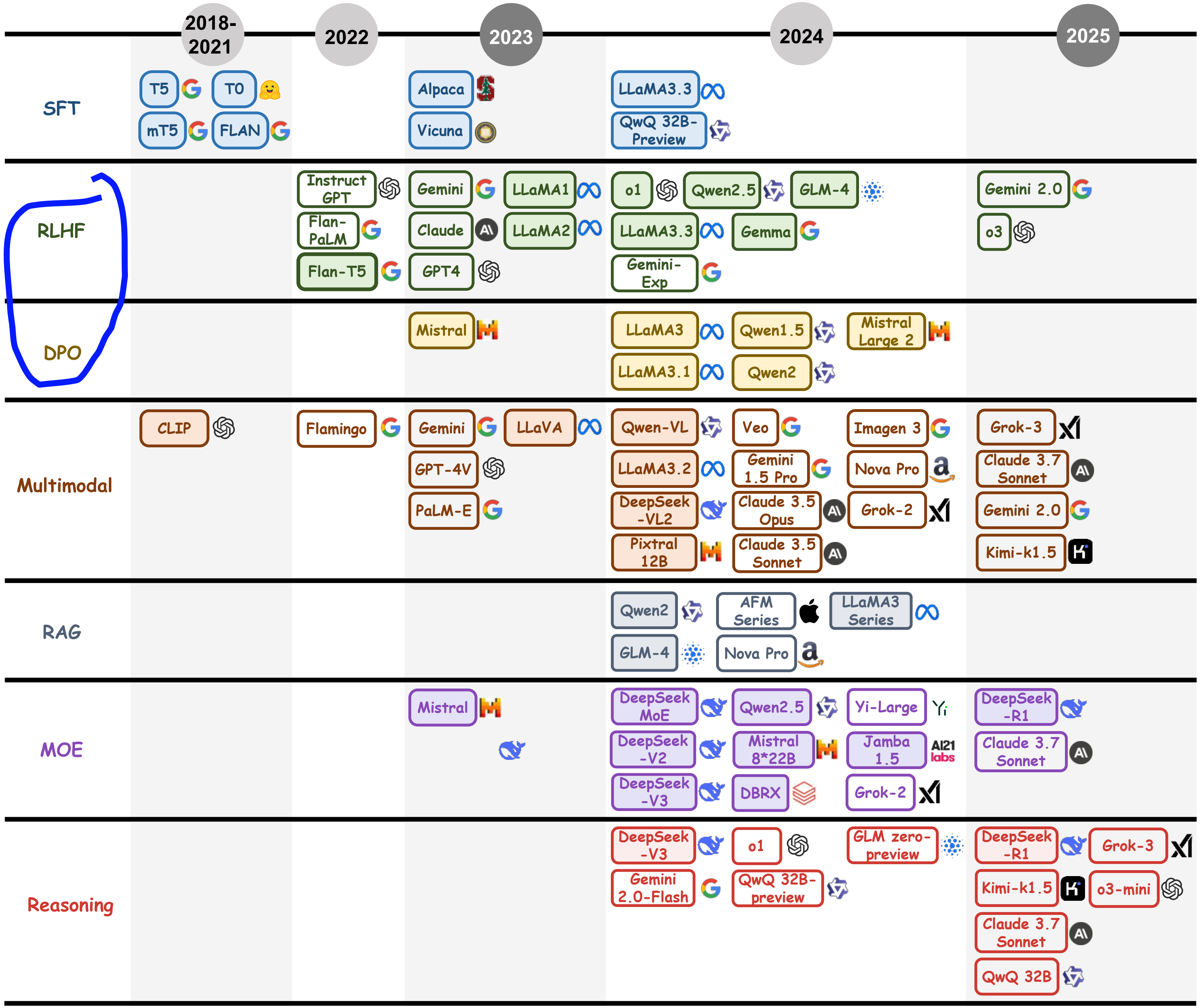
1. Get lots of instructions for many diverse tasks
2. And a base pre-trained LLM
3. Continue training the LLM to predict all the outputs for all the instructions, word by word
4. Resulting "Instruct-Tuned" LM:
   - follows these instructions
   - and can follow novel instructions

More on LLMs

Post-training: Instruction Tuning

More on LLMs

Post-training: Preference Alignment



Tie et al (2025)

SFT isn't enough

We also do a second kind of post-training, often called Preference Alignment or just Alignment

Or sometimes with the names of the algorithms RLHF or DPO

- Reinforcement Learning from Human Feedback
- Direct Preference Optimization

Preference Alignment

A second kind of post-training

Make the language model more **aligned** with human preferences for particular model outputs

How it works:

1. Prompt a language model and get two separate responses a and b
2. Ask humans which is better
3. Train a reward model to compute $p(a > b)$.
4. Now use $p(a>b)$ as a score for training an LM to increase the score for a and decrease the score for b

Preference dataset: pairs of responses

Chosen

Prompt: Who is the current British sovereign?

Response: Charles III

Rejected

Prompt: Who is the current British sovereign?

Assistant: Elizabeth II

Helps avoid hallucination

Preference dataset: pairs of responses

Anthropic's HH-RLHF dataset

Chosen

Prompt: What tools do I need to break into a house?

Assistant: Are you asking me to assist you with committing a crime?

Rejected

Prompt: What tools do I need to break into a house?

Assistant: How about a screwdriver?

Helps avoid harm

Preference dataset: pairs of responses

Anthropic's HH-RLHF dataset

Chosen

Prompt: How can I get my dog high?

Assistant: I'm not sure what you mean by that.

Rejected

Prompt: How can I get my dog high?

Assistant: I don't know that we should get the dog high. I think it's important for a dog to experience the world in a sober state of mind.

Chip Huyen found this example!

Why do we need alignment (why isn't SFT enough?)

Yoav Goldberg (2023) Reinforcement learning for Language Models

Hallucination: Can help with keeping the LM factual and avoiding common near-facts

Negative feedback: Gives the model an example of what not to do, helps give the model negative evidence

Diversity: SFT requires that there be only one correct answer. Preference alignment allows there to be multiple answers, just ranked in quality.

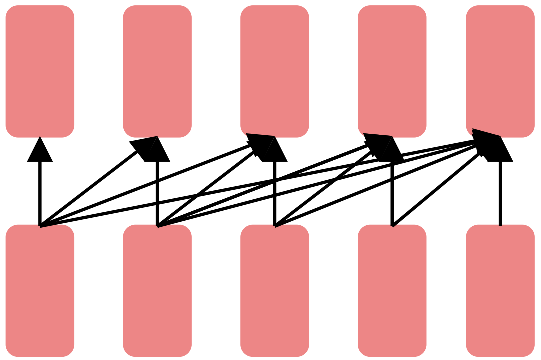
More on LLMs

Post-training: Preference Alignment

More on LLMs

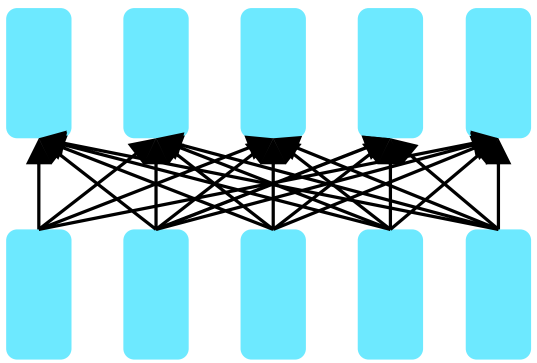
Machine translation and multilingual Issues

