

# Domain Specialization of Large Language Models

- The Key to Make Large Language Models Disruptive

**Chen Ling**

PhD Candidate

Emory University



**EMORY**  
UNIVERSITY

---

**Department of  
Computer Science**

# Collaborations

Computer Science > Computation and Language

[Submitted on 30 May 2023 (v1), last revised 29 Mar 2024 (this version, v7)]

## Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, Liang Zhao

Large language models (LLMs) have significantly advanced the field of natural language processing (NLP), providing a highly useful, task-agnostic foundation for a wide range of applications. However, directly applying LLMs to solve sophisticated problems in specific domains meets many hurdles, caused by the heterogeneity of domain data, the sophistication of domain knowledge, the uniqueness of domain objectives, and the diversity of the constraints (e.g., various social norms, cultural conformity, religious beliefs, and ethical standards in the domain applications). Domain specification techniques are key to make large language models disruptive in many applications. Specifically, to solve these hurdles, there has been a notable increase in research and practices conducted in recent years on the domain specialization of LLMs. This emerging field of study, with its substantial potential for impact, necessitates a comprehensive and systematic review to better summarize and guide ongoing work in this area. In this article, we present a comprehensive survey on domain specification techniques for large language models, an emerging direction critical for large language model applications. First, we propose a systematic taxonomy that categorizes the LLM domain-specialization techniques based on the accessibility to LLMs and summarizes the framework for all the subcategories as well as their relations and differences to each other. Second, we present an extensive taxonomy of critical application domains that can benefit dramatically from specialized LLMs, discussing their practical significance and open challenges. Last, we offer our insights into the current research status and future trends in this area.

This work is also cited by *2024 Economic Report of the President*

Collaborators from:

- BlackRock Aladdin Financial Engineering IAI Team
- NEC Labs America
- Microsoft Inc.
- Duke University, UCLA, Emory University



# Outline



## Background

1. History of Language Modeling
2. Cons of general LLMs
3. Domain-specific LLM



## Methods

1. External Augmentation
2. Prompt Crafting
3. Model Fine-tuning



## Conclusion and Thoughts

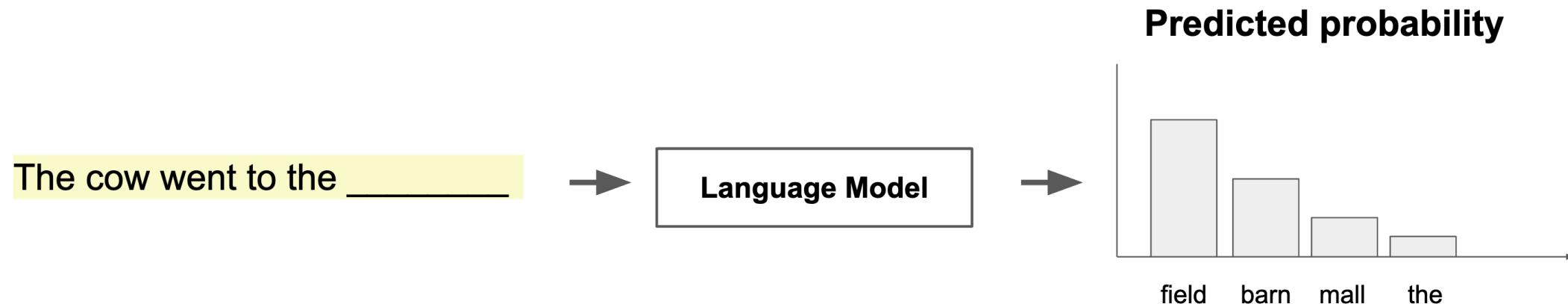
# Why do we need Domain-specific Large Language Models?

