

# CSCSE 638 Natural Language Processing Foundation and Techniques

## Lecture 10: Large Language Models

Kuan-Hao Huang

Spring 2026



(Some slides adapted from Graham Neubig)

# Quiz 1

- Average: 82.7
- Std: 9.04
- Q1: 78
- Median: 84
- Q3: 90

Check [Gradescope](#) for details. For questions, send emails to [csce638-ta-26s@lists.tamu.edu](mailto:csce638-ta-26s@lists.tamu.edu) with “[CSCE 638] Subject ...” or check with TA in office hours

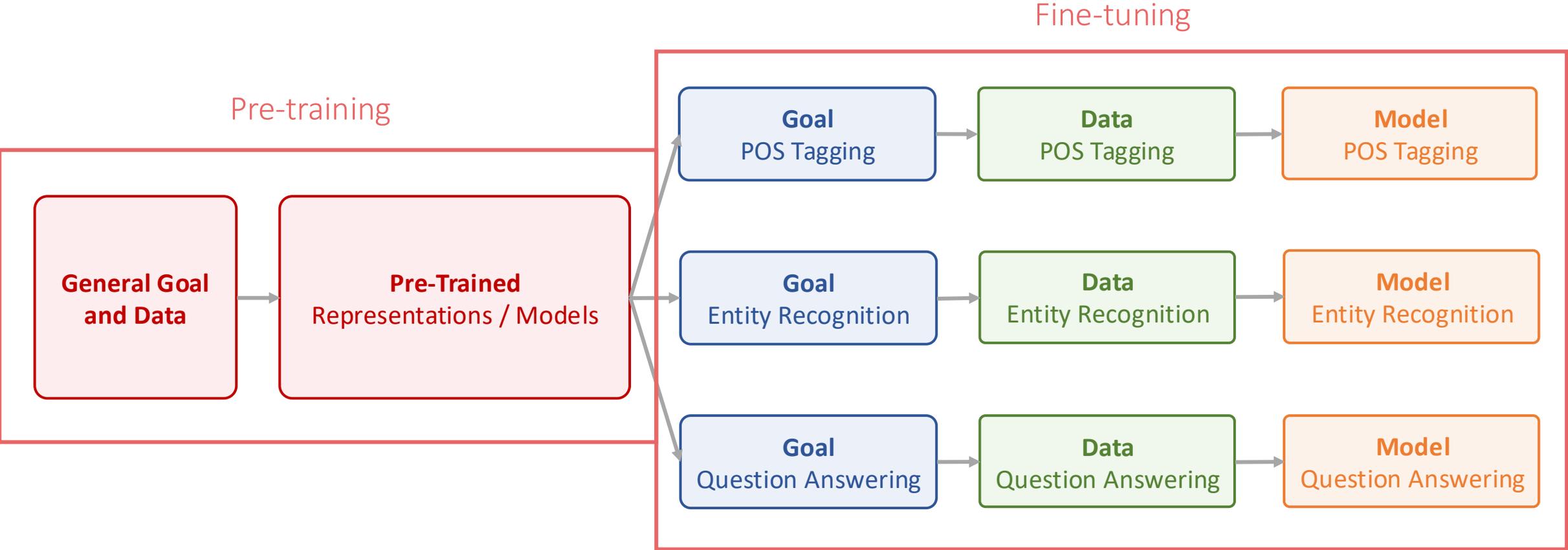
# Quiz 2

- Feb 25 (Wednesday)
- Coverage: mainly Lecture 6 to 10
  - Naturally include some concepts in Lecture 1 to 5
- In-class, 20 minutes, closed book, no cheat sheet
- Written quiz, 5 questions
  - Please bring a pen
- Tips
  - Get familiar with formula
  - Understand the intuition behind the formula and the design
  - Know the pros and cons of different approaches

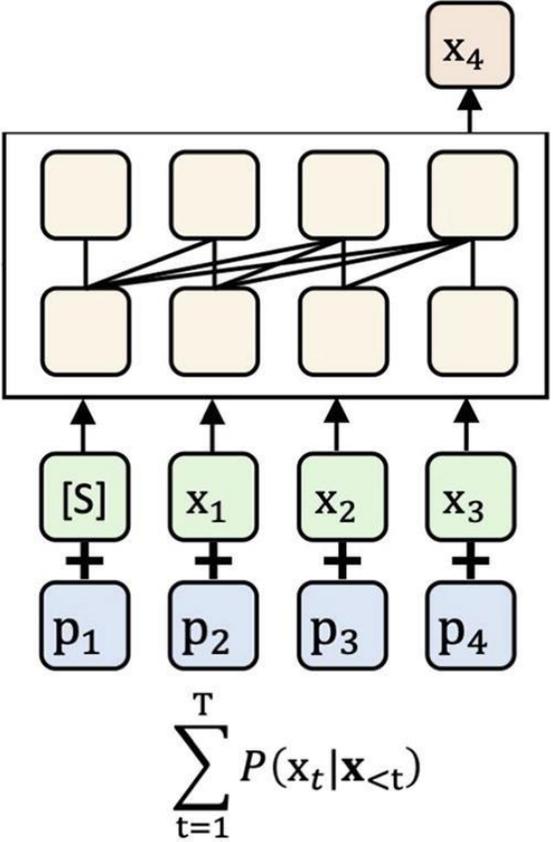
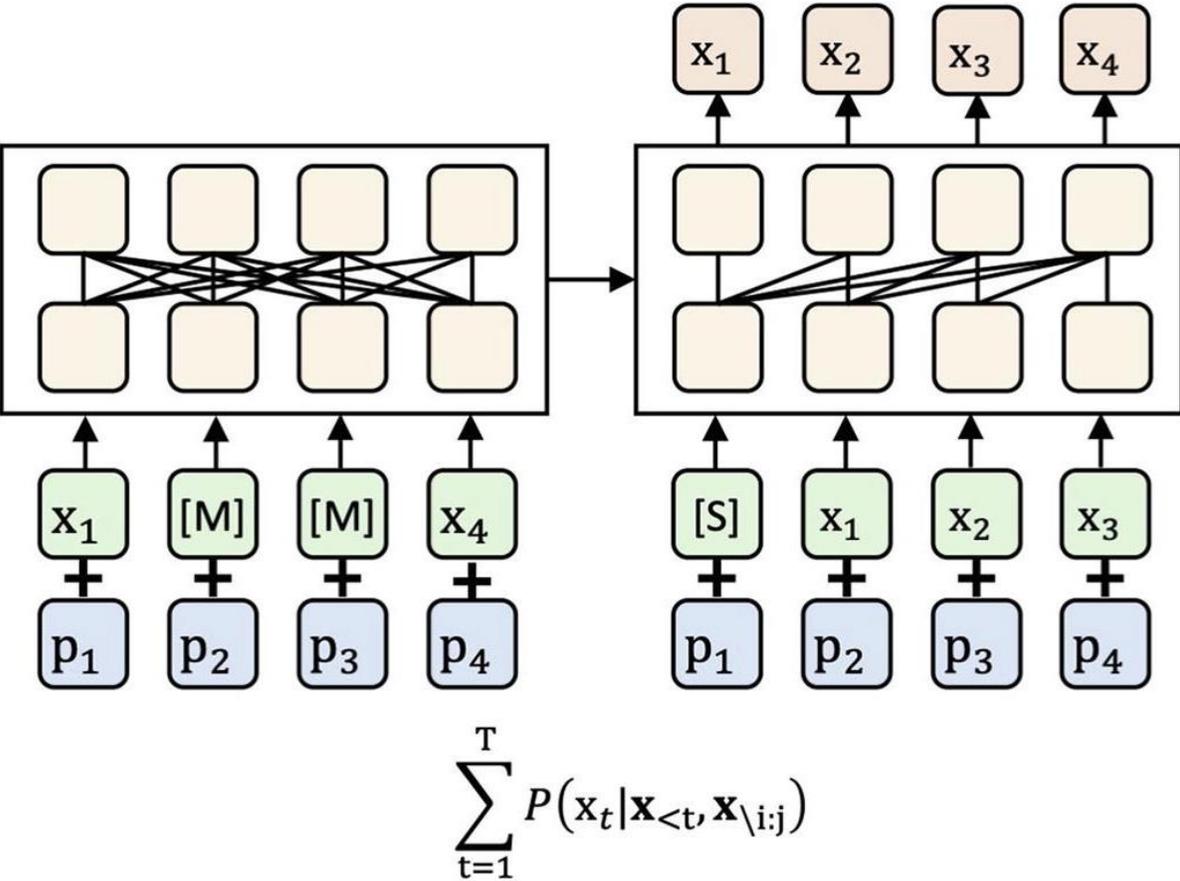
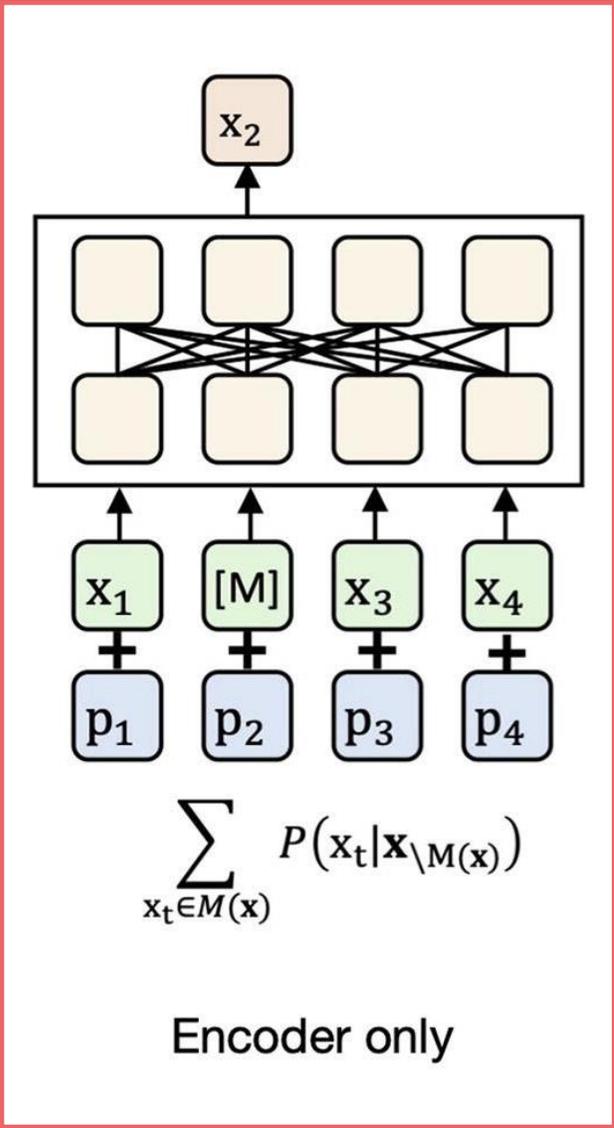
# Lecture Plan

- Large Language Models (LLMs)
  - In-Context Learning
  - Chain-of-Thought
  - Popular LLMs