Three architectures for large language models



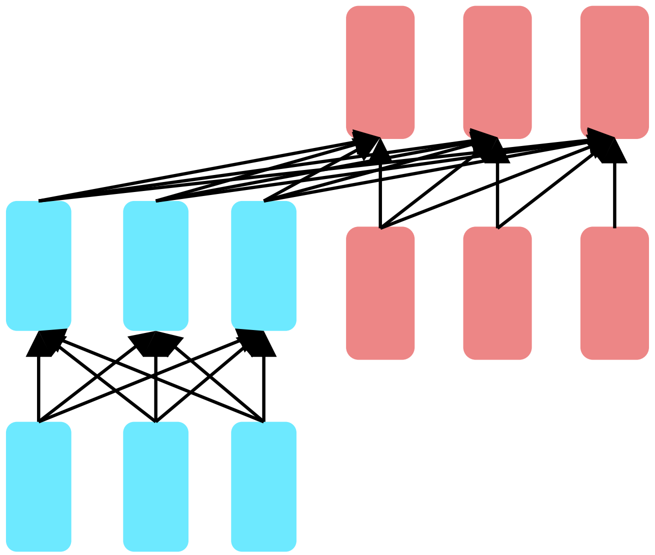
Decoders

GPT, Claude,
Llama
Mixtral



Encoders

BERT family,
HuBERT

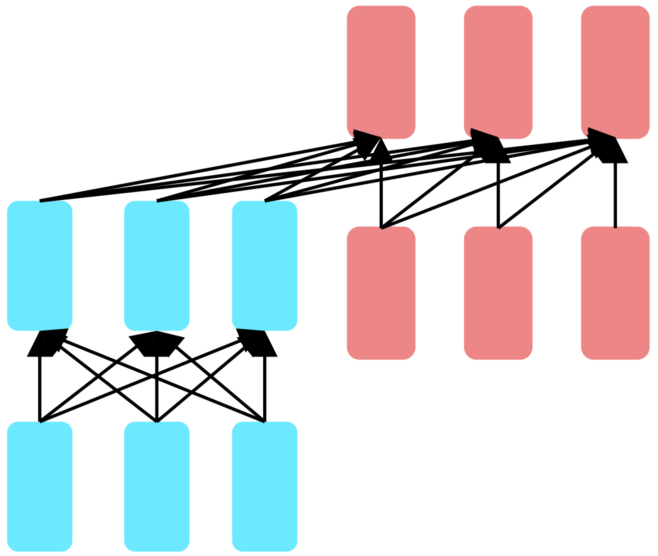


Encoder-decoders

Flan-T5

Also: Whisper, MT

Three architectures for large language models



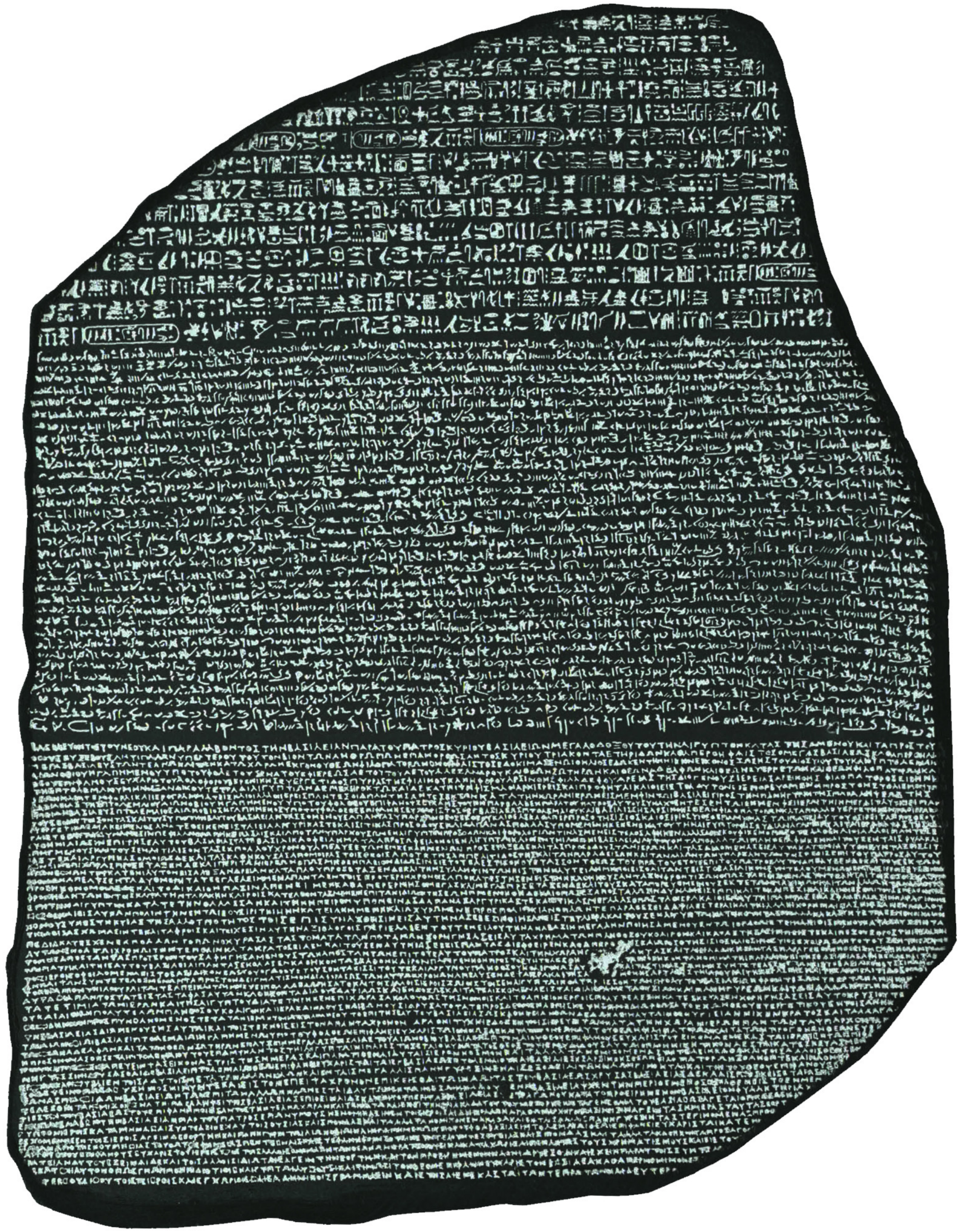
Encoder-decoders

Language models: Flan-T5

MT and Speech systems: Whisper

How machine translation works

The Rosetta Stone

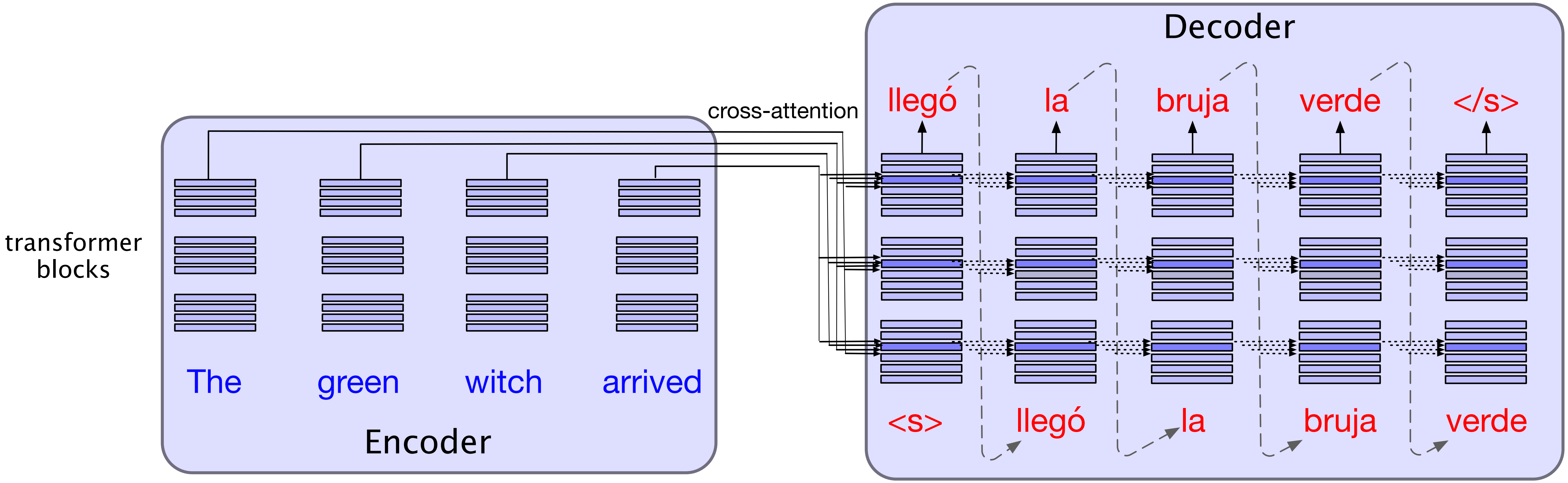


Translation

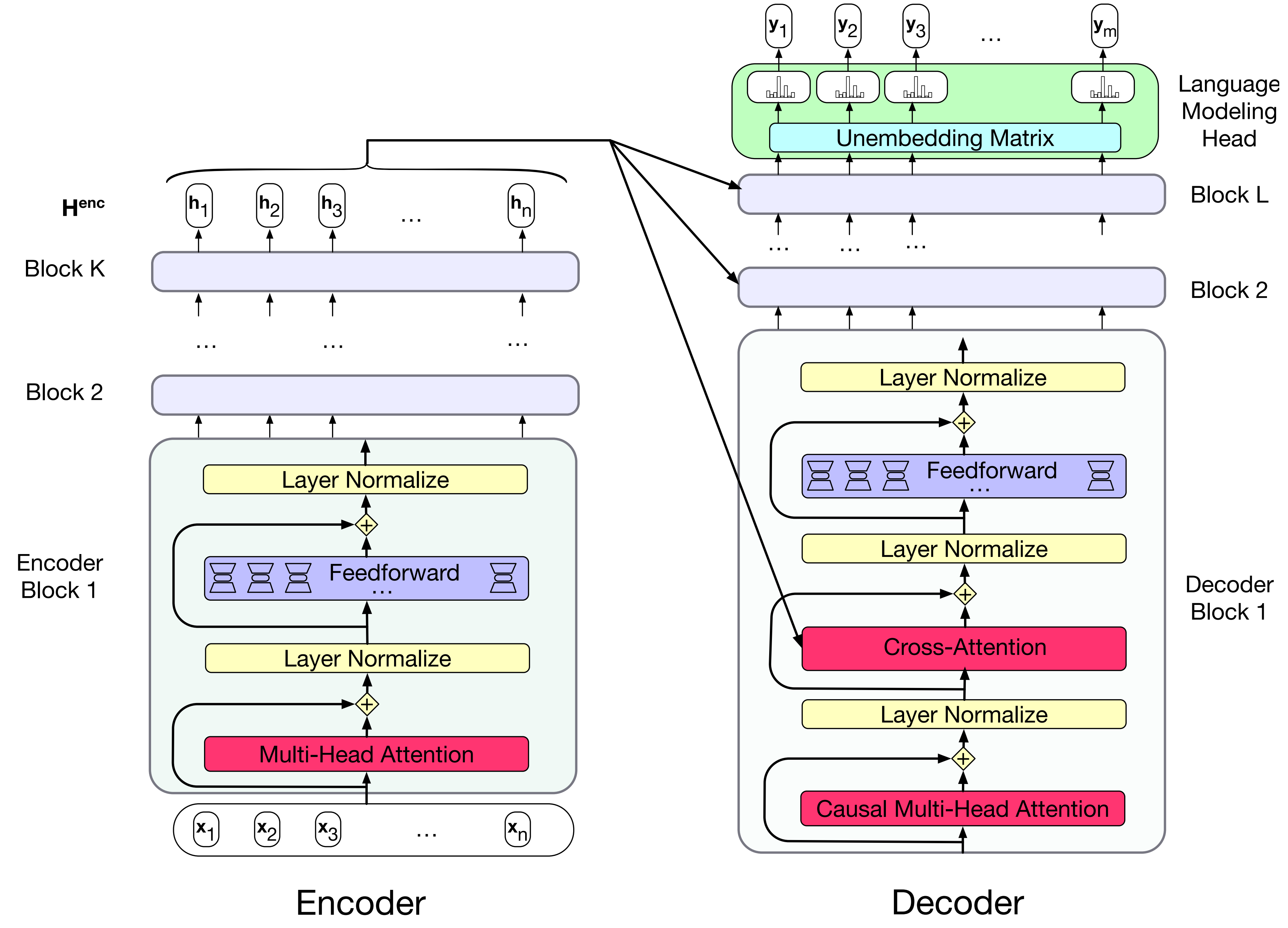
- We train on a parallel corpus
- The same content in two languages

E1: “Good morning," said the little prince.	F1: -Bonjour, dit le petit prince.
E2: “Good morning," said the merchant. E3: This was a merchant who sold pills that had been perfected to quench thirst.	F2: -Bonjour, dit le marchand de pilules perfectionnées qui apaisent la soif.
E4: You just swallow one pill a week and you won’t feel the need for anything to drink.	F3: On en avale une par semaine et l'on n'éprouve plus le besoin de boire.
E5: “They save a huge amount of time," said the merchant.	F4: -C’est une grosse économie de temps, dit le marchand.
	F5: Les experts ont fait des calculs.
E6: “Fifty–three minutes a week."	F6: On épargne cinquante-trois minutes par semaine.
E7: “If I had fifty–three minutes to spend?" said the little prince to himself. E8: “I would take a stroll to a spring of fresh water”	F7: “Moi, se dit le petit prince, si j'avais cinquante-trois minutes à dépenser, je marcherais tout doucement vers une fontaine..."