---

**Background**

# What is Language Modeling?

**Given** a context of words, **produce** the probability distribution of the next word.

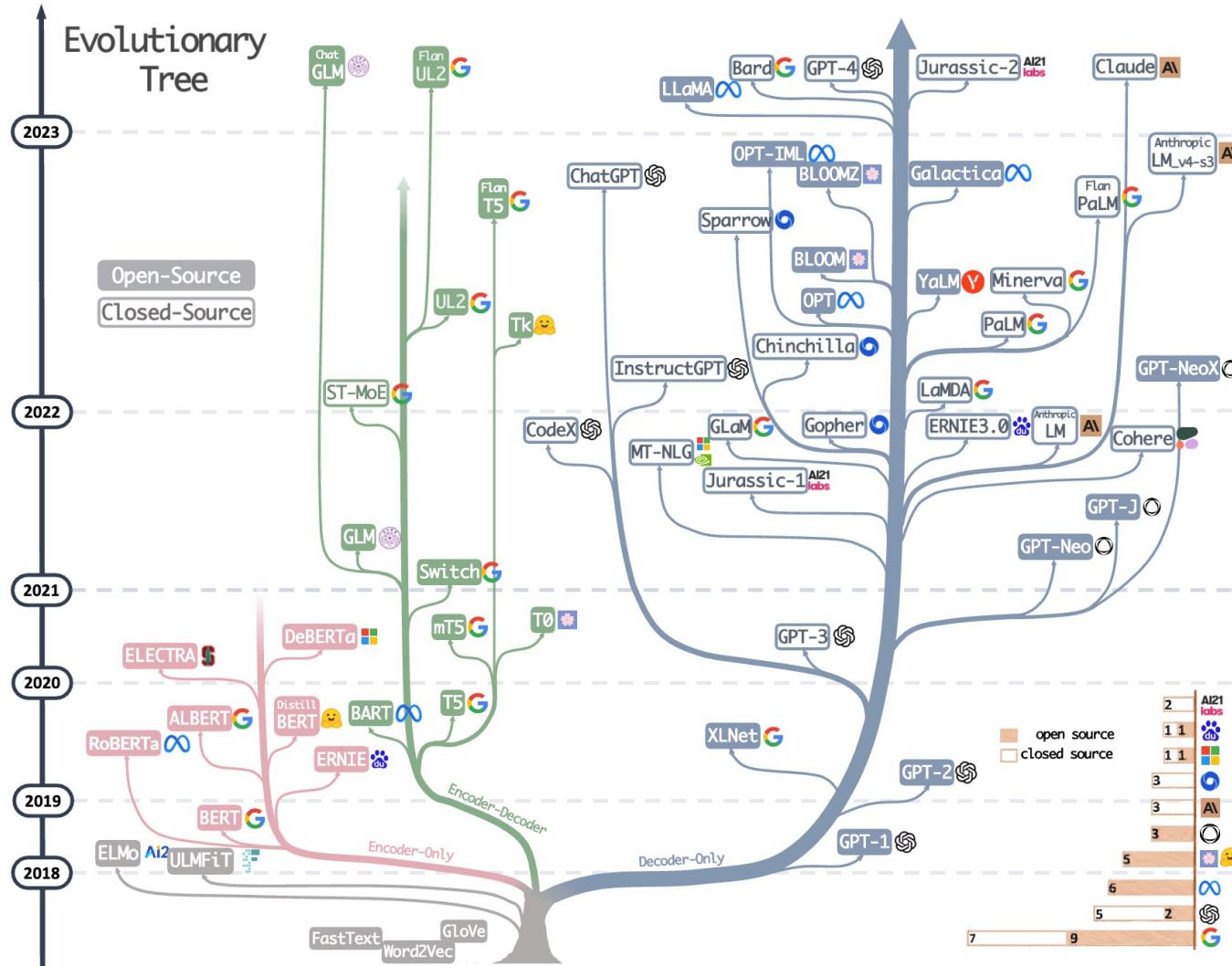


The model is trained to accomplish this task on **massive amount of text**.