# Recap: Fine-Tuning with Pre-Training

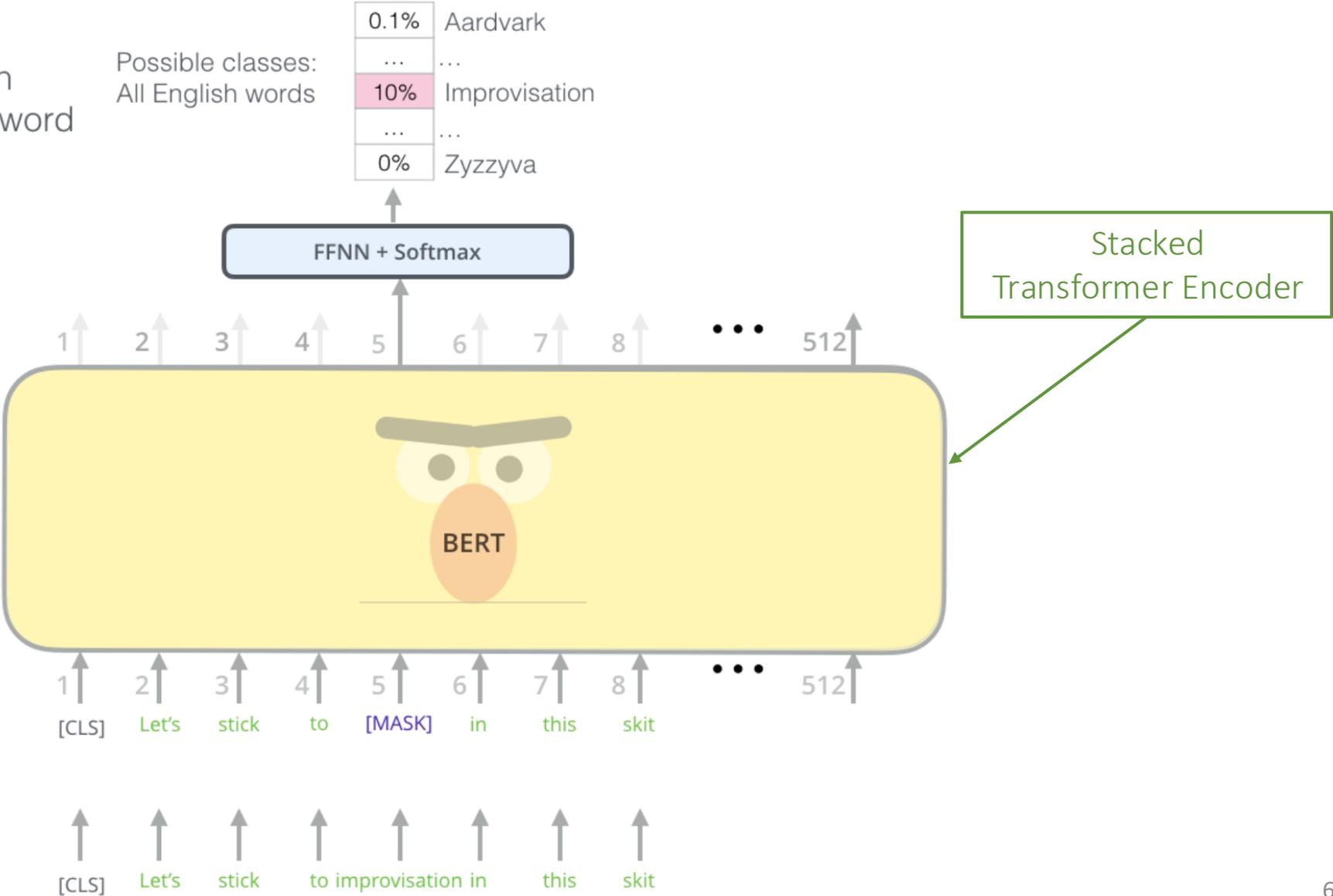


# Recap: Types of Pre-Training



# Recap: BERT – Masked Language Modeling

Use the output of the masked word's position to predict the masked word

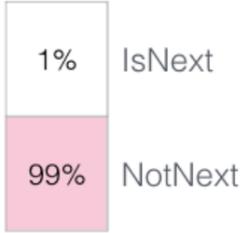


Randomly mask 15% of tokens

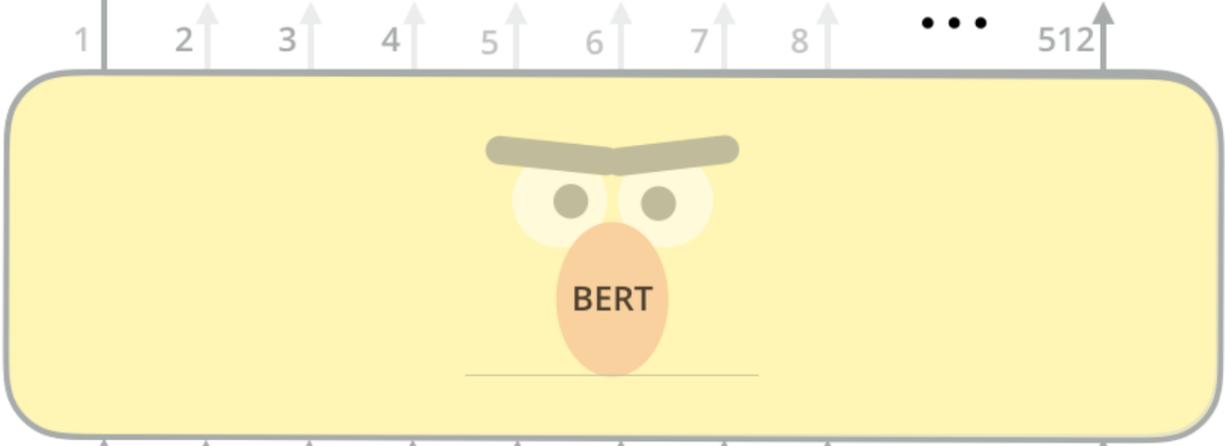
Input

# Recap: BERT – Next Sentence Prediction

Predict likelihood that sentence B belongs after sentence A



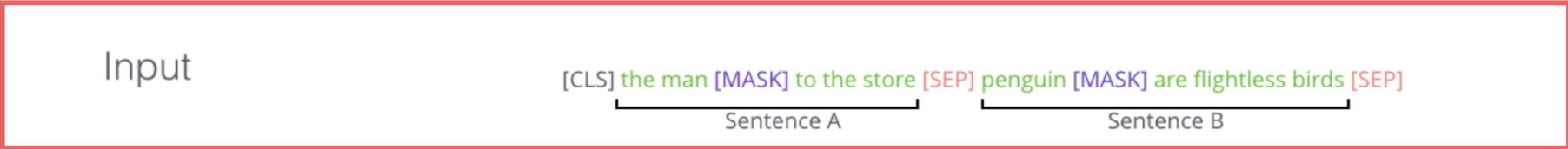
FFNN + Softmax



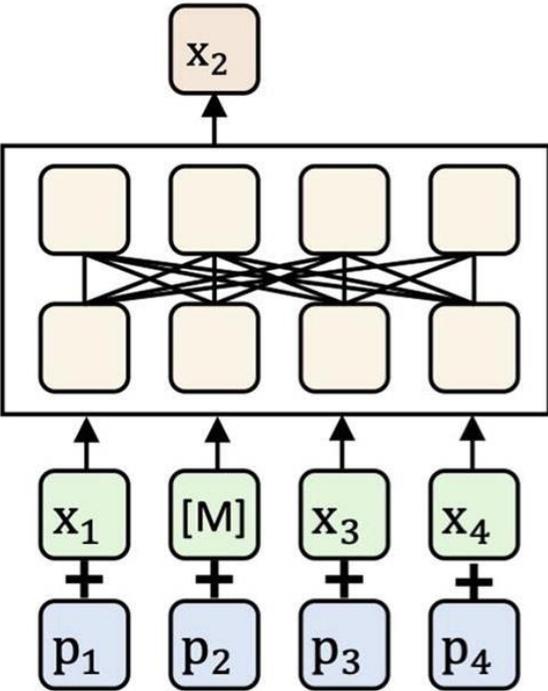
Positive example: real next sentence  
Negative example: random sentence

Tokenized Input

1 [CLS] 2 the 3 man 4 [MASK] 5 to 6 the 7 store 8 [SEP] ... 512

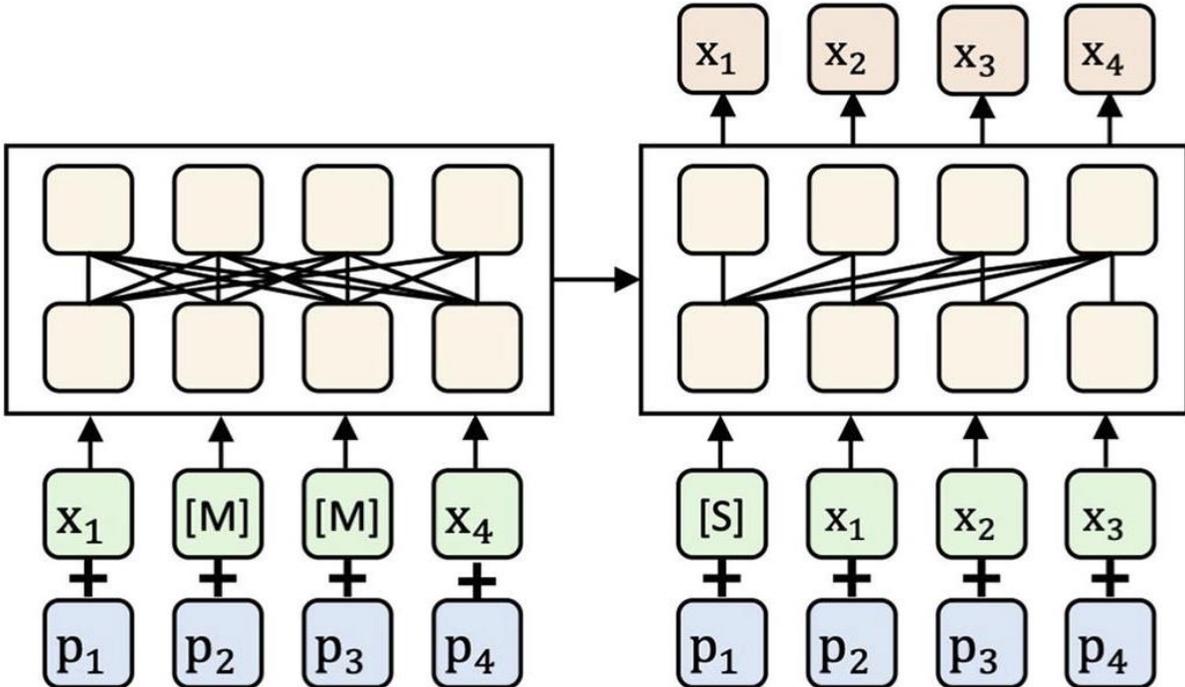


# Recap: Types of Pre-Training



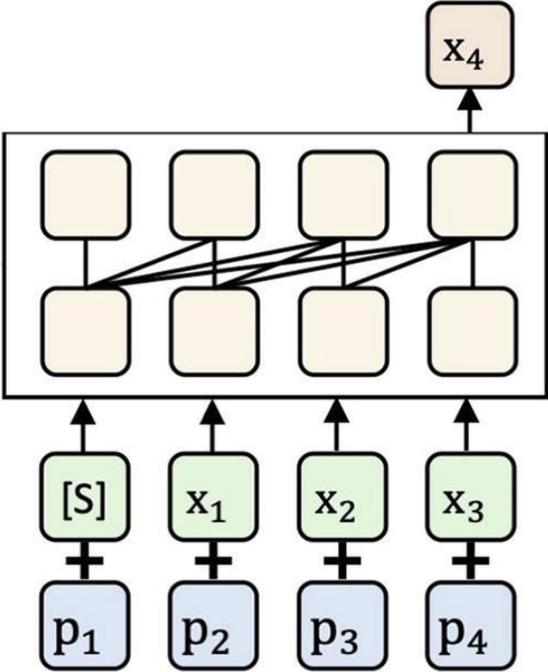
$$\sum_{x_t \in M(x)} P(x_t | \mathbf{x}_{\setminus M(x)})$$