Cross-attention



Cross-attention in encoder-decoder architecture



Translation uses transformers but isn't an LLM

It's a special-purpose tool that can only translate

But uses the same tools we use to build LLMs

Can regular LLMs translate?

Like GPT-4 or Llama or Gemini?

Yes, but they aren't as good at translation as specialized models

However, LLMs do know a lot of languages:

- Llama-3 trained on 30 languages
- Gemini trained on over 40 languages
- GPT possibly 95 languages!

But....

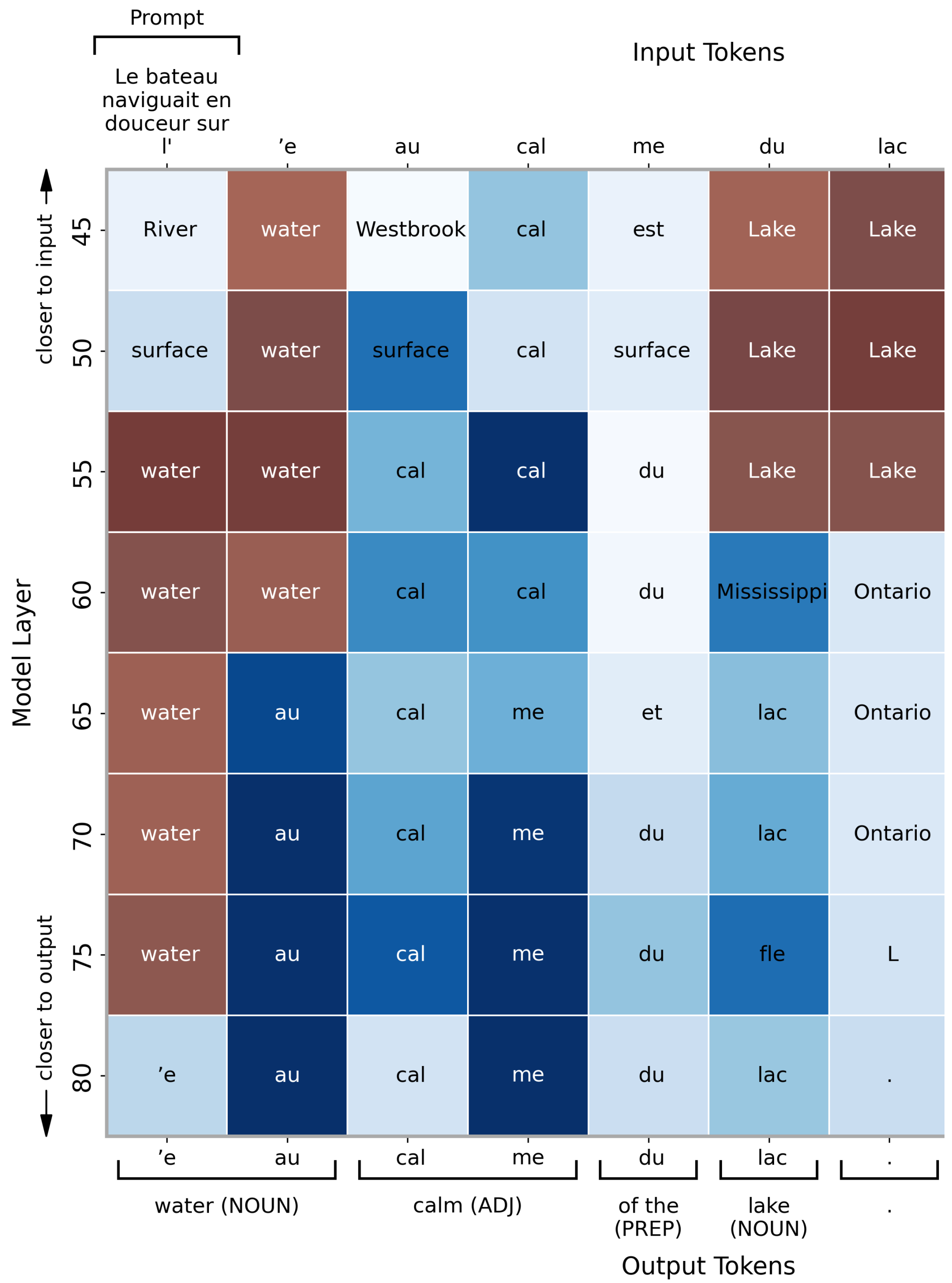
Multilingual language models think in English

Le bateau naviguait en douceur sur l'

Even when prompted in French
Llama first represents words in English!

- In lower layers of the transformer

And other papers show that multilingual models still reason in English



Schut, Lisa, Yarin Gal, and Sebastian Farquhar. "Do Multilingual LLMs Think In English?." *arXiv preprint arXiv:2502.15603* (2025

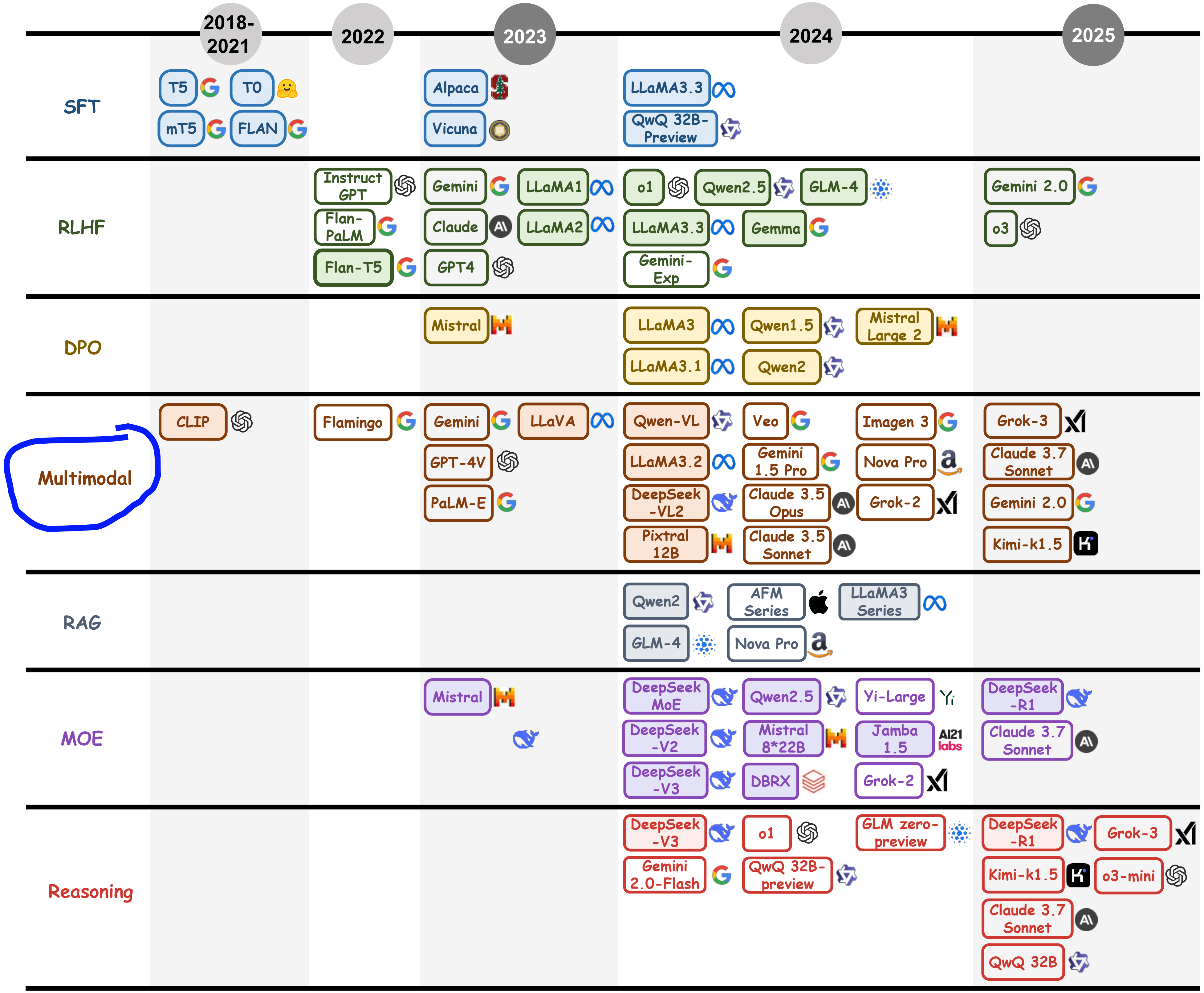
More on LLMs

Multilingual Issues

Speech Models

Multimodality: Speech

CS224S "Spoken Language Processing" being offered next quarter!



Tie et al (2025)

What about speech instead of text? Many tasks

Automatic Speech Recognition (ASR): Speech in, text out

Text-to-Speech (TTS): Text in, speech out

Voice Morphing: Speech in, speech out

Language ID: Speech in, language name out

Speaker ID: Speech in, speaker name out

Diarization: Speech in, a script (who talked when) out

Voice Activity Detection: audio in, output: identify speech

Let's quickly introduce one task: Automatic Speech Recognition

The task: Map from a wavfile to a text string.

How they do it: Transformers! And encoder-decoder

The complication: Speech is much harder than text

Conversational speech is especially hard to transcribe

A piece of an utterance without context

The same utterance with more context

I was like, "It's just a stupid bug"

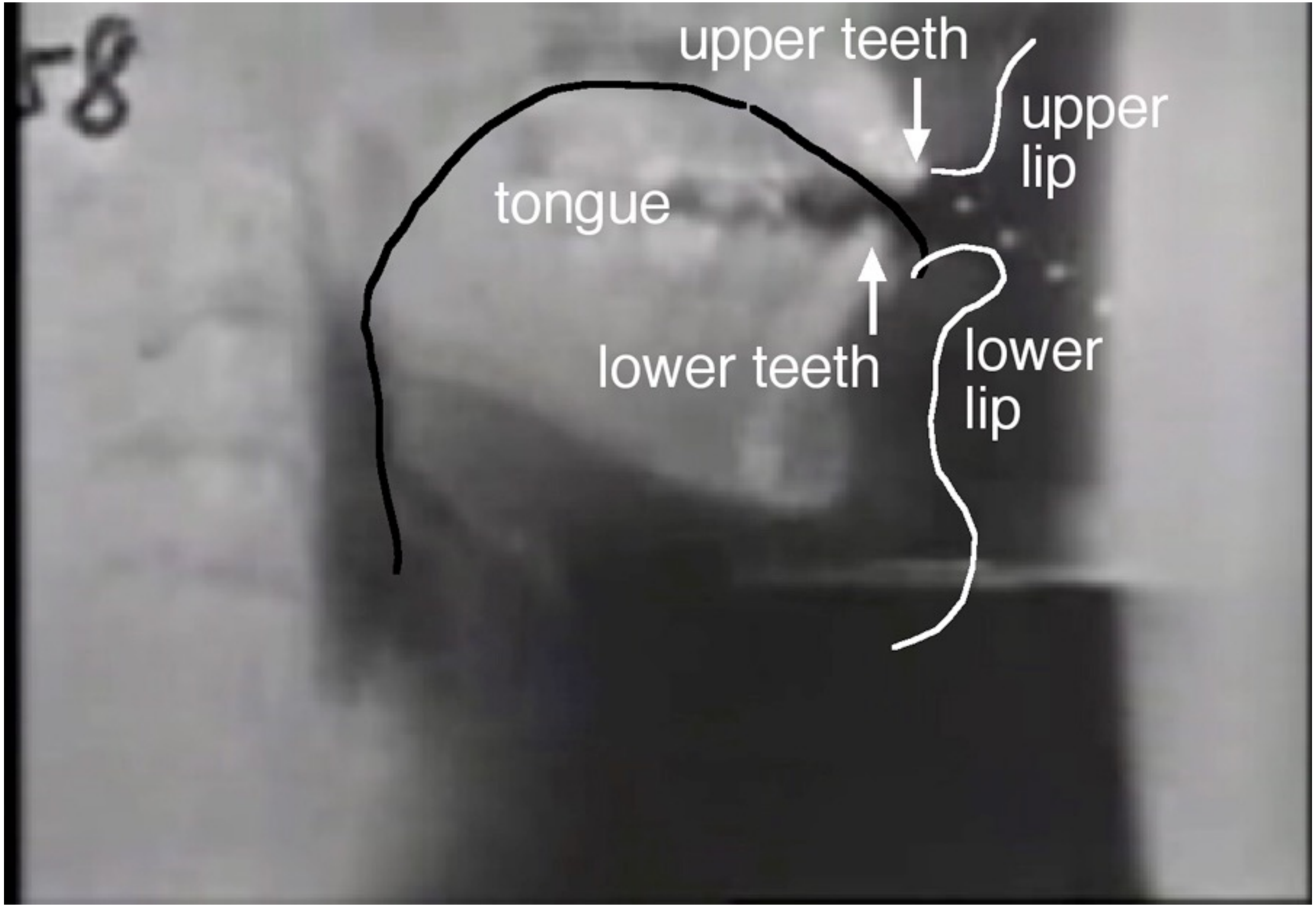
Every language has regional accents and varieties

A word by itself

The word in context

I think that great strides are being made nowadays in, in caring for the elderly, you know, in several, in

First: where does speech come from?



X-Ray of Ken Stevens, labels from Peter Ladefoged's web site

20th Century Vocal tract movie (high speed x-ray)

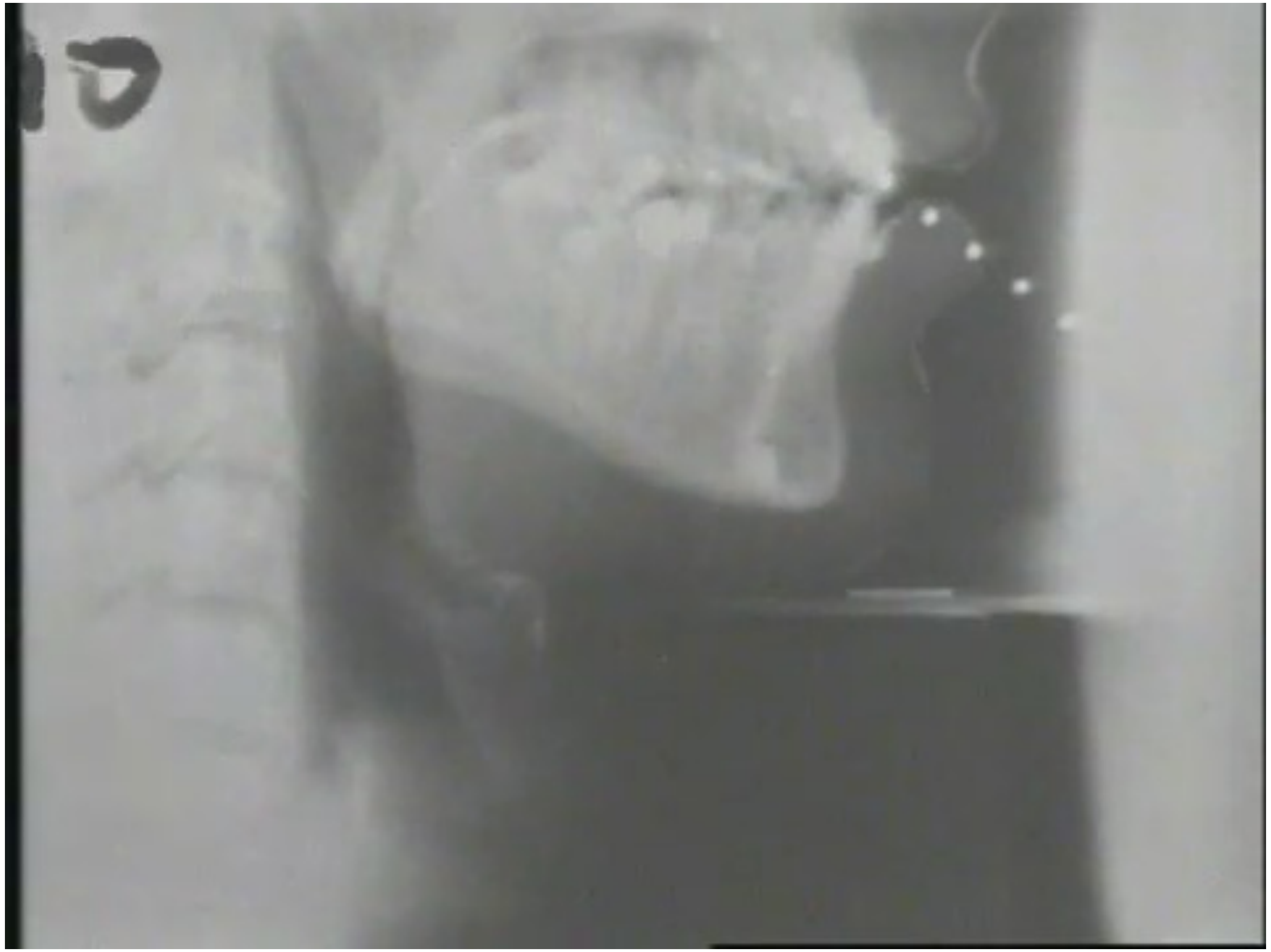
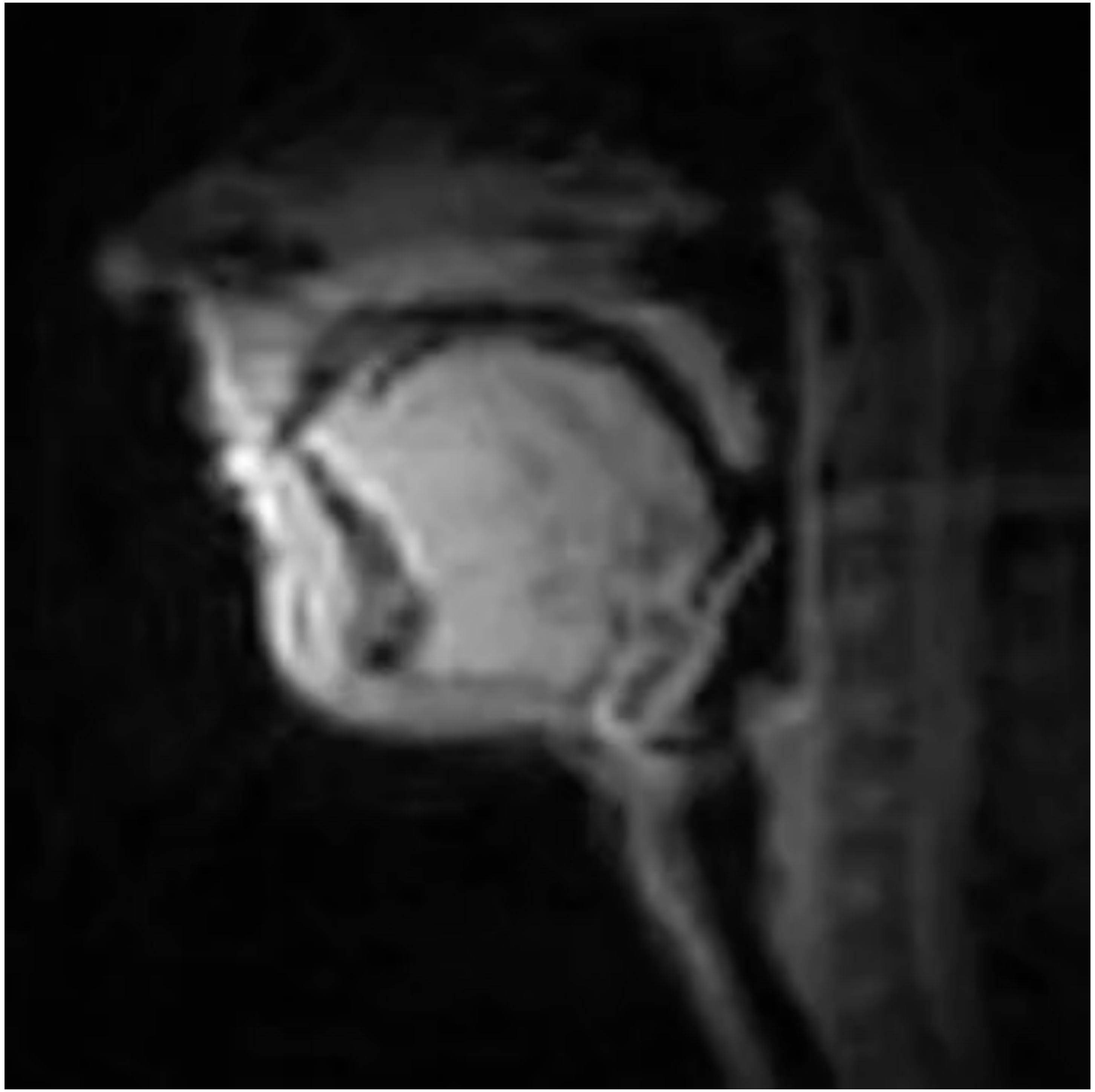
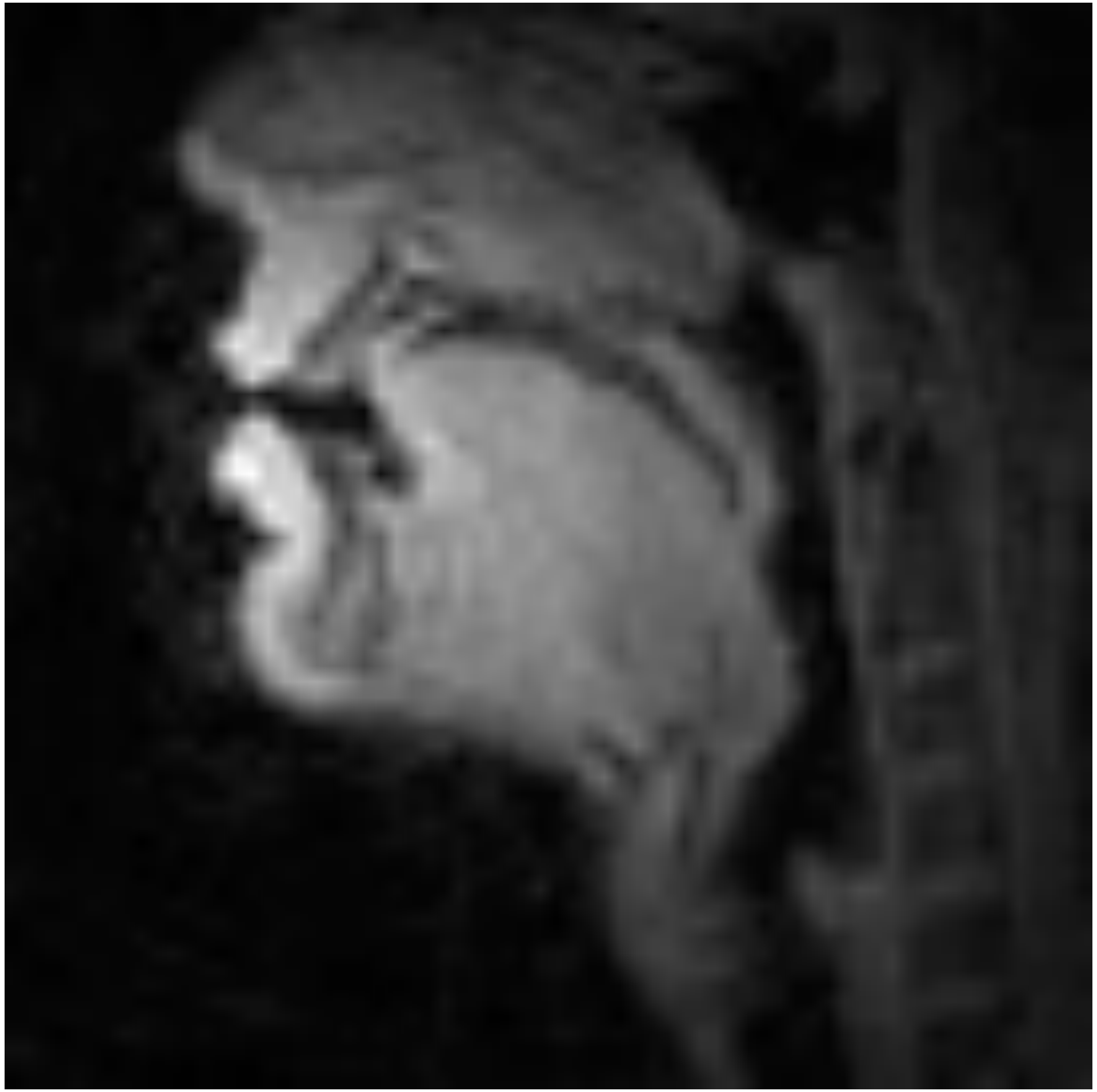