Encoder only



$$\sum_{t=1}^T P(x_t | \mathbf{x}_{<t}, \mathbf{x}_{\setminus i;j})$$

Encoder-decoder

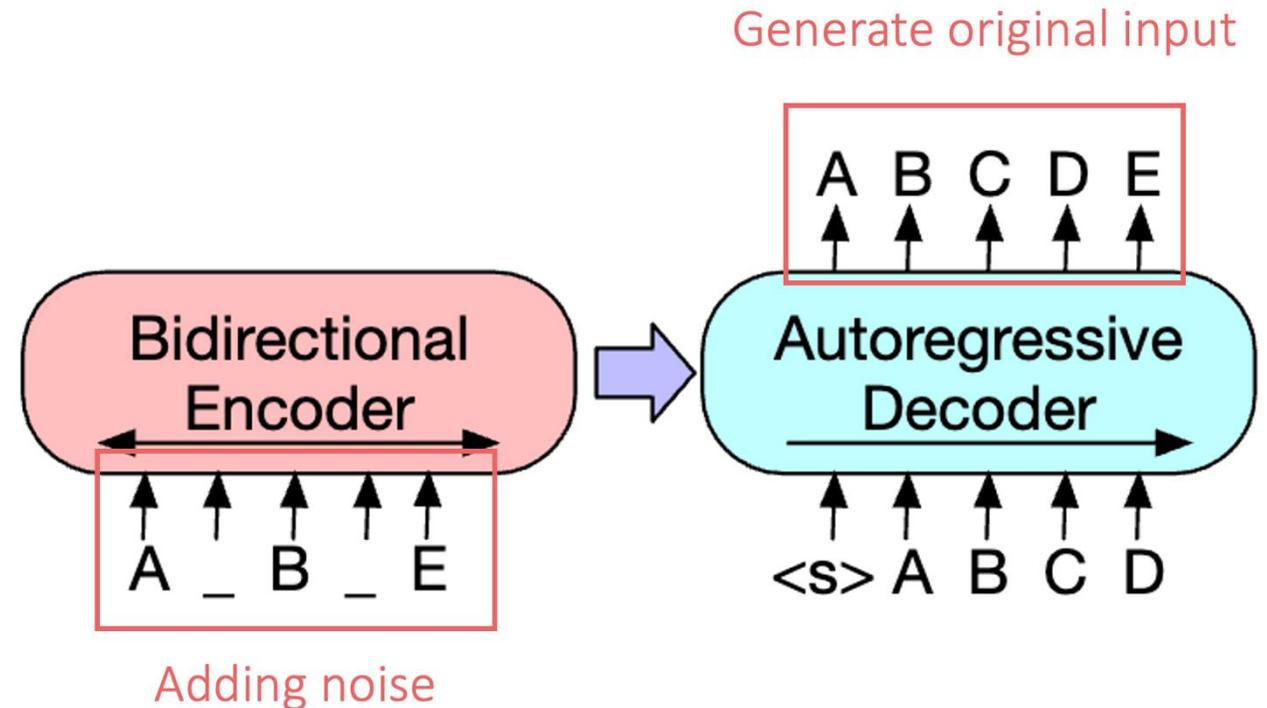


$$\sum_{t=1}^T P(x_t | \mathbf{x}_{<t})$$

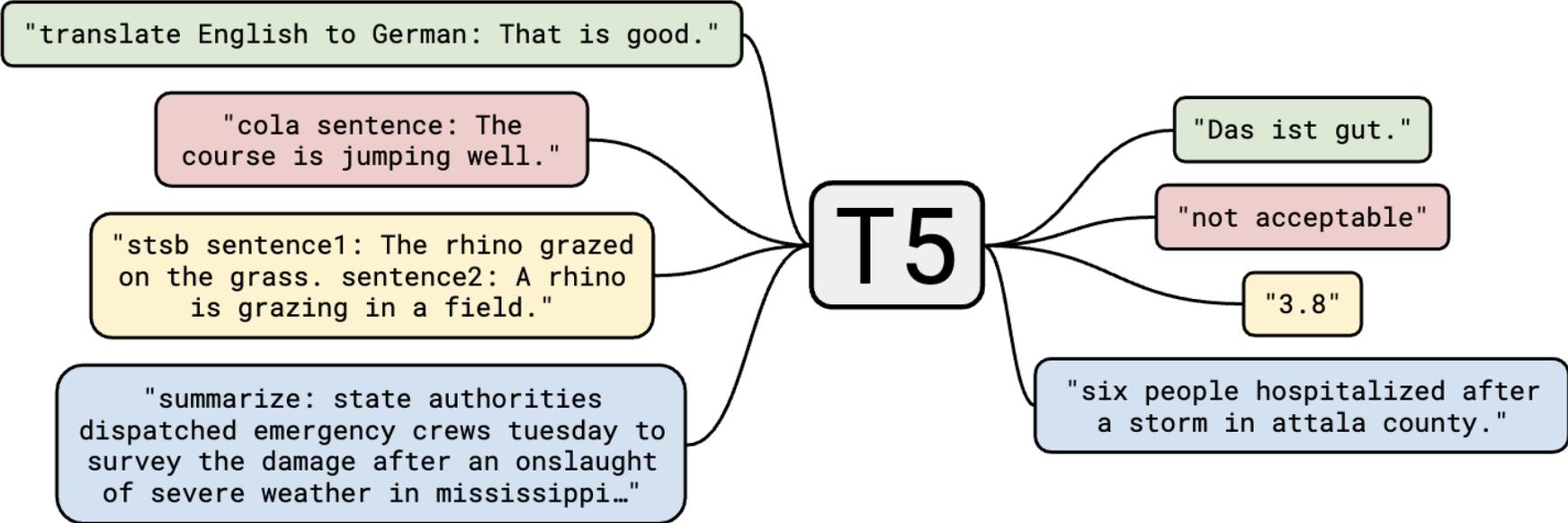
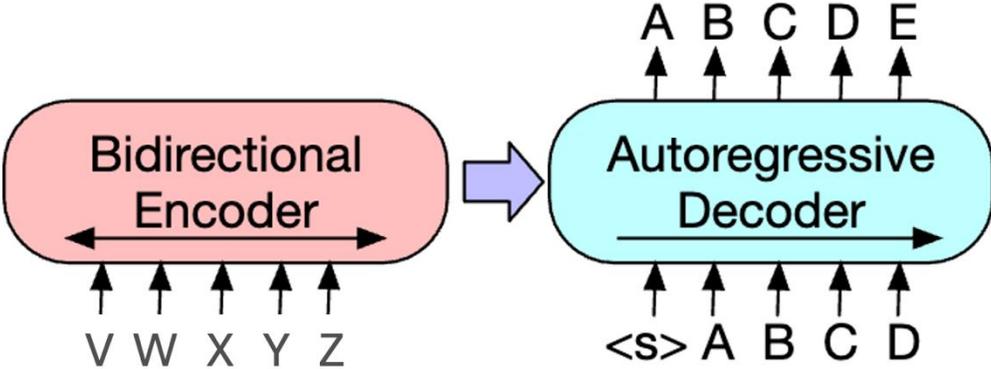
Decoder only

# Recap: BART – Denoising Objective

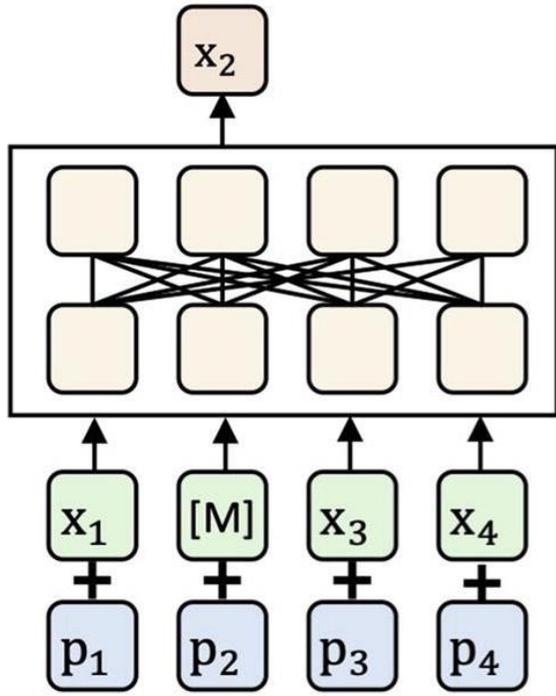
- Token Masking
  - A<mask>CD<mask>F. → ABCDEF
- Token Deletion
  - ACDF. → ABCDEF.
- Text Infilling
  - A<mask>D<mask>F. → ABCDEF.
- Sentence Permutation
  - FG. ABC. DE. → ABC. DE. FG.
- Document Rotation
  - E. FG. ABC. D → ABC. DE. FG.



# Recap: T5 – Convert Everything to Text-to-Text Tasks

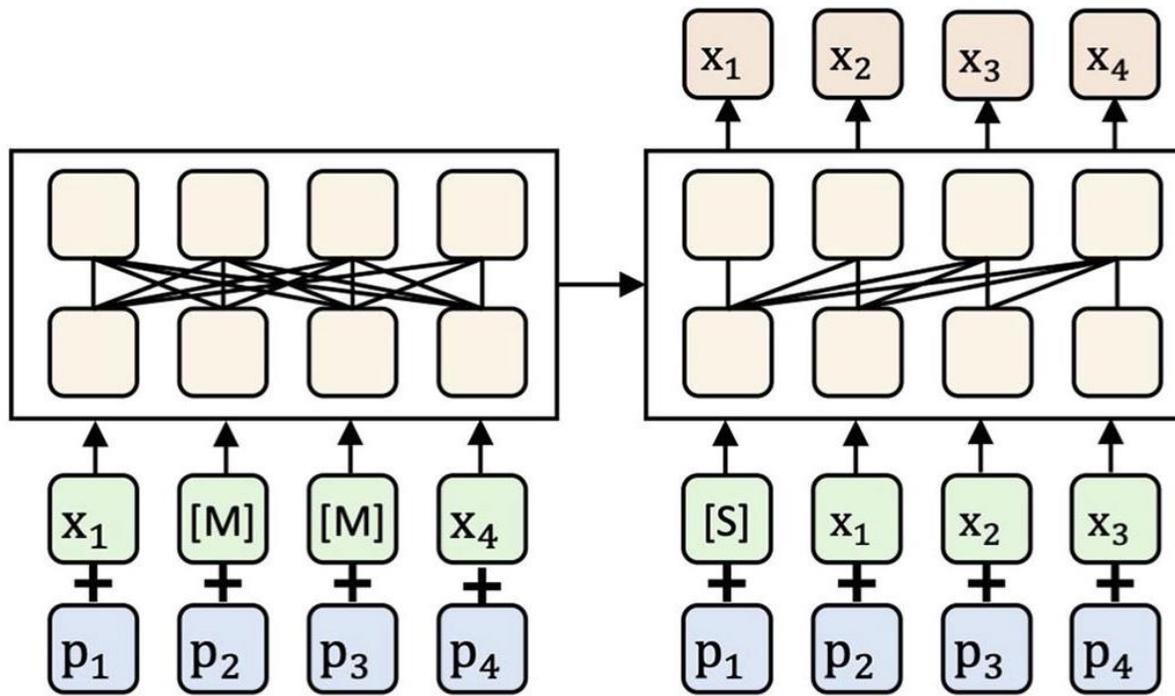


# Types of Pre-Training



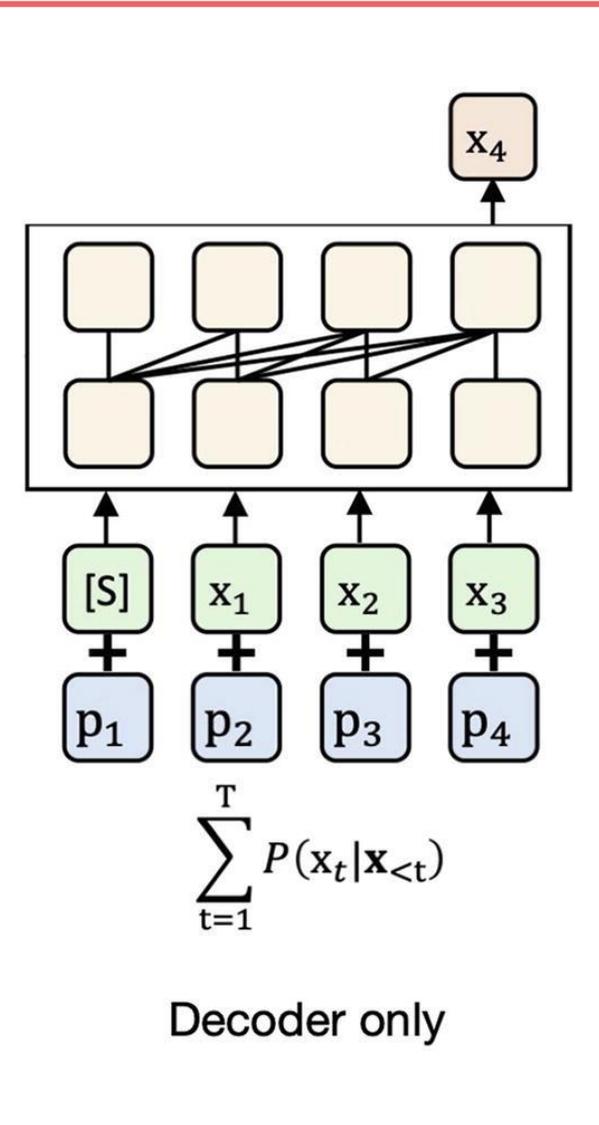
$$\sum_{x_t \in M(x)} P(x_t | \mathbf{x}_{\setminus M(x)})$$

Encoder only



$$\sum_{t=1}^T P(x_t | \mathbf{x}_{<t}, \mathbf{x}_{\setminus i:j})$$

Encoder-decoder

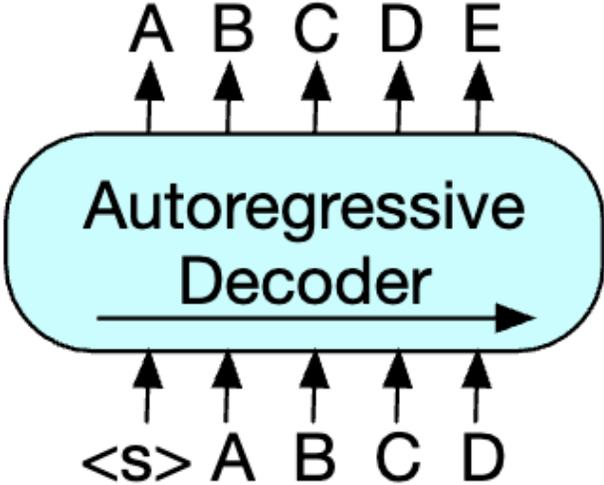


$$\sum_{t=1}^T P(x_t | \mathbf{x}_{<t})$$

Decoder only

# Language Modeling

- Next word prediction
- Trained with large corpus



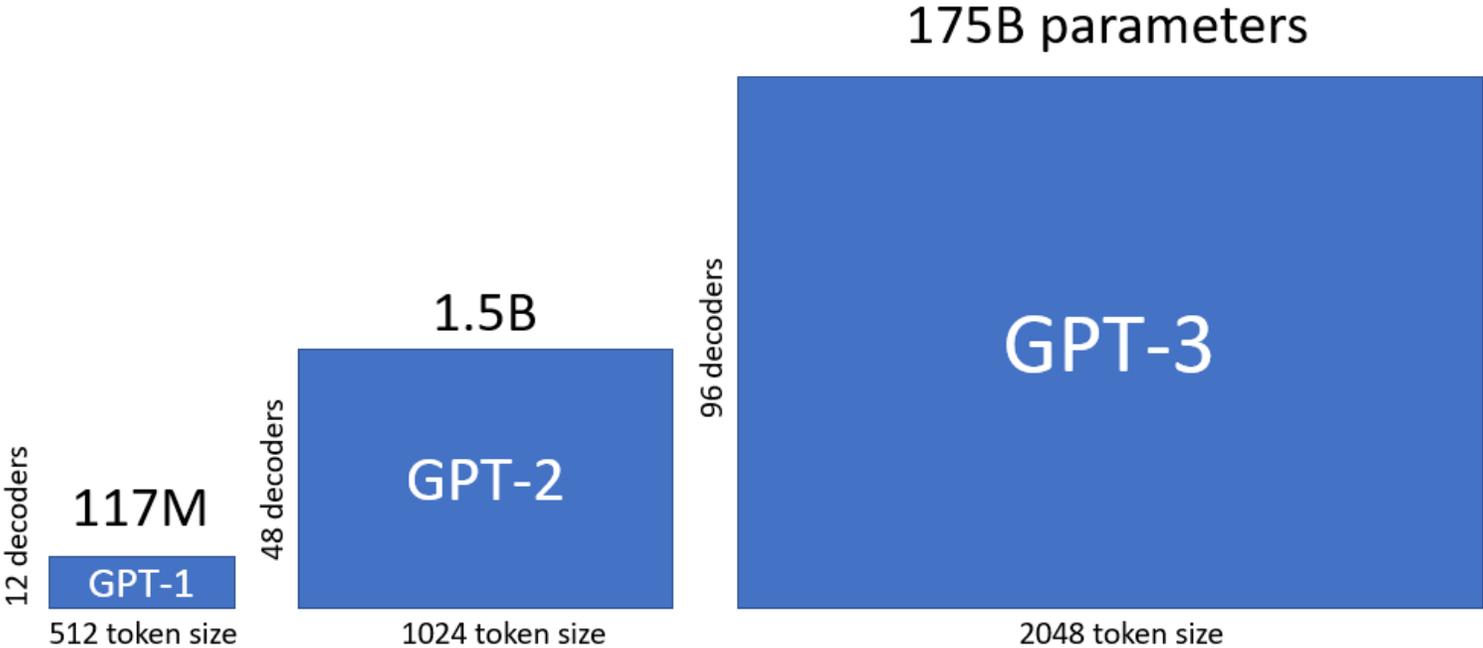
Binge ... on | - | and | of | is  
Binge **drinking** ... is | and | had | in | was  
Binge drinking **may** ... be | also | have | not | increase  
Binge drinking may **not** ... be | have | cause | always | help  
Binge drinking may not **necessarily** ... be | lead | cause | results | have  
Binge drinking may not necessarily **kill** ... you | the | a | people | your  
Binge drinking may not necessarily kill **or** ... even | injure | kill | cause | prevent  
Binge drinking may not necessarily kill or **even** ... kill | prevent | cause | reduce | injure  
Binge drinking may not necessarily kill or even **damage** ... your | the | a | you | someone  
Binge drinking may not necessarily kill or even damage **brain** ... cells | functions | tissue | neurons  
Binge drinking may not necessarily kill or even damage brain **cells,** ... some | it | the | is | long

# Decoder-Only: Generative Pre-trained Transformer (GPT)

- Improving Language Understanding by Generative Pre-Training, OpenAI 2018
  - Generative Pre-trained Transformer (GPT)
- Language Models are Unsupervised Multitask Learners, OpenAI 2019
  - GPT-2
- Language Models are Few-Shot Learners, OpenAI 2020
  - GPT-3

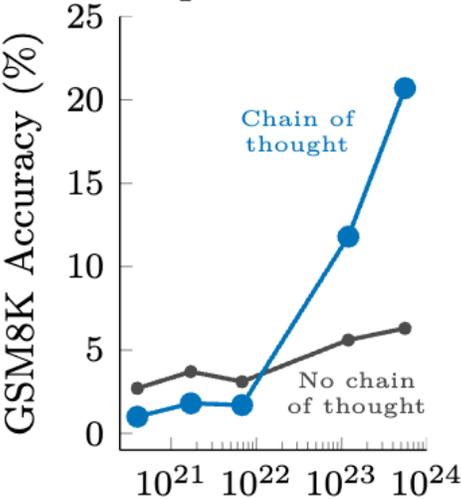
# GPT-3

- Even larger training data, even larger model size

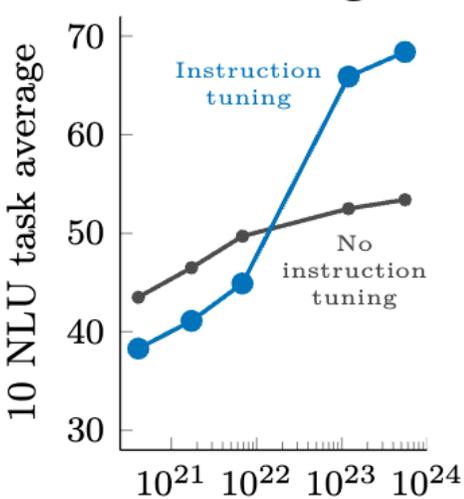


# Scaling Is The Key

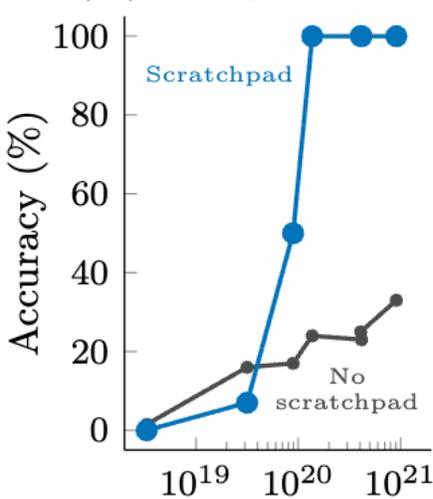
(A) Math word problems



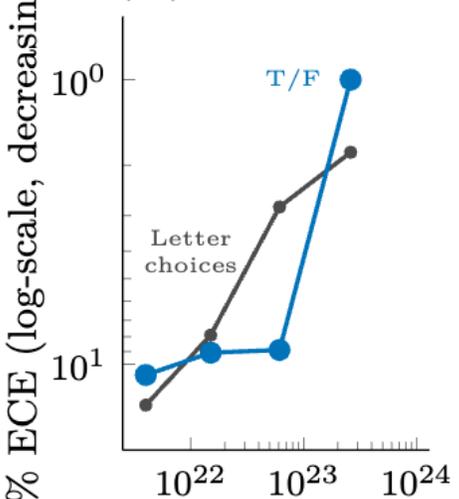
(B) Instruction following



(C) 8-digit addition



(D) Calibration



# Zero-Shot Prompting

- Prompt → Completion
  - Continue writing

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

# Zero-Shot Prompting

- Prompt → Completion
  - Continue writing

Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

sport.

Completion

# Zero-Shot Prompting

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

sport.

Completion

# A New Way to Use NLP Models

- Task-specific features + task-specific model
- General embeddings + task-specific model
- General embeddings + general model + task-specific fine-tuning
- General embeddings + general model + **task-specific prompting**

# Zero-Shot Prompting

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

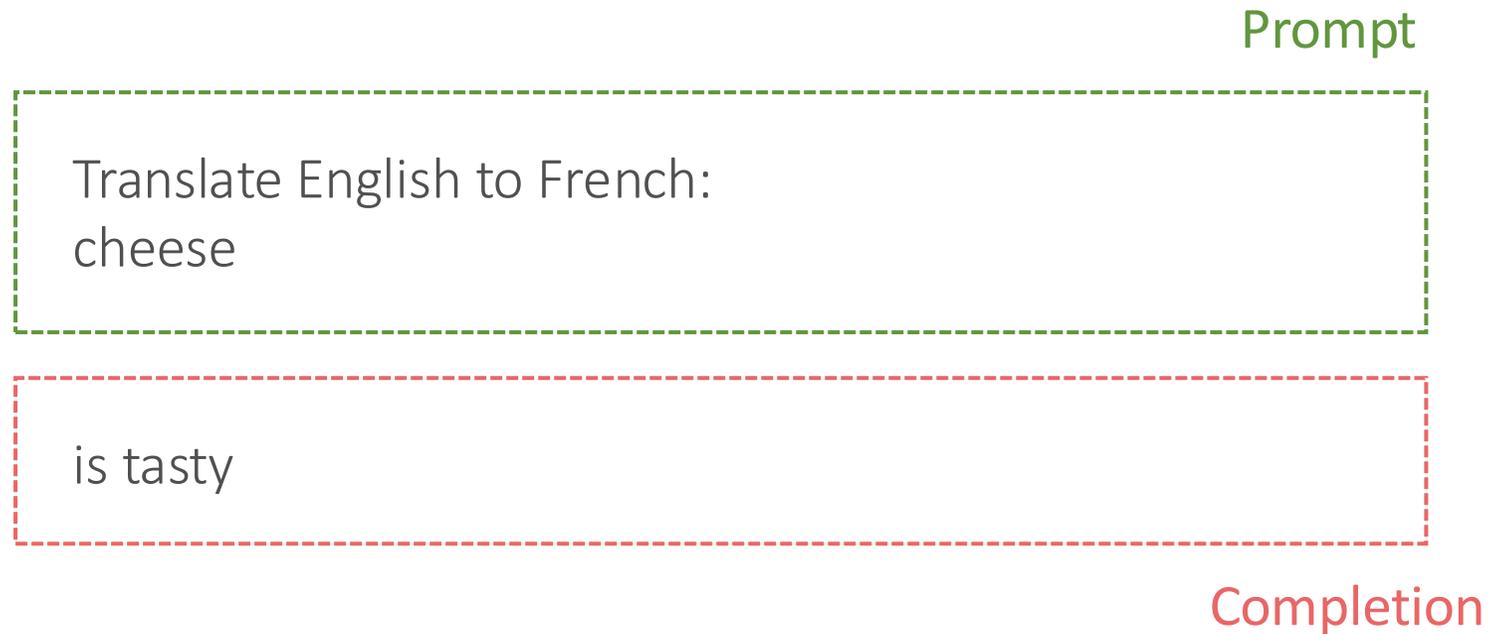
sport.

Completion

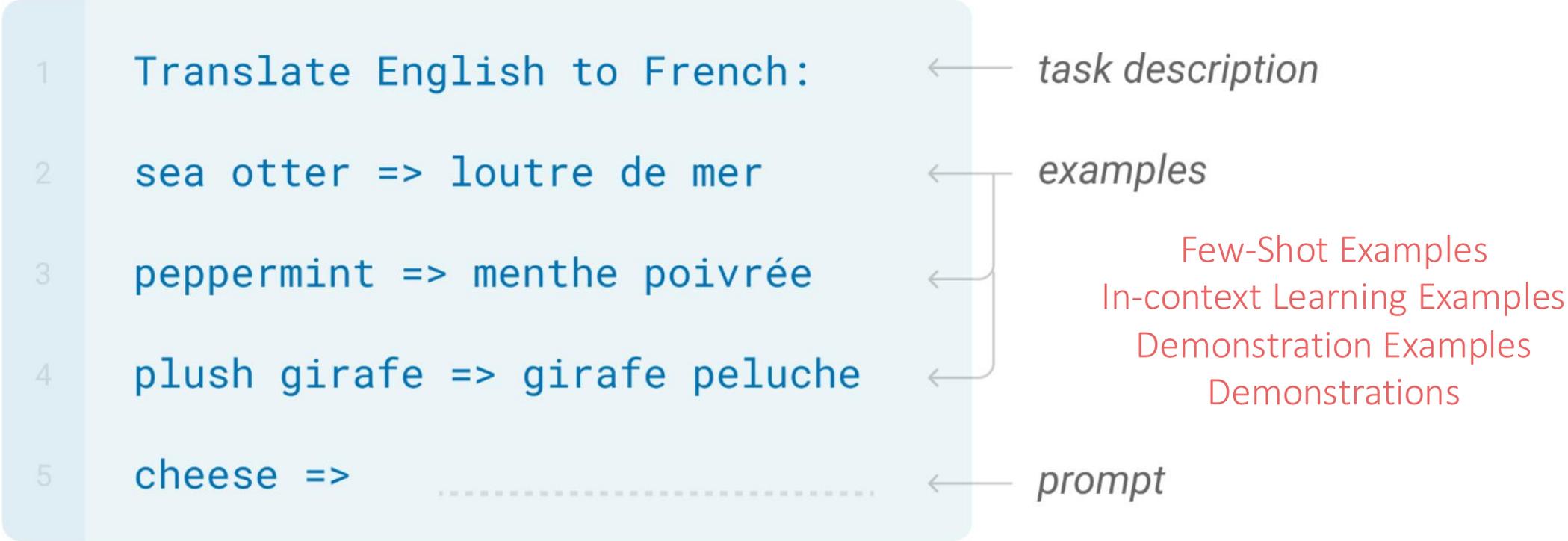
Any Issues?