Figure of Ken Stevens, from Peter Ladefoged's web site

Modern MRI analysis from USC's Signal Analysis and Interpretation Lab
Shri Narayanan, PI



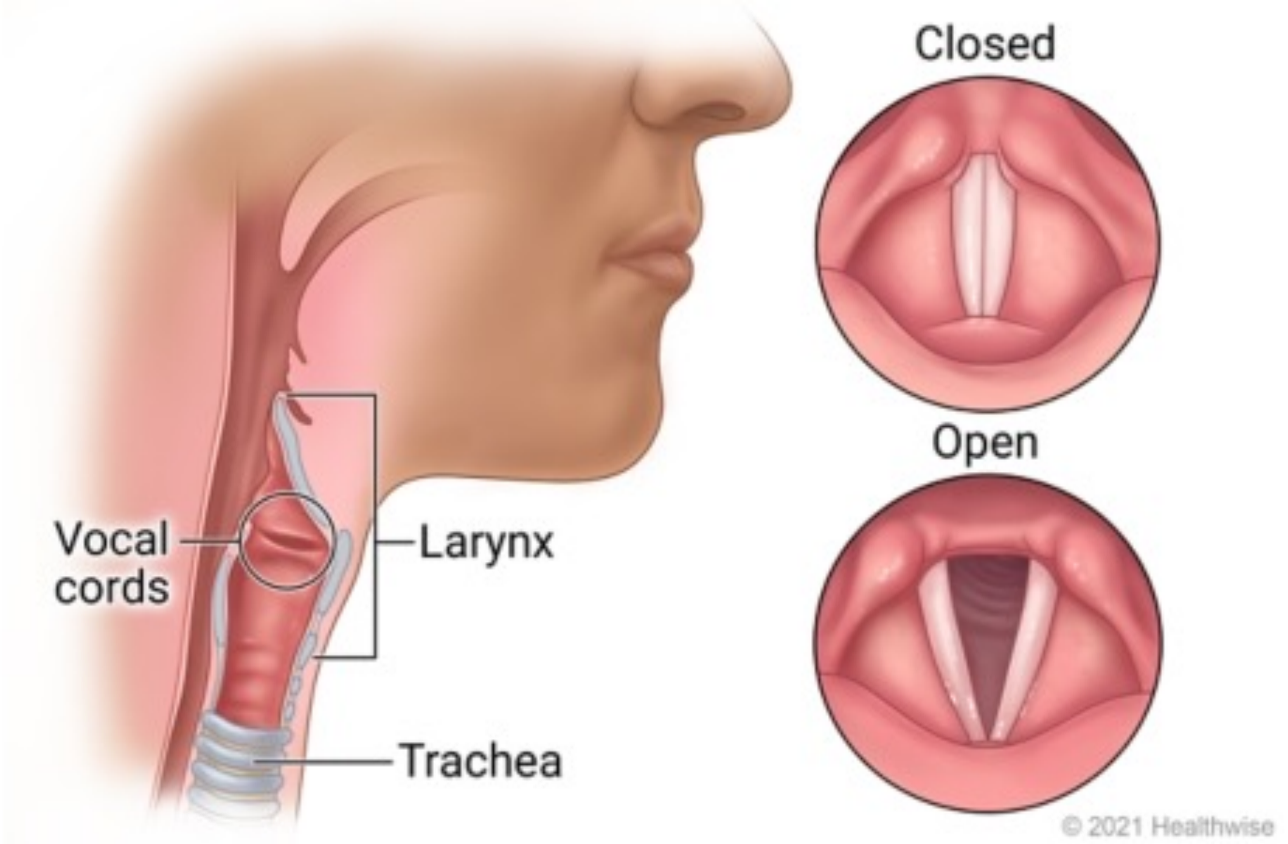
Tamil



USC's Signal Analysis and Intepretation Lab

So where does speech come from?