# Zero-Shot Prompting

- Prompt → Completion
  - Continue writing



# Few-Shot Prompting / In-Context Learning



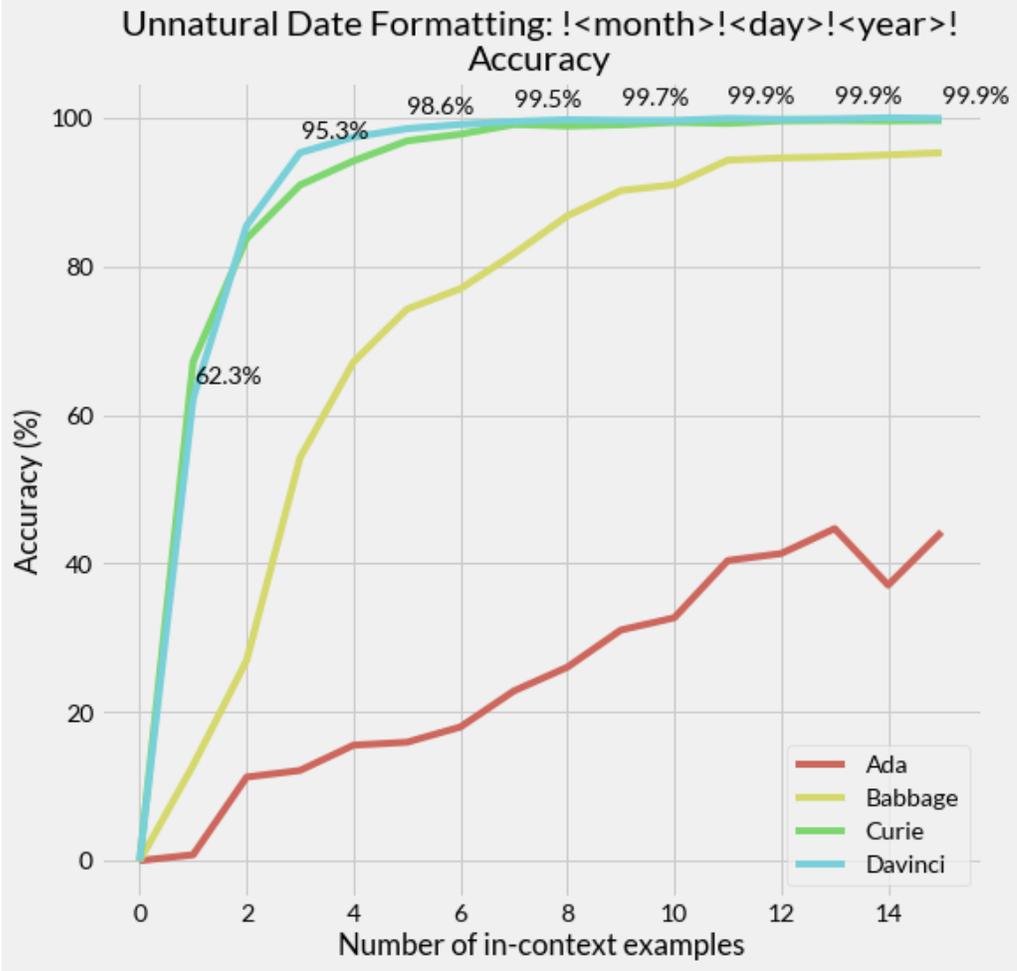
# Few-Shot Prompting / In-Context Learning

Input: 2014-06-01  
Output: !06!01!2014!  
Input: 2007-12-13  
Output: !12!13!2007!  
Input: 2010-09-23  
Output: !09!23!2010!  
Input: **2005-07-23**  
Output: **!07!23!2005!**

*in-context examples*

*test example*

*model completion*



# Few-Shot Prompting / In-Context Learning

## *Demonstrations*

Circulation revenue has increased by 5% in Finland.	\n	Positive
Panostaja did not disclose the purchase price.	\n	Neutral
Paying off the national debt will be extremely painful.	\n	Negative
The acquisition will have an immediate positive impact.	\n	_____

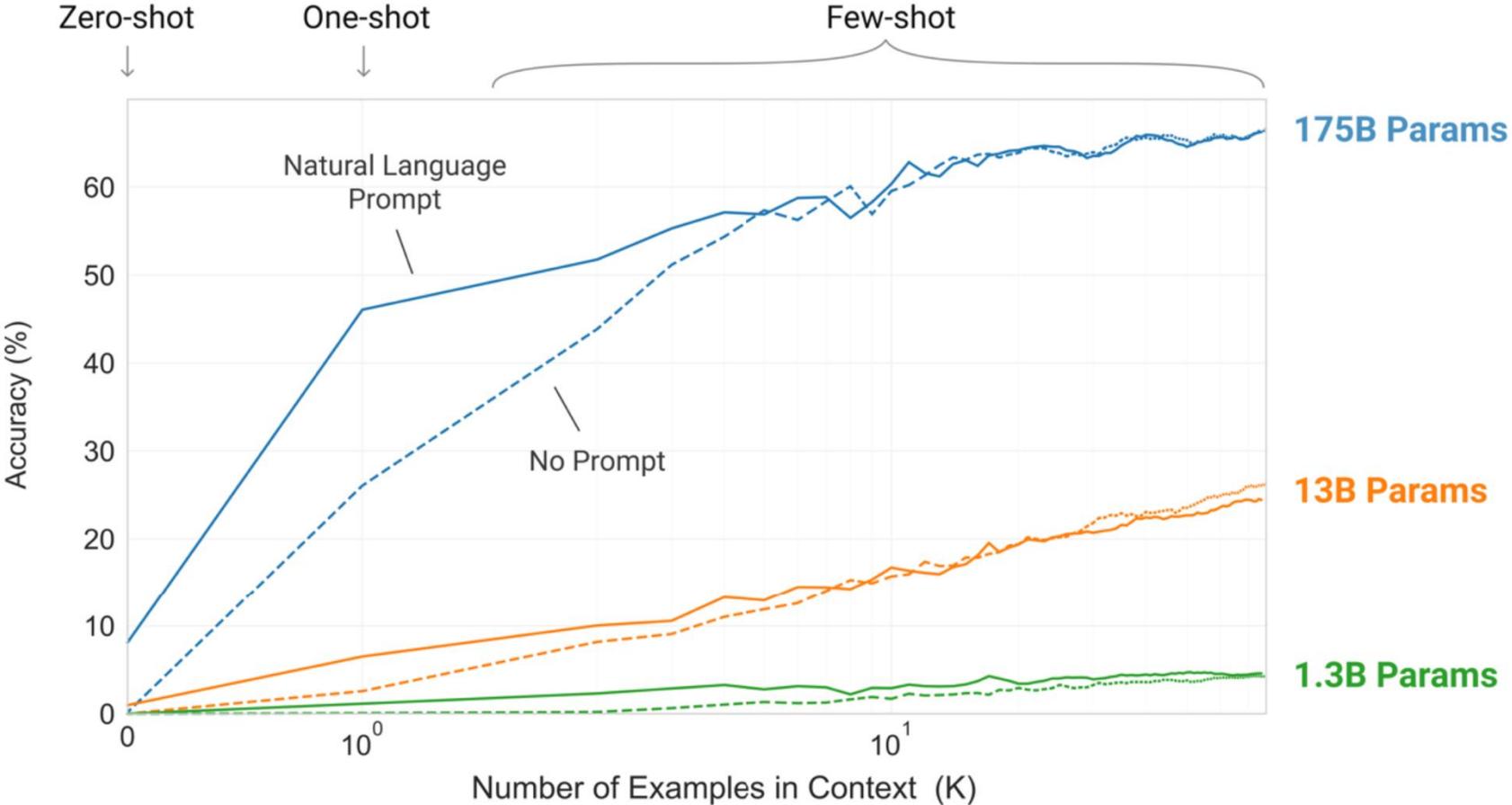
*Test input*



*Prediction*

Positive

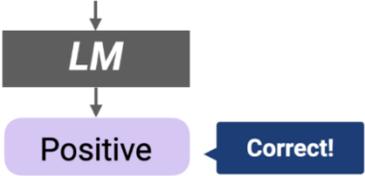
# Few-Shot Prompting / In-Context Learning



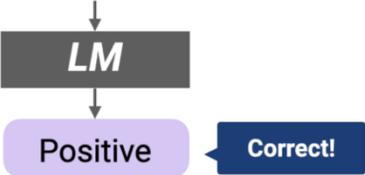
# Does In-Context Really Learn?

- Provide information more about format?
  - Give wrong in-context learning examples

Circulation revenue has increased by 5% in Finland. \n Positive  
Panostaja did not disclose the purchase price. \n Neutral  
Paying off the national debt will be extremely painful. \n Negative  
The company anticipated its operating profit to improve. \n \_\_\_\_\_

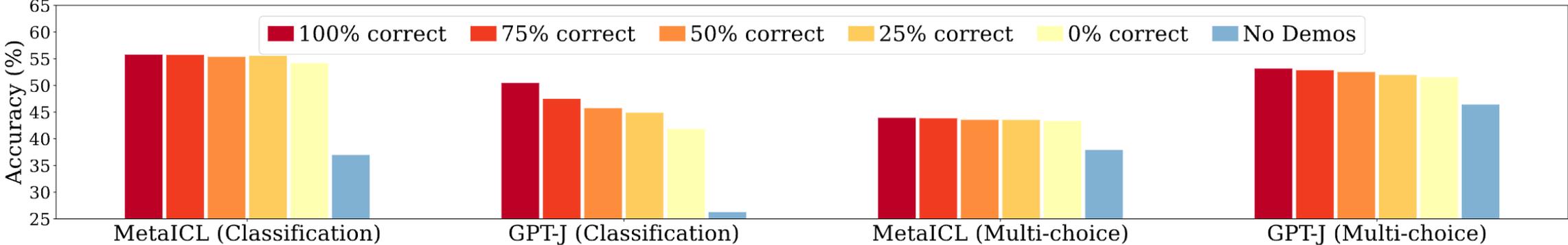


Circulation revenue has increased by 5% in Finland. \n **Neutral**  
Panostaja did not disclose the purchase price. \n **Negative**  
Paying off the national debt will be extremely painful. \n **Positive**  
The company anticipated its operating profit to improve. \n \_\_\_\_\_



# Does In-Context Really Learn?

- Provide information more about format?
  - Give wrong in-context learning examples

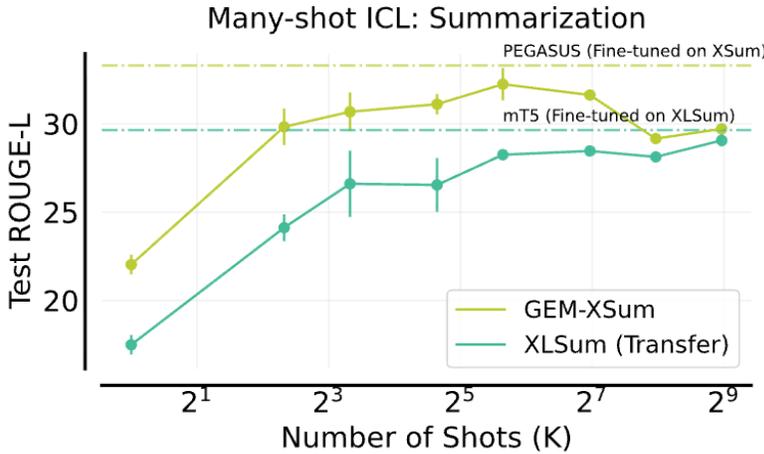
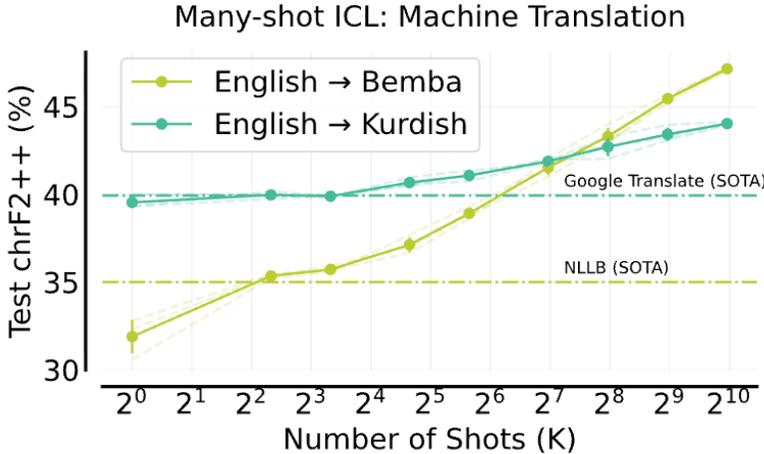