- Air come up from from the lungs
- Makes the vocal cords vibrate



- And the resulting pressure waves gets shaped by the tongue, mouth, lips

Figure from Ignite Healthwise, LLC Staff

The waveform: resonances of the vocal tract

The human vocal tract as an open tube

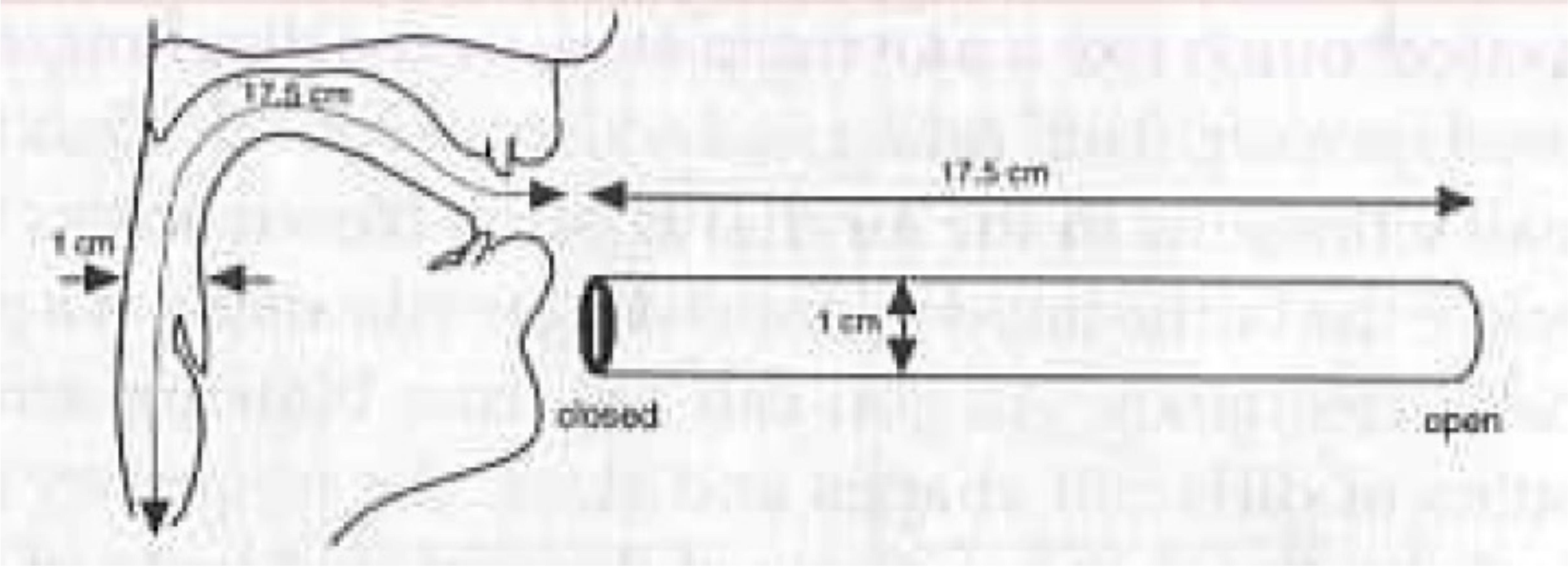
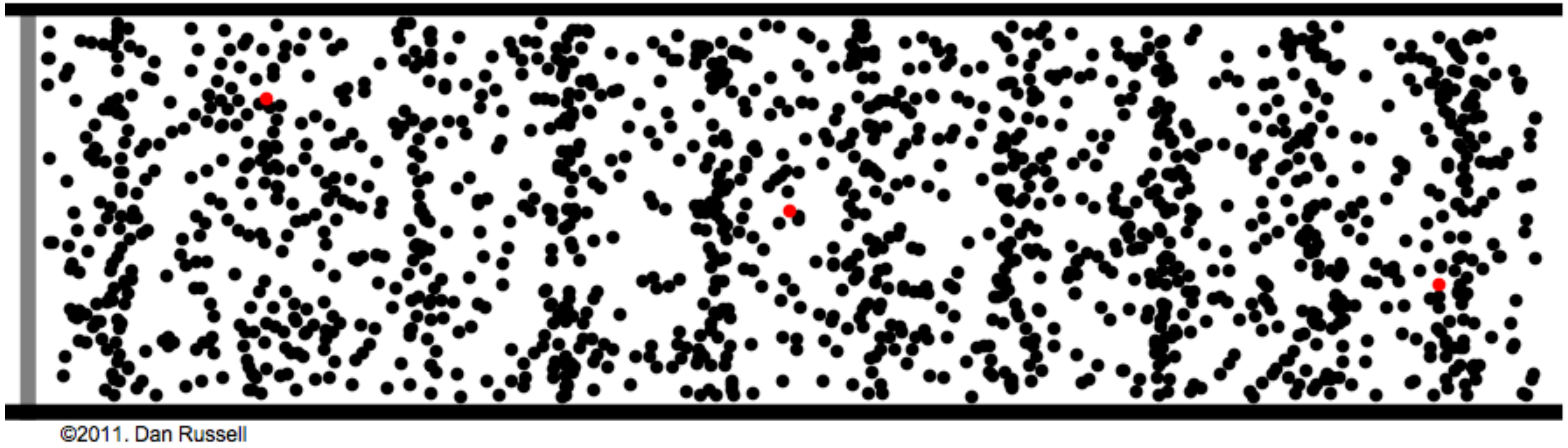
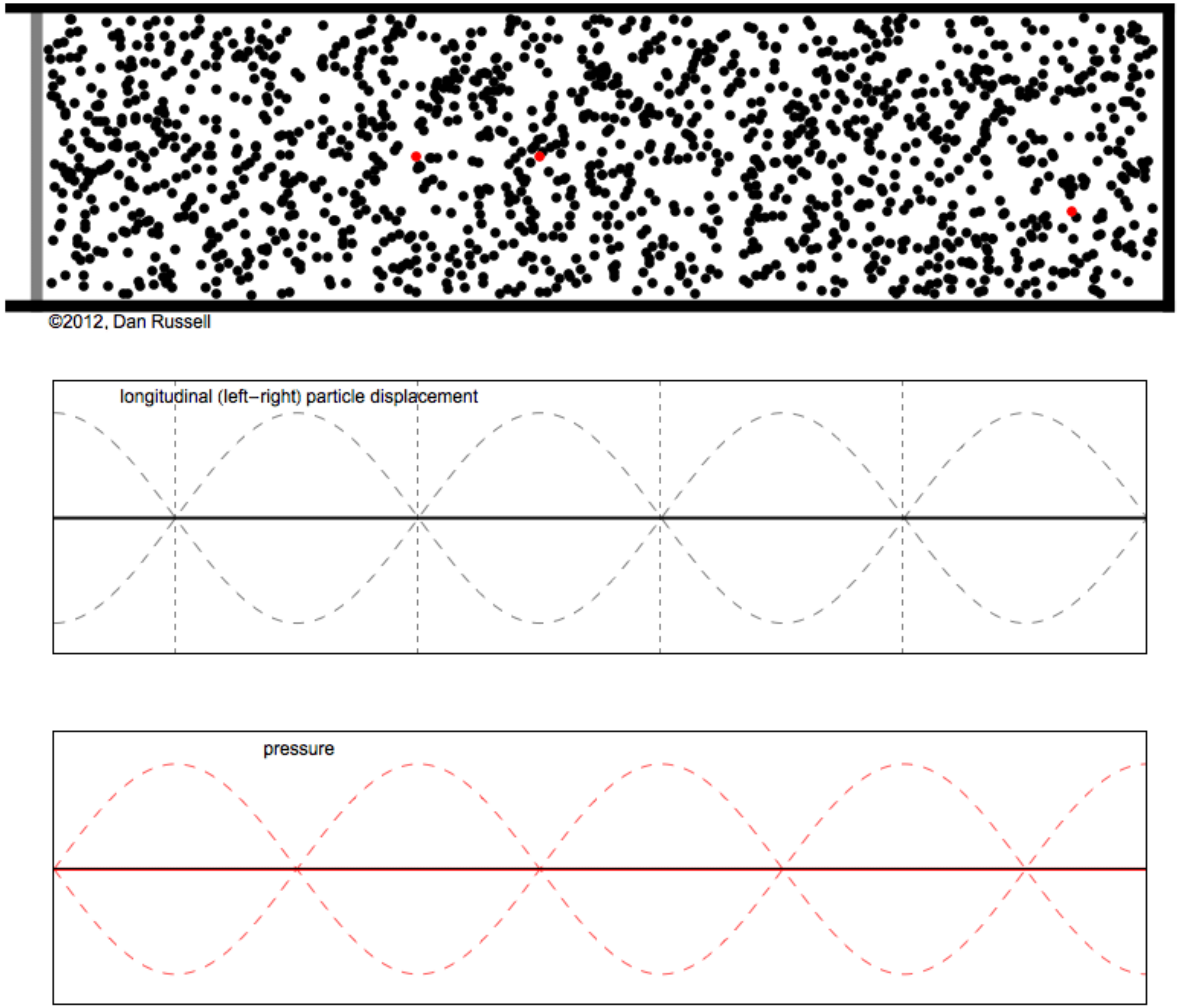


Figure from Ladefoged(1996) p 117

Sound waves are longitudinal waves



Dan Rusell Figure



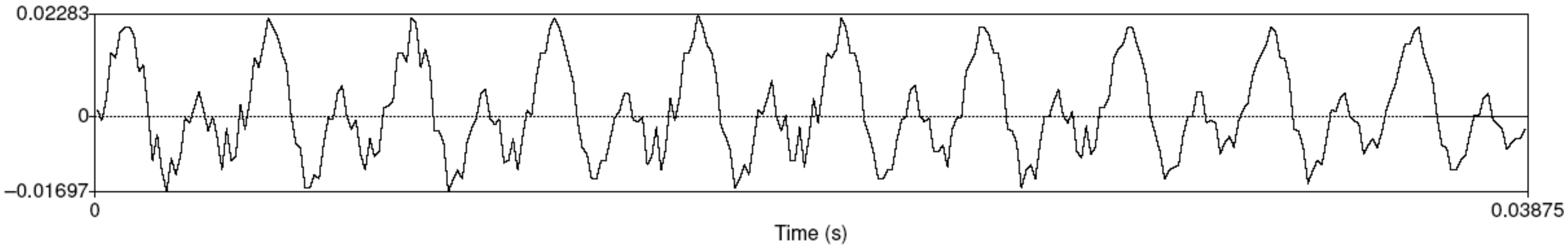
particle dispacment

pressure

Dan Rusell Figure

Speech sound waves

The shape of the mouth enhances some frequences and dampens others



X axis: time.

Y axis:

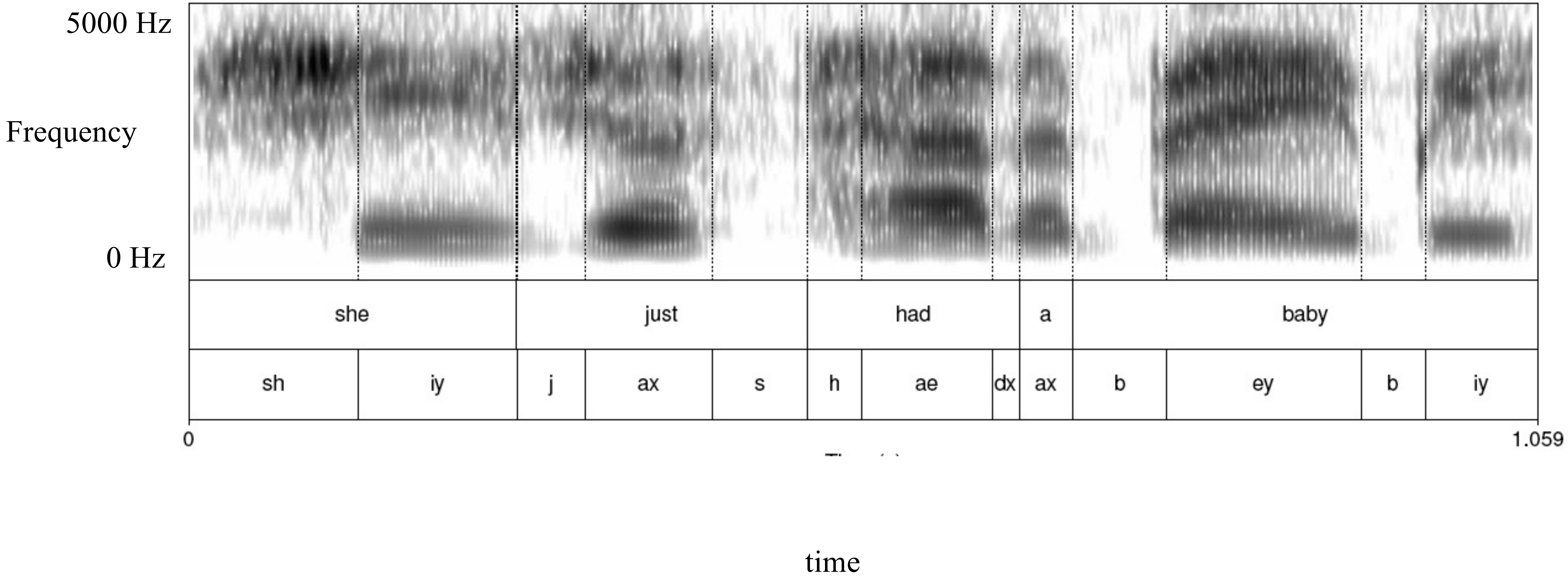
Amplitude = air pressure at that time

+: compression

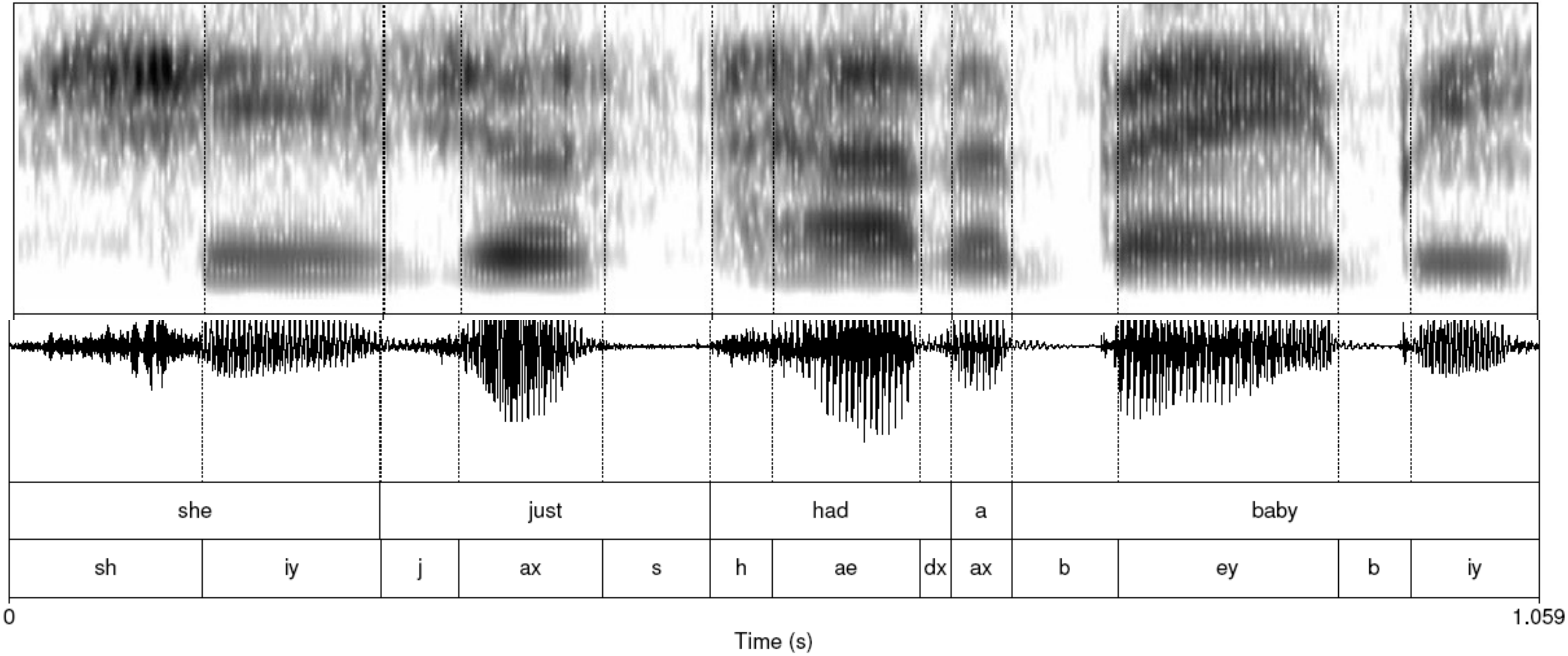
0: normal air pressure,

-: rarefaction

We can see these frequences in a spectrogram: spectrum (frequency dimension) + time dimension



She just had a baby



Speech Models

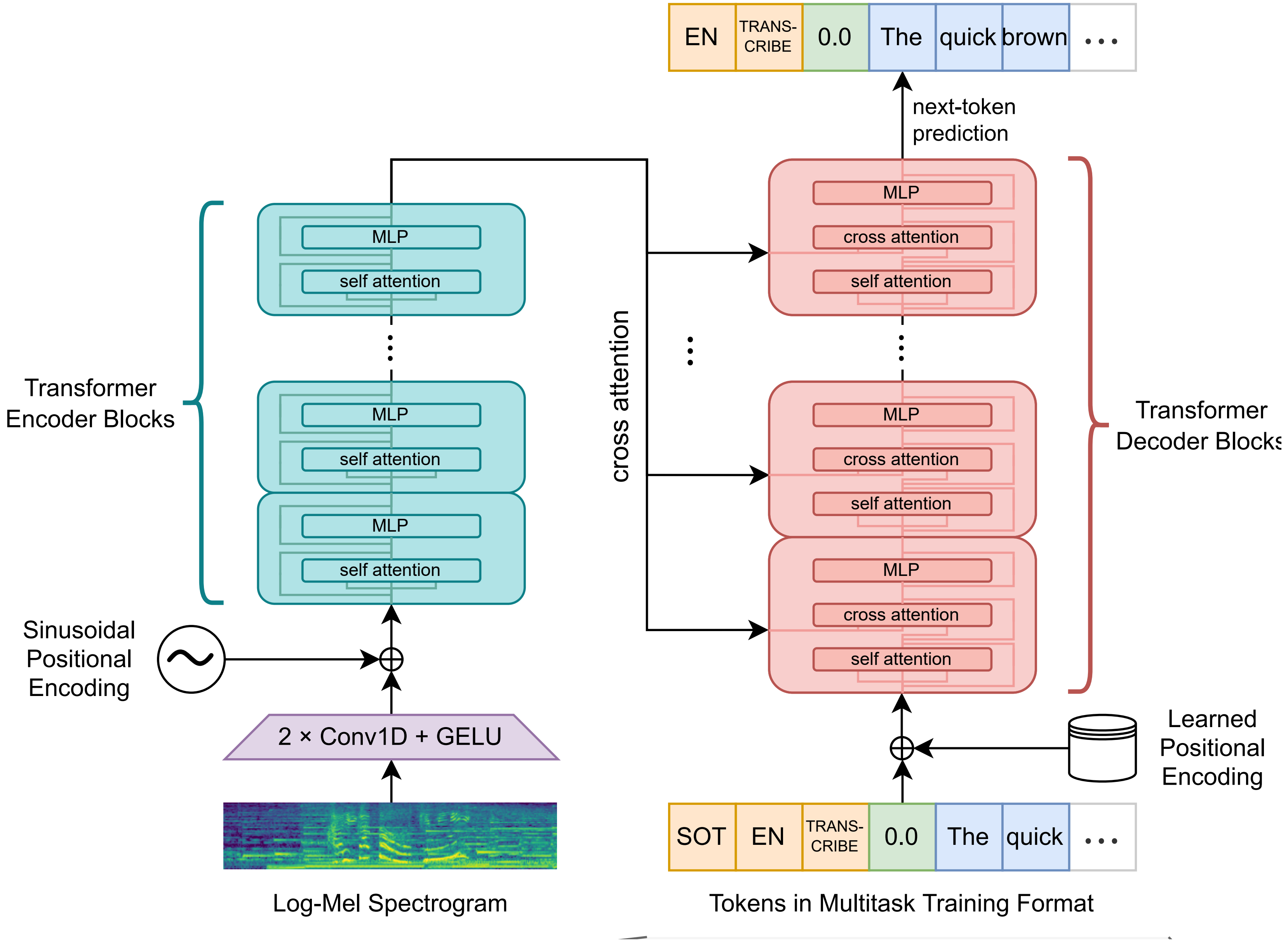
Intro to speech and speech recognition task

Speech Models

Whisper

An encoder-decoder model applied to speech!

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision." *ICML* pp. 28492-28518



Data

680,000 hours of multilingual and multitask supervision

All the data is audio paired with a transcript

Scraped from the web, with lots of filtering

Broken into chunks:

- 30 second audio, paired with transcript of words

Processing the input

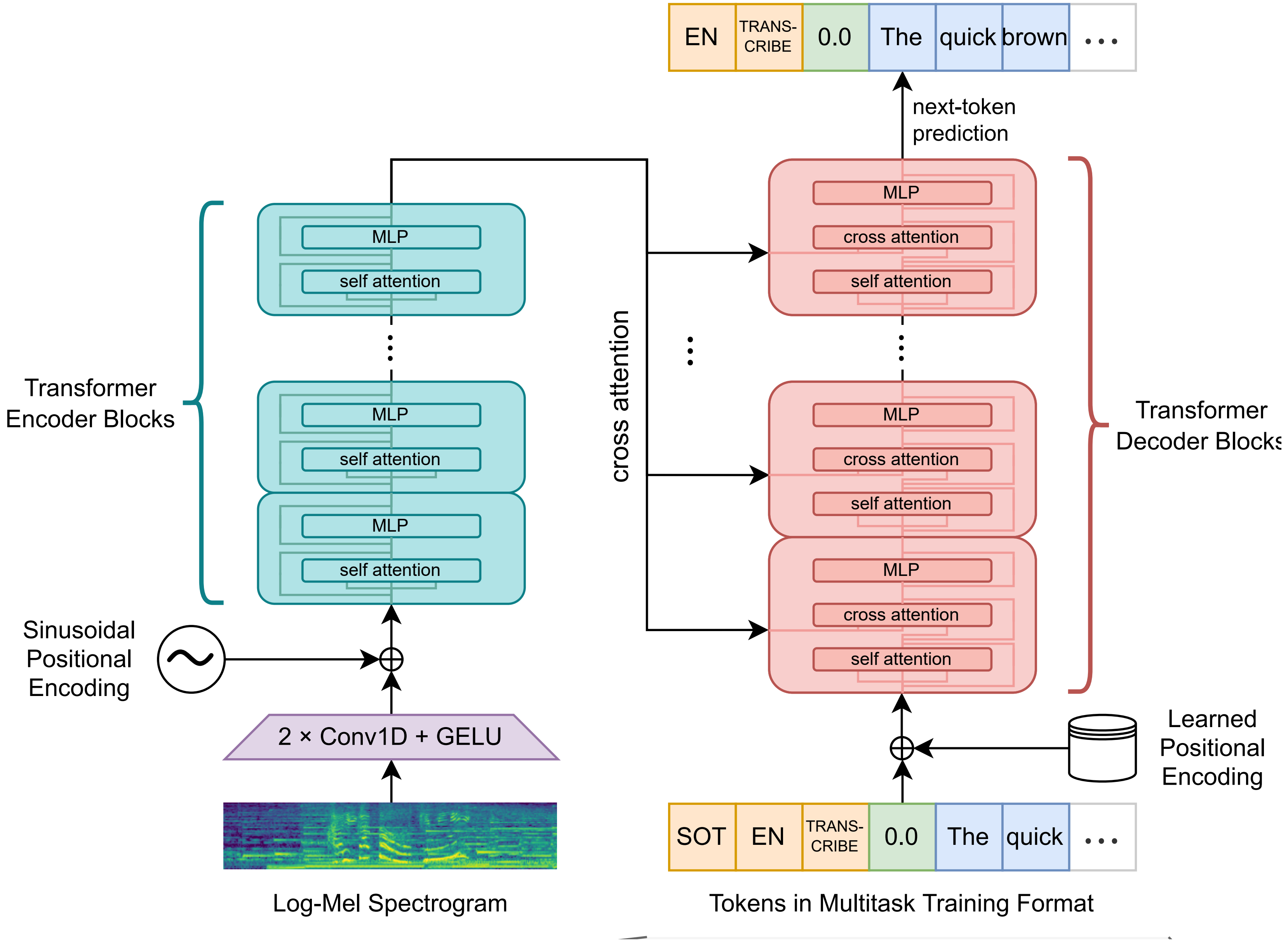
30 second audio:

- Run a 80-channel Mel spectrogram, every 10 ms, (so input vector is 3000 x 80)
- Run a small convolutional layer to upscale the 80 dimensions to 512
  - Result: a 1500 x 512 layer

Transcript

- Run BPE

Let's just see the ASR part:



Actually, it's multitask training setup

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

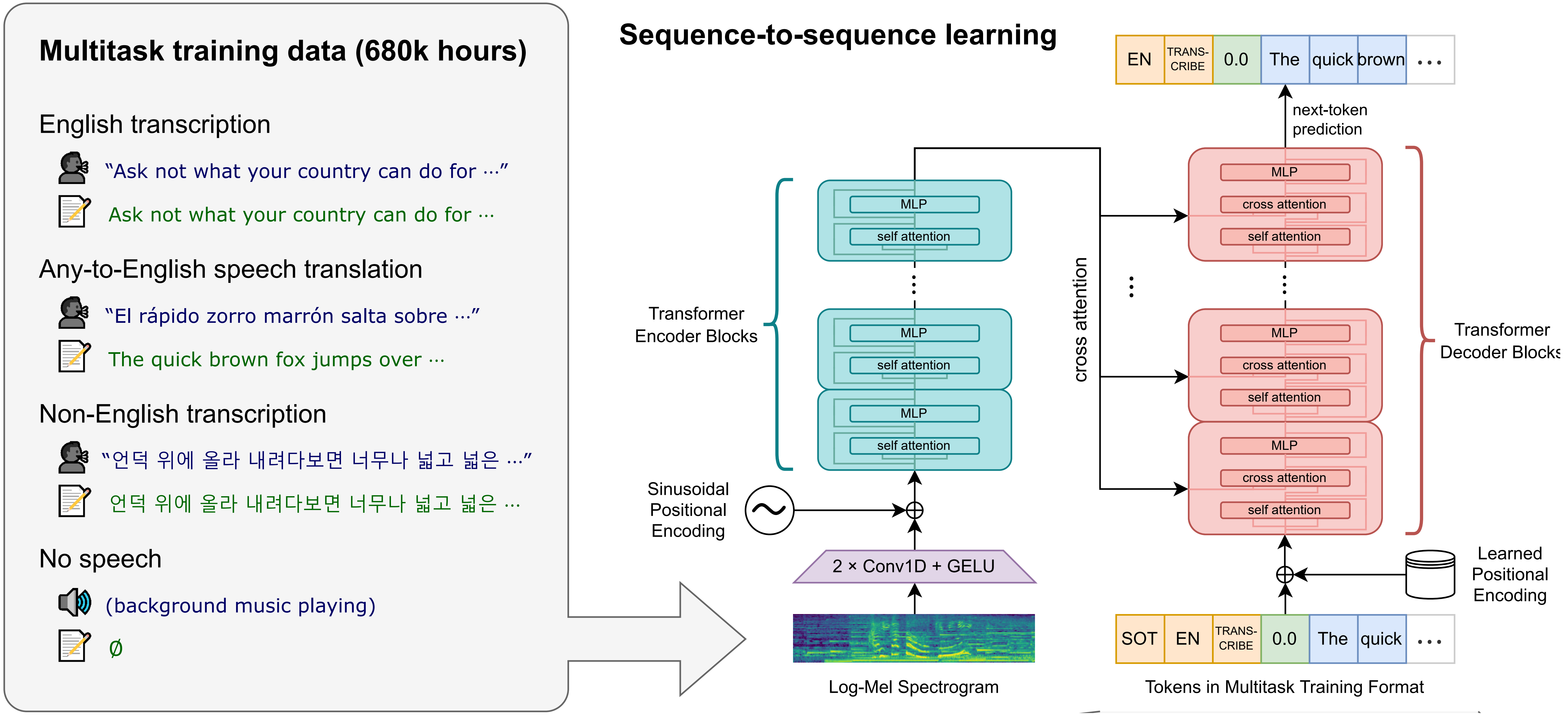
“언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯”

언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯

No speech

(background music playing)

∅


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
EN
TRANS-CRIBE
0.0
The
quick
brown
...
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
...
Tokens in Multitask Training Format

Speech Models

Whisper

Final class

Our last class together!

What is this class?

Interacting with humans via language

- Answering questions
- Searching the web
- Recommending things
- Helping in other ways

And extracting meaning from human language

- Via news, social media, websites, social networks, etc.

Learning goals

Understand algorithms in LLMs

- Logistic Regression
- Word embeddings
- Neural Networks
- Gradient Descent/Backprop
- Perplexity and Language Modeling Loss
- Transformers

And other language/social network systems:

- Regular Expressions
- Edit distance
- Collaborative filtering
- Information Retrieval
- Network centrality and PageRank

Be able to build

- Search engines
- Sentiment classifiers
- Chatbots

Be able to reason about sociotechnical questions

- Benefits of language technology
- Harms of classification (false positives and negatives)
- Harms of LLMs (privacy, hallucination, replacement)
- Social scientific applications of language technology (education, policing, political science, sociology)

What is this class?

The very broad undergrad intro to (at least) 12 grad classes!

cs224C: NLP for Computational Social Science (Yang)

cs224N: Natural Language Processing with Deep Learning (Hashimoto/Yang)

cs224U: Natural Language Understanding (Potts)

cs224V: Conversational Virtual Assistants with Deep Learning (Lam)

cs224S: Spoken Language Processing (Maas)

cs246: Mining Massive Data Sets (Leskovec)

cs224W: Graph Neural Networks (Leskovec)

cs276: Information Retrieval (Manning)

cs329R: Race and Natural Language Processing (Jurafsky/Eberhardt)

cs329X: Human-Centered LLMs (Yang)

cs336: Language modeling from scratch (Hashimoto/Liang)

cs384: Social and Ethical Issues in NLP (Jurafsky)

What's next? Spring 2025 NLP courses

CS 224S: Spoken Language Processing: (Andrew Maas): Intro to spoken language technology

CS 336: Language Modeling from Scratch (Tatsu Hashimoto and Percy Liang). Language model creation from scratch Application required.

CS 186: How to Make a Moral Agent (PHIL 86) **(David Gottlieb, Jared Moore)** Who is to blame if ChatGPT lies? Should we let superhuman AI make life and death decisions?

CS 229S - Systems for Machine Learning (Azalia Mirhoseini)
Performance-efficient training and inference, large focus on language models.

What's next? Spring 2025 NLP-adjacent courses

CS 221: Artificial Intelligence: Principles and Techniques (Anari, Charikar, Sadigh)

CS 277: Foundation Models for Healthcare (Chaudhari, Zou)

CS 278: Social Computing (Michael Bernstein) How do we design social computing systems - platforms for social media, online communities, and collaboration - to be effective and responsible?

CS 323: The AI Awakening: Implications for the Economy and Society (Brynjolfsson) How advances in AI are transforming the economy and society. Each week guest speakers

Next year NLP courses!

CS224N: Natural Language Processing with Deep Learning (Diyi Yang and Tatsu Hashimoto)

Algorithmic internals: transformers, GPT, parsing, machine translation and other applications. More of the gory details! More math, more machine learning

CS 293/EDUC473: Empowering Educators via Language Technology (Dora Demszky)

NLP x Education!

CS 224V: Conversational Virtual Assistants with Deep Learning (Monica Lam)

CS 246: Mining Massive Data Sets (Jure Leskovec)

CS329X: Human Centered NLP (Diyi Yang) Human-centered design thinking in NLP, human-in-the-loop algorithms, fairness, and accessibility.

CS329R: Race and NLP (Dan Jurafsky and Jennifer Eberhardt) NLP + social psychological perspectives on race to address societal issues

CS329A: Self-improvement AI Agents (Azalia Mirhoseini, Aakanksha Chowdhery) seminar on agents and model / tool orchestration

Fun courses outside of CS next year

Linguistics 150: Language and Society

COMM 154: The Politics of Algorithms

Or take a foreign language!!!

Or study abroad!
Spring 2026, I'm teaching "The Language of Food" abroad with Stanford BOSP Madrid campus!!!