# Many-Shot In-Context Learning

## Many-Shot In-Context Learning

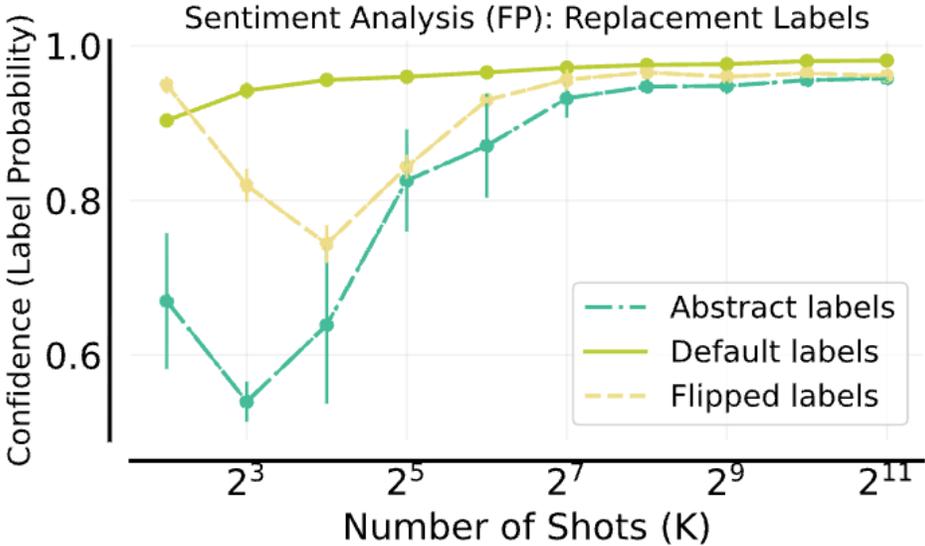
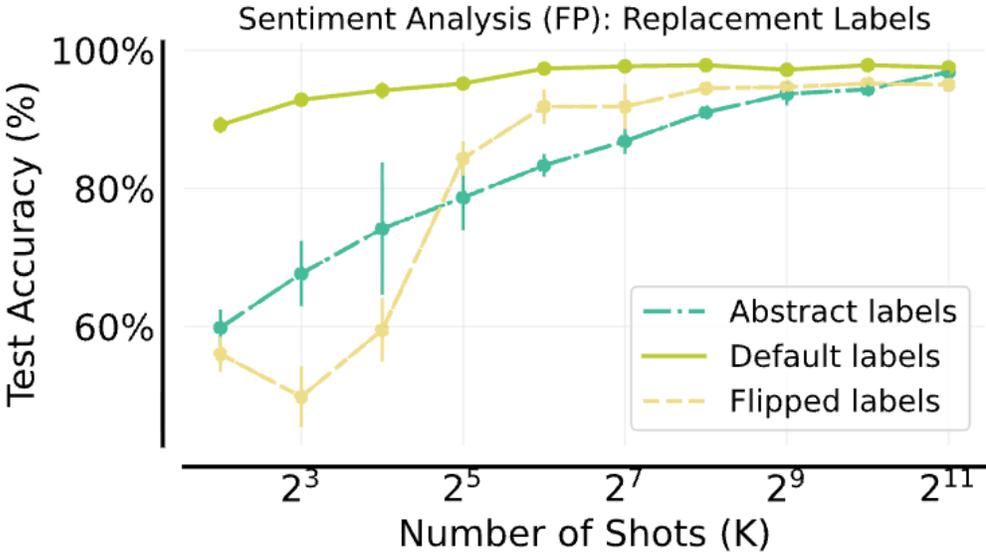
Rishabh Agarwal\*, Avi Singh\*, Lei M. Zhang†, Bernd Bohnet†, Luis Rosias†, Stephanie C.Y. Chan†, Biao Zhang†, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust and Hugo Larochelle

\*Contributed equally, †Key contribution



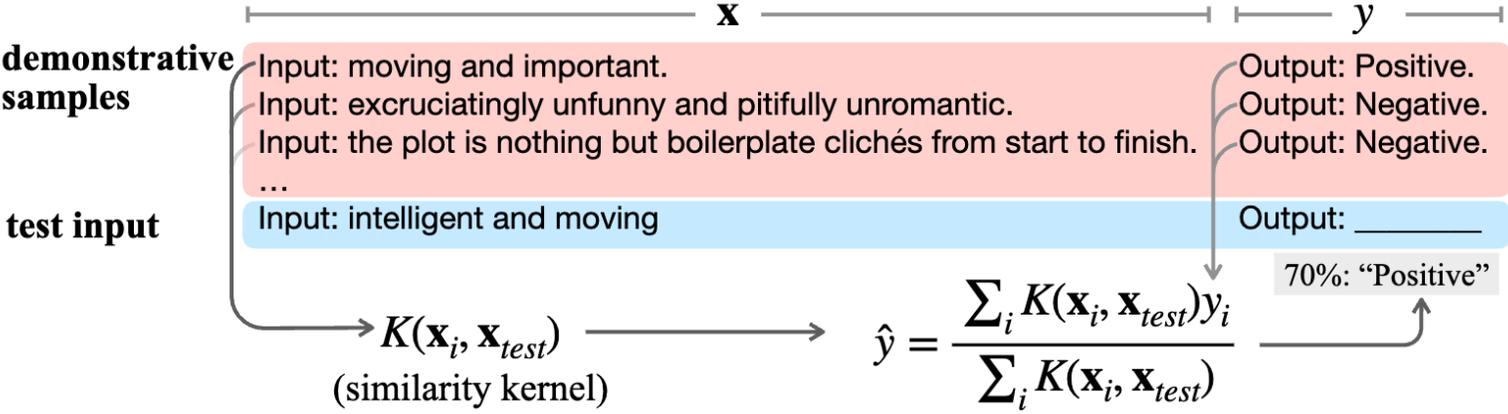
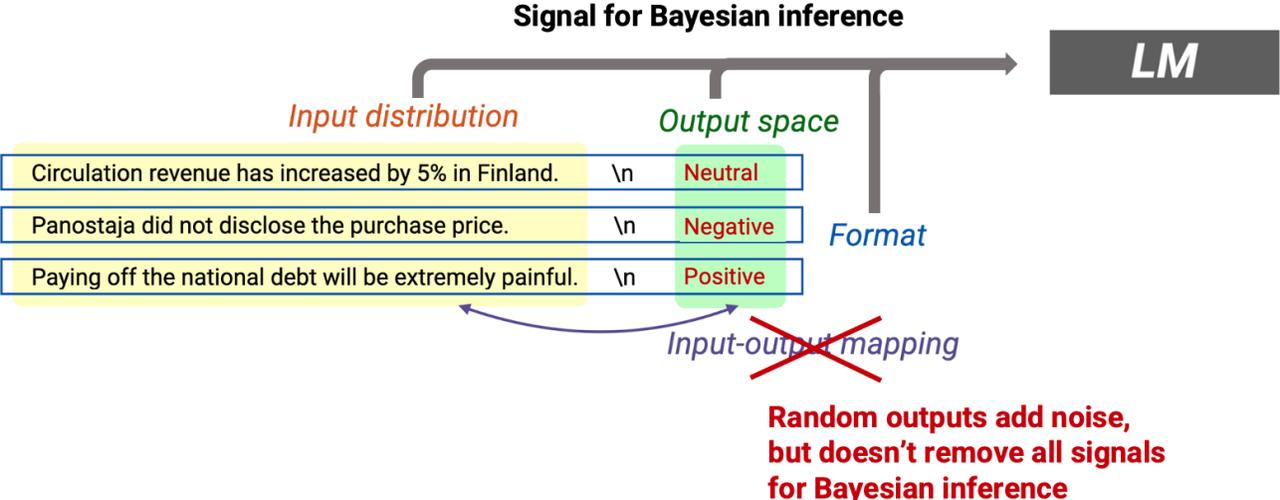
# Many-Shot In-Context Learning

- Overcoming Pre-training Biases with many-shot in-context learning



# What Makes In-Context Learning Work?

- Still an open research problem



# What Makes In-Context Learning Work?

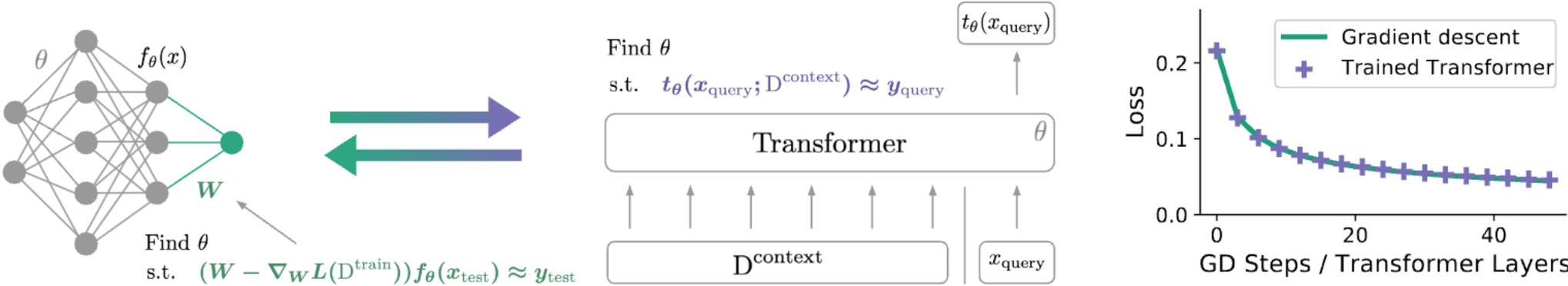
- Still an open research problem

---

## Transformers Learn In-Context by Gradient Descent

---

Johannes von Oswald<sup>1,2</sup> Eyvind Niklasson<sup>2</sup> Ettore Randazzo<sup>2</sup> João Sacramento<sup>1</sup>  
Alexander Mordvintsev<sup>2</sup> Andrey Zhmoginov<sup>2</sup> Max Vladymyrov<sup>2</sup>



# Lecture Plan

- Large Language Models (LLMs)
  - In-Context Learning
  - Chain-of-Thought
  - Popular LLMs

# Chain-of-Thought (CoT) Prompting

- Provide **reasoning chain** improves performance

### Standard Prompting

**Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Model Output**

A: The answer is 27. ❌

### Chain-of-Thought Prompting

**Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Model Output**

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅

# Chain-of-Thought (CoT) Prompting

### Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

### Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?  
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500.  $9 + 90(2) + 401(3) = 1392$ . The answer is (b).

### CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?  
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

### StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about  $0.6 \text{ g/cm}^3$ , which is less than water. Thus, a pear would float. So the answer is no.

### Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

### Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

### SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

### Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

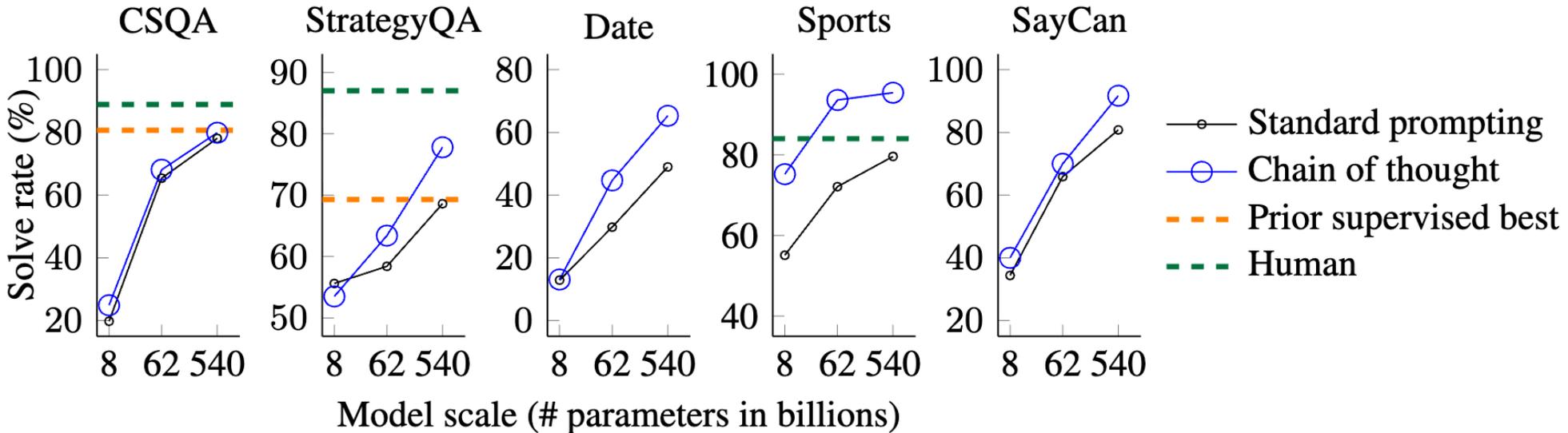
A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

### Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

# Chain-of-Thought (CoT) Prompting



# What Makes Chain-of-Thought Work?

- Explicit reasoning steps
  - Models can think
- Knowledge expansion
  - Model can retrieve and use internal knowledge
- Possibility to refine answers
  - Model can do self-correction

# Zero-Shot Chain-of-Thought Prompting

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

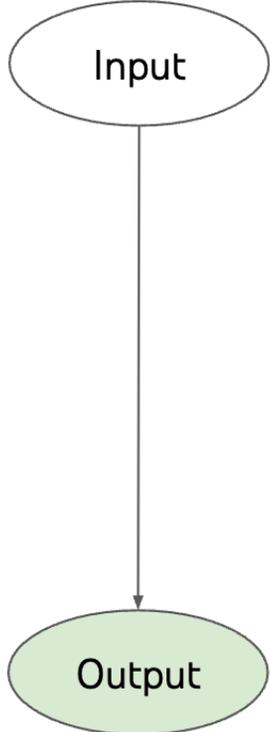
A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

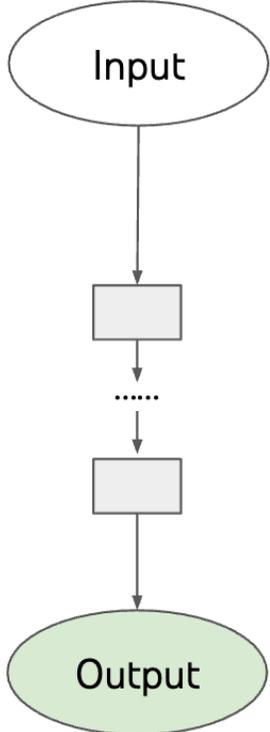
# Zero-Shot Chain-of-Thought Prompting

	MultiArith	GSM8K
<b>Zero-Shot</b>	<b>17.7</b>	<b>10.4</b>
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
<b>Zero-Shot-CoT</b>	<b>78.7</b>	<b>40.7</b>
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
<b>Zero-Plus-Few-Shot-CoT (8 samples) (*2)</b>	<b>92.8</b>	<b>51.5</b>
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
<b>PaLM 540B: Zero-Shot</b>	<b>25.5</b>	<b>12.5</b>
<b>PaLM 540B: Zero-Shot-CoT</b>	<b>66.1</b>	<b>43.0</b>
<b>PaLM 540B: Zero-Shot-CoT + self consistency</b>	<b>89.0</b>	<b>70.1</b>
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

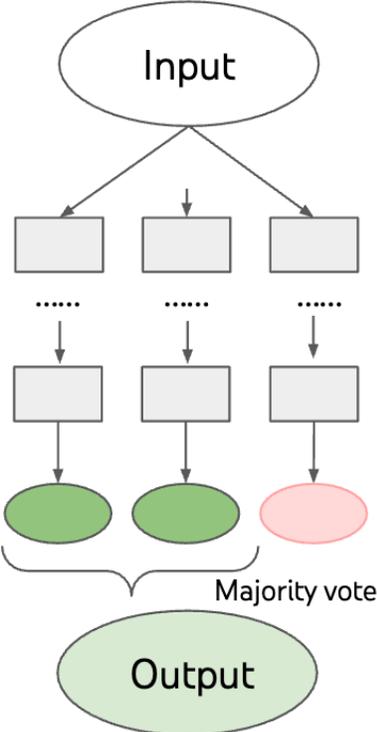
# Tree-of-Thoughts



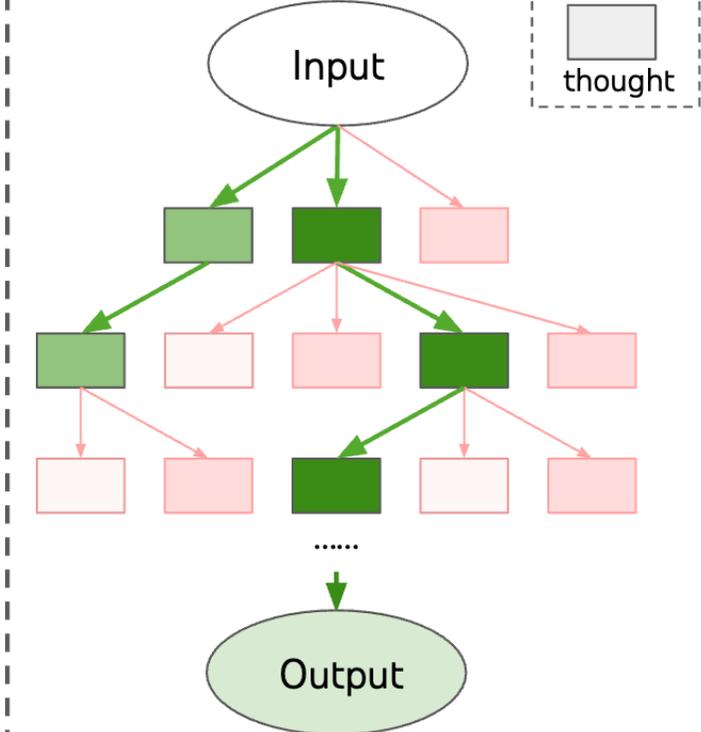
(a) Input-Output Prompting (IO)



(c) Chain of Thought Prompting (CoT)



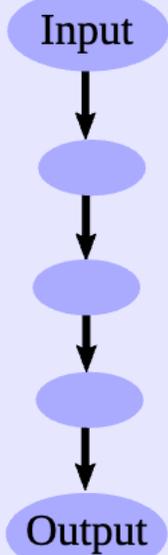
(c) Self Consistency with CoT (CoT-SC)



**(d) Tree of Thoughts (ToT)**

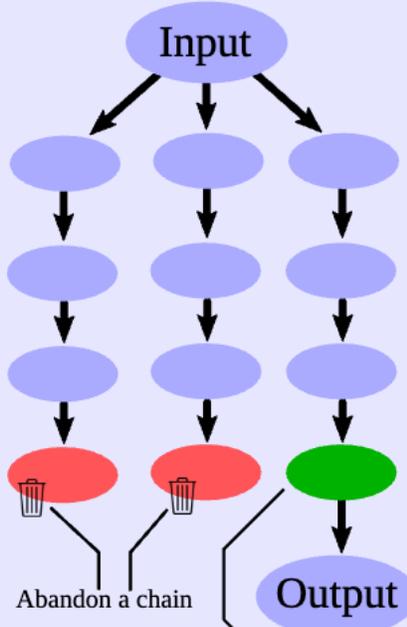
# Graph-of-Thoughts

## Chain-of-Thought (CoT)



**Key novelty:** Intermediate LLM thoughts within a chain

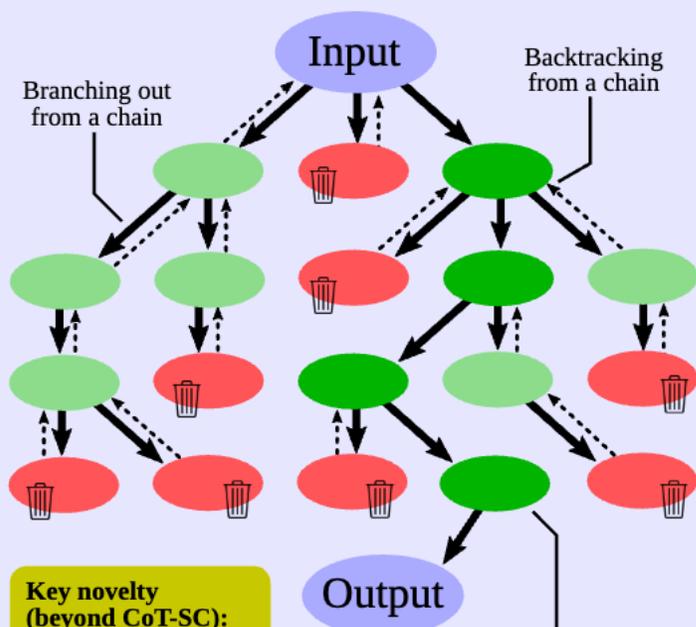
## Multiple CoTs (CoT-SC)



**Key novelty (beyond CoT):** Harnessing multiple independent chains of thoughts

Selecting a chain with the best score

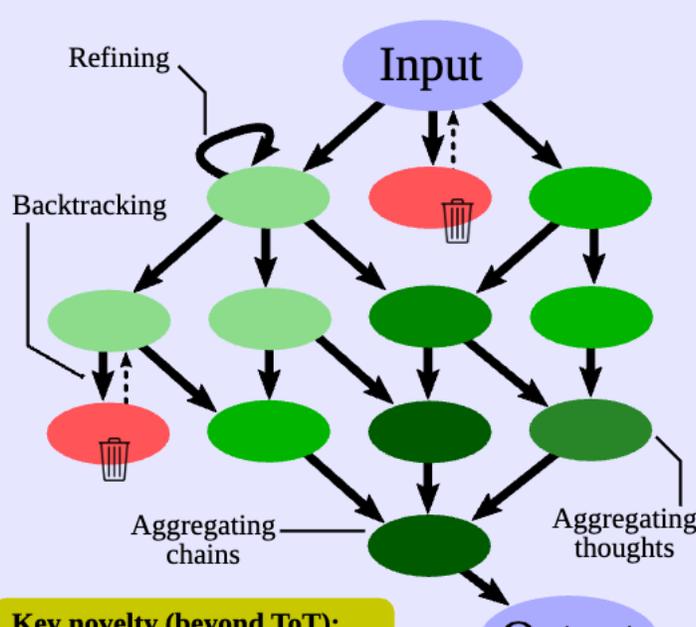
## Tree of Thoughts (ToT)



**Key novelty (beyond CoT-SC):** Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it

Intermediate thoughts are also scored

## Graph of Thoughts (GoT) [This work]



**Key novelty (beyond ToT):** Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

# Model Reasoning is The Key

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

*(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

How to encourage model to generate better reasoning chain?

# Reasoning Models

Question: If  $a > 1$ , then the sum of the real solutions of  $\sqrt{a - \sqrt{a+x}} = x$  is equal to

Response: <think>

To solve the equation  $\sqrt{a - \sqrt{a+x}} = x$ , let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a+x \implies a^2 - 2ax^2 + (x^2)^2 = a+x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

**Wait, wait. Wait. That's an aha moment I can flag here.**

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

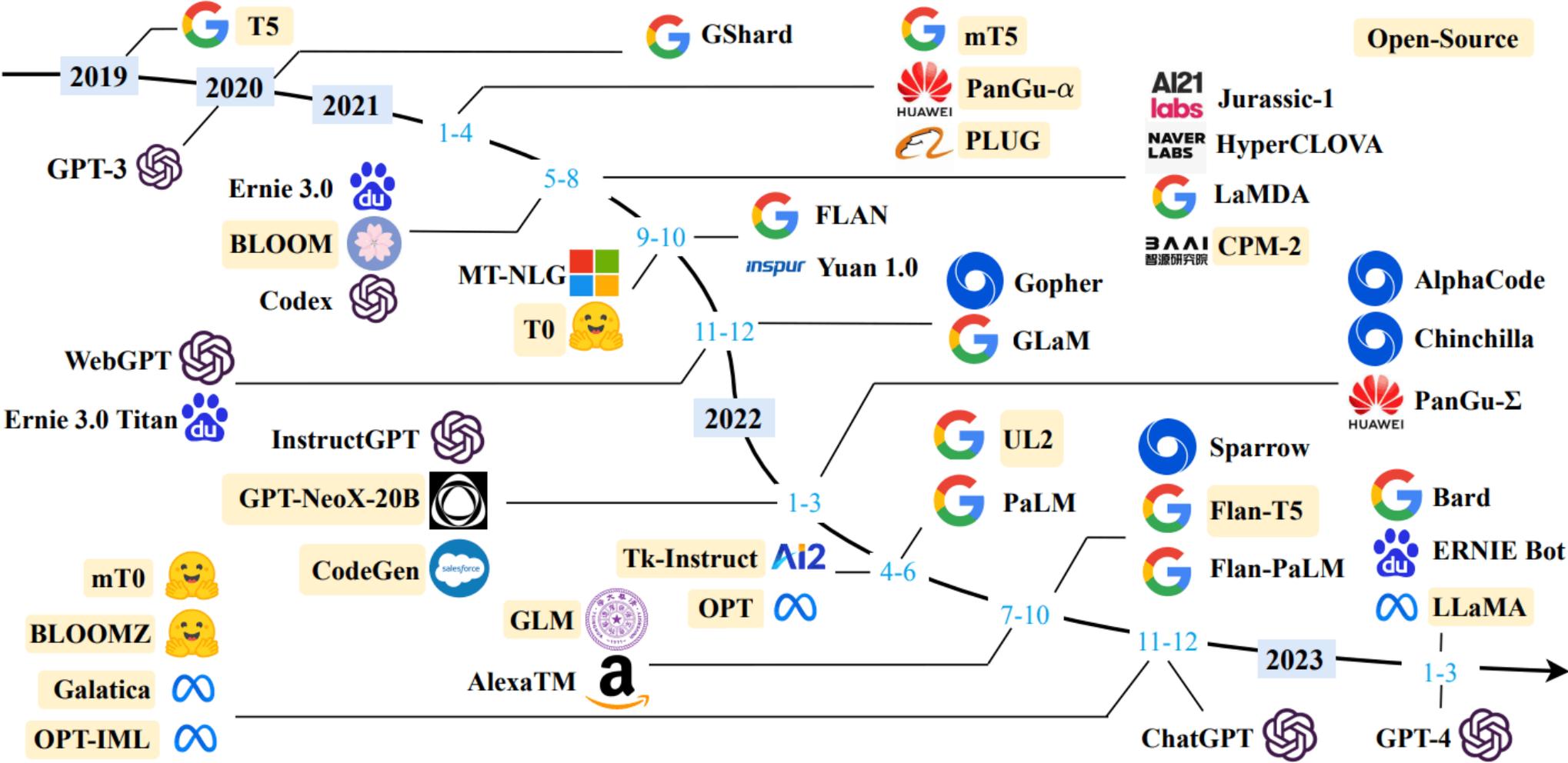
$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

GPT-5.2	GPT-4.1	o4-mini
The best model for coding and agentic tasks across industries	Smartest non-reasoning model	Fast, cost-efficient reasoning model, succeeded by GPT-5 mini
<a href="#">Learn more</a>	<a href="#">Learn more</a>	<a href="#">Learn more</a>
<a href="#">Playground</a>	<a href="#">Playground</a>	<a href="#">Playground</a>
Reasoning: ●●●●●	Intelligence: ●●●●●	Reasoning: ●●●●●
Speed: ⚡⚡⚡	Speed: ⚡⚡⚡	Speed: ⚡⚡⚡
Input: [🗣️] [📄] [📁] [📧]	Input: [🗣️] [📄] [📁] [📧]	Input: [🗣️] [📄] [📁] [📧]
Output: [🗣️] [📄] [📁] [📧]	Output: [🗣️] [📄] [📁] [📧]	Output: [🗣️] [📄] [📁] [📧]
Reasoning tokens: ✓	Reasoning tokens: ☒	Reasoning tokens: ✓

# Large Language Models



# Proprietary LLMs: GPT Series



<https://openai.com/>

GPT-5.2	GPT-4.1	o4-mini
The best model for coding and agentic tasks across industries	Smartest non-reasoning model	Fast, cost-efficient reasoning model, succeeded by GPT-5 mini
<a href="#">Learn more</a>	<a href="#">Learn more</a>	<a href="#">Learn more</a>
<a href="#">Playground</a>	<a href="#">Playground</a>	<a href="#">Playground</a>
Reasoning: 5/5	Intelligence: 5/5	Reasoning: 4/5
Speed: 4/4	Speed: 4/4	Speed: 4/4
Input: 4/4	Input: 4/4	Input: 4/4
Output: 4/4	Output: 4/4	Output: 4/4
Reasoning tokens: ✓	Reasoning tokens: ✗	Reasoning tokens: ✓

# Proprietary LLMs: Gemini Series



<https://gemini.google.com/>

OUR MOST INTELLIGENT MODEL

## Gemini 3 Pro

The best model in the world for multimodal understanding, and our most powerful agentic and vibe-coding model yet, delivering richer visuals and deeper interactivity, all built on a foundation of state-of-the-art reasoning.

[+ Expand to learn more](#)

OUR MOST BALANCED MODEL, DESIGNED TO SCALE

## Gemini 3 Flash

Our most balanced model built for speed, scale, and frontier intelligence.

[+ Expand to learn more](#)

# Proprietary LLMs: Claude Series



<https://claude.ai/>

## Claude Sonnet 4.6

The image is a promotional banner for Claude Sonnet 4.6. It features a light beige background. At the top left is the orange starburst icon. The main title "Claude Sonnet 4.6" is in a large, black, serif font. Below the title is a subtitle in a smaller, black, sans-serif font. At the bottom are two buttons: a dark grey "Try Claude" button and a white "Get API access" button with a black border. Below the buttons is a section titled "Announcements" in a bold, black, sans-serif font. Underneath is a "NEW" badge in a grey box, followed by the text "Claude Sonnet 4.6 Feb 17, 2026". The main body of text describes the model's capabilities, with the word "coding" highlighted by a red box.

Hybrid reasoning model with superior intelligence  
for agents, featuring a 1M context window

Try Claude

Get API access

### Announcements

**NEW** Claude Sonnet 4.6 Feb 17, 2026

Sonnet 4.6 delivers frontier performance across **coding**, agents, and professional work at scale. It can compress multi-day coding projects into hours and deliver production-ready solutions.

# Proprietary LLMs: Grok Series



<https://grok.com/>

Model	Modalities ⓘ	Capabilities ⓘ	Context ⓘ
Language models			
grok-4-1-fast-reasoning	T 🗒️ → T 🗒️	📄 {} ⚡	2,000,000
grok-4-1-fast-non-reasoning	T 🗒️ → T 🗒️	📄 {} 🗑️	2,000,000
grok-code-fast-1	T 🗒️ → T 🗒️	📄 {} ⚡	256,000
grok-4-fast-reasoning	T 🗒️ → T 🗒️	📄 {} ⚡	2,000,000
grok-4-fast-non-reasoning	T 🗒️ → T 🗒️	📄 {} 🗑️	2,000,000
grok-4-0709	T 🗒️ → T 🗒️	📄 {} ⚡	256,000
grok-3-mini	T 🗒️ → T 🗒️	📄 {} ⚡	131,072
grok-3	T 🗒️ → T 🗒️	📄 {} 🗑️	131,072
grok-2-vision-1212	T 🗒️ → T 🗒️	📄 {} 🗑️	32,768

# Proprietary LLMs: Grok Series



Amazon Nova

<https://nova.amazon.com/>

Service

**Nova Act**  
Nova Act is a new AWS service to build, deploy, and manage fleets of AI agents for automating production UI workflows. Nova Act provides high...

→ +

Service

**Nova Forge**  
Nova Forge is a new service to build your own frontier models using Nova. Organizations of all sizes can now experience the benefits of a purpos...

→ +

Models

**Nova Models**  
Amazon Nova models deliver frontier intelligence and industry-leading price-performance. With the most comprehensive suite of customization...

→ +

# Open-Weight LLMs: Llama Series



<https://llama.com/>

### Llama 4 models

Multi-image Text

#### Llama 4 Scout

- Superior text and visual intelligence
- Class-leading 10M context window
- **17B active params x 16 experts, 109B total params**
- Llama Guard 4 12B is included
- Llama Prompt Guard 2 22M and Llama Prompt Guard 2 86M are included

\*Licensed under Llama 4 Community License Agreement

Multi-image Text

#### Llama 4 Maverick

- Our most powerful open source multimodal model
- Industry-leading intelligence and fast responses at a low cost
- **17B active params x 128 experts, 400B total params**
- Llama Guard 4 12B is included
- Llama Prompt Guard 2 22M and Llama Prompt Guard 2 86M are included

\*Licensed under Llama 4 Community License Agreement

### Llama 3 models

Text

#### Llama 3.3: 70B

- Multilingual open source large language model
- Experience 405B performance and quality at a fraction of the cost

\*Licensed under Llama 3.3 Community License Agreement

Lightweight

#### Llama 3.2: 1B & 3B

- Lightweight and most cost-efficient models you can run anywhere on mobile and on edge devices
- Llama Guard 3 1B is included
- Quantized models available

\*Licensed under Llama 3.2 Community License Agreement

Text

#### Llama 3.1: 405B & 8B

- Multilingual open source large language model
- Llama Guard 3 8B and Llama Prompt Guard 2 are included

\*Licensed under Llama 3.1 Community License Agreement

Multimodal

#### Llama 3.2: 11B & 90B

- Open multimodal models that are flexible and can reason on high resolution images and output text
- Llama Guard 3 11B Vision is included

\*Licensed under Llama 3.2 Community License Agreement

Widely used in academic research  
Multilingual and multimodal support

# Open-Weight LLMs: Qwen Series



<https://qwen.ai/>

**Qwen3.5: Towards Native Multimodal Agents**

**Qwen3.5: Towards Native Multimodal Agents**

We are delighted to announce the official release of Qwen3.5, introducing the open-weight of the first model in the Qwen3.5 series, namely Qwen3.5-397B-A17B. As a native vision-language model,...

Open-Source | 2026/02/15

**Qwen-Image-2.0: Professional infographics, exquisite photorealism**

**Qwen-Image-2.0: Professional infographics, exquisite p...**

We are launching Qwen-Image-2.0, a next-generation foundational image generation model. The key highlights of Qwen-Image-2.0 include: Professional Typography Rendering: Supports 1k-token instructions for...

Open-Source | 2026/02/09

**Qwen3-Coder-Next: Pushing Small Hybrid Models on Agentic Coding**

**Qwen3-Coder-Next: Pushing Small Hybrid Models on A...**

--- We introduce Qwen3-Coder-Next, an open-weight language model designed specifically for coding agents and local development. Built on top of Qwen3-Next-80B-A3B-Base, which adopts a novel architecture...

Open-Source | 2026/02/02

**Qwen3-ASR & Qwen3-ForcedAligner is Now Open Sourced: Robust, Streaming and Multilingual!**

**Qwen3-ASR & Qwen3-ForcedAligner is Now Open Sour...**

Qwen3-ASR family includes two powerful all-in-one speech recognition models and a novel non-autoregressive speech forced alignment model. Qwen3-ASR-1.7B and Qwen3-ASR-0.6B are ASR models that support...

Open-Source | 2026/01/28

Widely used in academic research  
Multilingual and multimodal support

My personal take: Probably the strongest  
open-weight LLM right not

# Open-Weight LLMs: Gemma Series



<https://deepmind.google/models/gemma/>



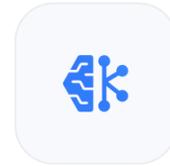
## Gemma 3

Multimodal capabilities and wide language support in sizes from 270M to 27B parameters



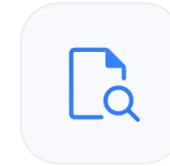
## Gemma 3n

Mobile-first architecture optimized for low-latency audio and visual understanding



## FunctionGemma

Specialized for fast, private agentic workflows



## MedGemma

Optimized for medical text and image comprehension

# Open-Weight LLMs: DeepSeek Series



<https://deepseek.com/>

---

Question: If  $a > 1$ , then the sum of the real solutions of  $\sqrt{a - \sqrt{a+x}} = x$  is equal to

---

Response: <think>

To solve the equation  $\sqrt{a - \sqrt{a+x}} = x$ , let's start by squaring both ...

$$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

**Wait, wait. Wait. That's an aha moment I can flag here.**

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

---

Strong reasoning capability

# Open-Weight LLMs: Mistral Series



<https://mistral.ai/>

## Mistral Large 3

One of the best OSS models in the world: open-weight, general-purpose, flagship multimodal and multilingual model.

→ Mistral Large 3 is our largest model to date, featuring 41B active parameters and 675B total parameters, with a large 256k context window, and offers powerful agentic capabilities.

[Learn more](#) >

## Ministral Family

3B, 8B, and 14B brings best-in-class frontier AI to the edge.

→ Combining compact efficiency with multimodal and multilingual capability. Engineered for edge devices, self-hosted systems, and robotics, these models seamlessly blend language, vision, and reasoning into highly efficient architectures

[Learn more](#) >

# Open-Weight LLMs: Kimi Series



<http://kimi.com/>

15	11 -> 22	<b>AI</b> <code>claude-sonnet-4-5-20250929</code> Anthropic · Proprietary	1450 ±4	44,667
16	12 -> 22	<code>gemini-2.5-pro</code> Google · Proprietary	1449 ±3	95,644
17	11 -> 26	<code>ernie-5.0-preview-1203</code> Baidu · Proprietary	1449 ±7	9,743
18	12 -> 24	<b>AI</b> <code>claude-opus-4-1-20250805-thinking-16k</code> Anthropic · Proprietary	1449 ±4	49,820
19	11 -> 27	<code>kimi-k2.5-thinking</code> Moonshot · Modified MIT	1448 ±7	9,156
20	12 -> 26	<b>AI</b> <code>claude-opus-4-1-20250805</code> Anthropic · Proprietary	1445 ±3	75,907
21	12 -> 29	<code>gpt-4.5-preview-2025-02-27</code> OpenAI · Proprietary	1444 ±6	14,549
22	17 -> 27	<code>chatgpt-4o-latest-20250326</code> OpenAI · Proprietary	1442 ±3	83,252

# Other Open-Weight Models

- Code

- StarCoder 2



- Code Llama  Meta

- DeepSeek-Coder-V2  deepseek

- Math

- LLeMa



- DeepSeek Math  deepseek

- Science

- Galactica  Meta

# Open-Source LLMs: OLMo Series



<https://allenai.org/olmo>

## Olmo 3

Our latest and greatest open language model provides the most complete model pipeline ever released, allowing for infinite customizability and contextual understanding of how to create, modify, and deploy LLMs.

Visit the Olmo page



Fully open-source models, including training data,  
training pipeline, evaluation framework

# Arena: LLM Battleground

🗙 Battle Mode ▾

Popular LLMs?

Assistant A



## Popular Large Language Models (LLMs)

### OpenAI

- GPT-4 / GPT-4o
- GPT-3.5 (ChatGPT)
- o1, o3 (reasoning models)

### Google

- Gemini (Ultra, Pro, Flash)
- PaLM 2

### Anthropic

- Claude 3.5 (Sonnet, Haiku, Opus)



Assistant B



Popular LLMs (as of 2024–2025), grouped by ecosystem:

### Commercial / hosted (closed-weight)

- **OpenAI:** GPT-4.1 / GPT-4o, o1 (reasoning)
- **Anthropic:** Claude 3.5 (Sonnet/Opus)
- **Google:** Gemini 1.5 / 2.x (varies by release)
- **Cohere:** Command R / R+
- **AI21:** Jamba (some variants)

### Open-weights (you can self-host)

- **Meta:** Llama 3 / 3.1 (very widely used)
- **Mistral AI:** Mistral Large (hosted), **Mixtral 8×7B / 8×22B**, Mistral 7B
- **Alibaba:** Qwen2 / Qwen2.5 (strong multilingual + coding variants)
- **Google:** Gemma 2
- **Microsoft:** Phi-3 / Phi-4 (small, efficient)



# Arena: LLM Battleground

Text <span>🕒 1 day ago</span>				Code <span>🕒 1 day ago</span>			
Rank ↕	Model ↕	Score ↓	Votes ↕	Rank ↕	Model ↕	Score ↓	Votes ↕
1	claude-opus-4-6-thinking	1506	4,745	1	claude-opus-4-6	1561	2,364
2	claude-opus-4-6	1503	5,540	2	claude-opus-4-6-thinking	1554	1,789
3	gemini-3-pro	1486	36,354	3	claude-opus-4-5-20251101-...	1501	10,354
4	grok-4.1-thinking	1474	35,833	4	gpt-5.2-high	1471	1,693
5	gemini-3-flash	1473	27,116	5	claude-opus-4-5-20251101	1469	10,491
6	dola-seed-2.0-preview <span>📄</span>	1472	3,164	6	glm-5	1458	2,048
7	claude-opus-4-5-20251101-...	1471	28,493	7	minimax-m2.5	1453	2,604
8	claude-opus-4-5-20251101	1467	33,328	8	gemini-3-pro	1444	16,457
9	grok-4.1	1463	39,912	9	kimi-k2.5-thinking	1442	3,342
10	gemini-3-flash (thinking-...	1461	18,480	10	gemini-3-flash	1441	12,079
<a href="#">View all</a>				<a href="#">View all</a>			