To use it to generate text, you can repeatedly:

send in the context → calculate next-word probability → sample one from the distribution

# The development of language models in recent years



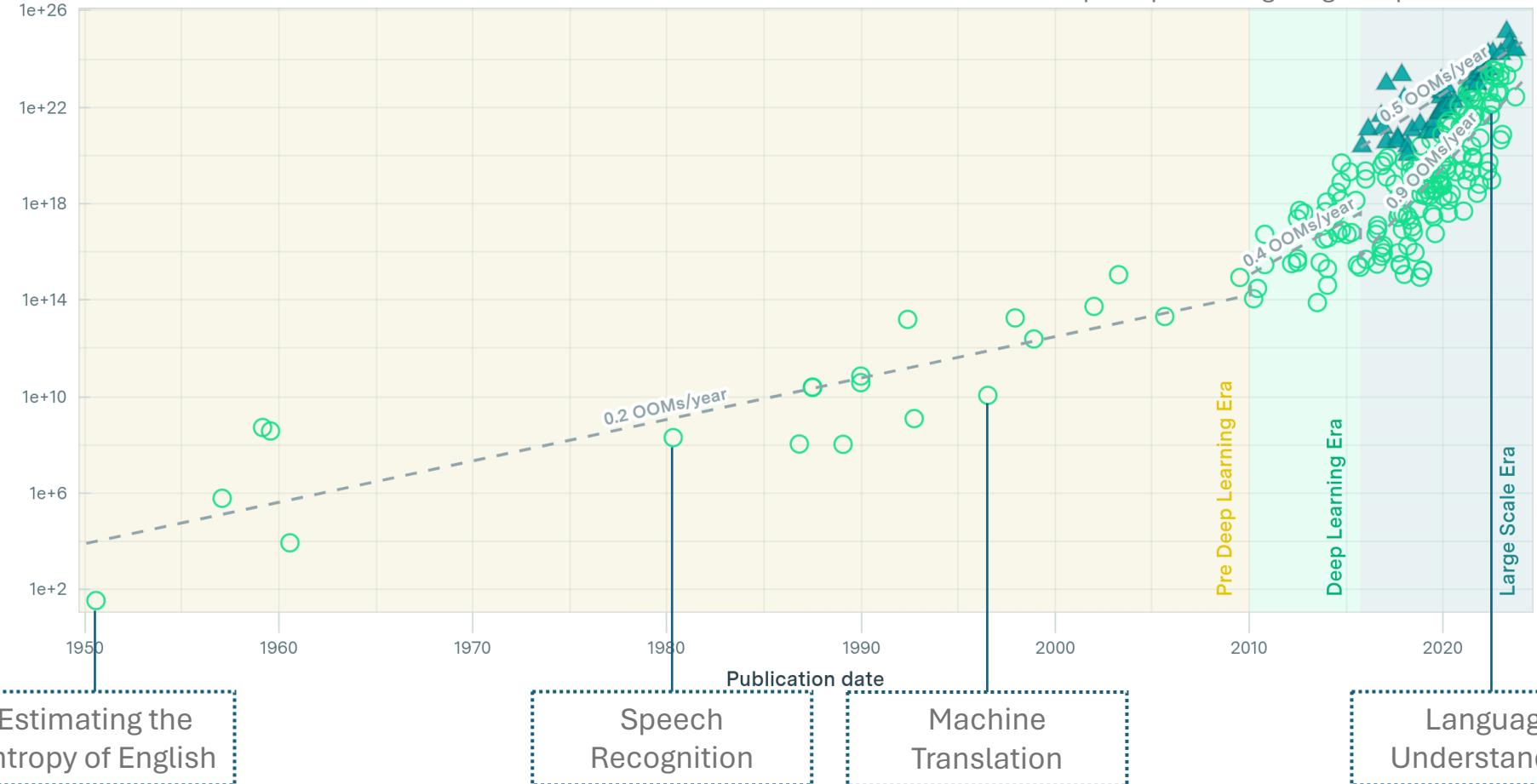
# Scale has driven the acceleration of AI developments

Training Compute of Notable Machine Learning Systems Over Time

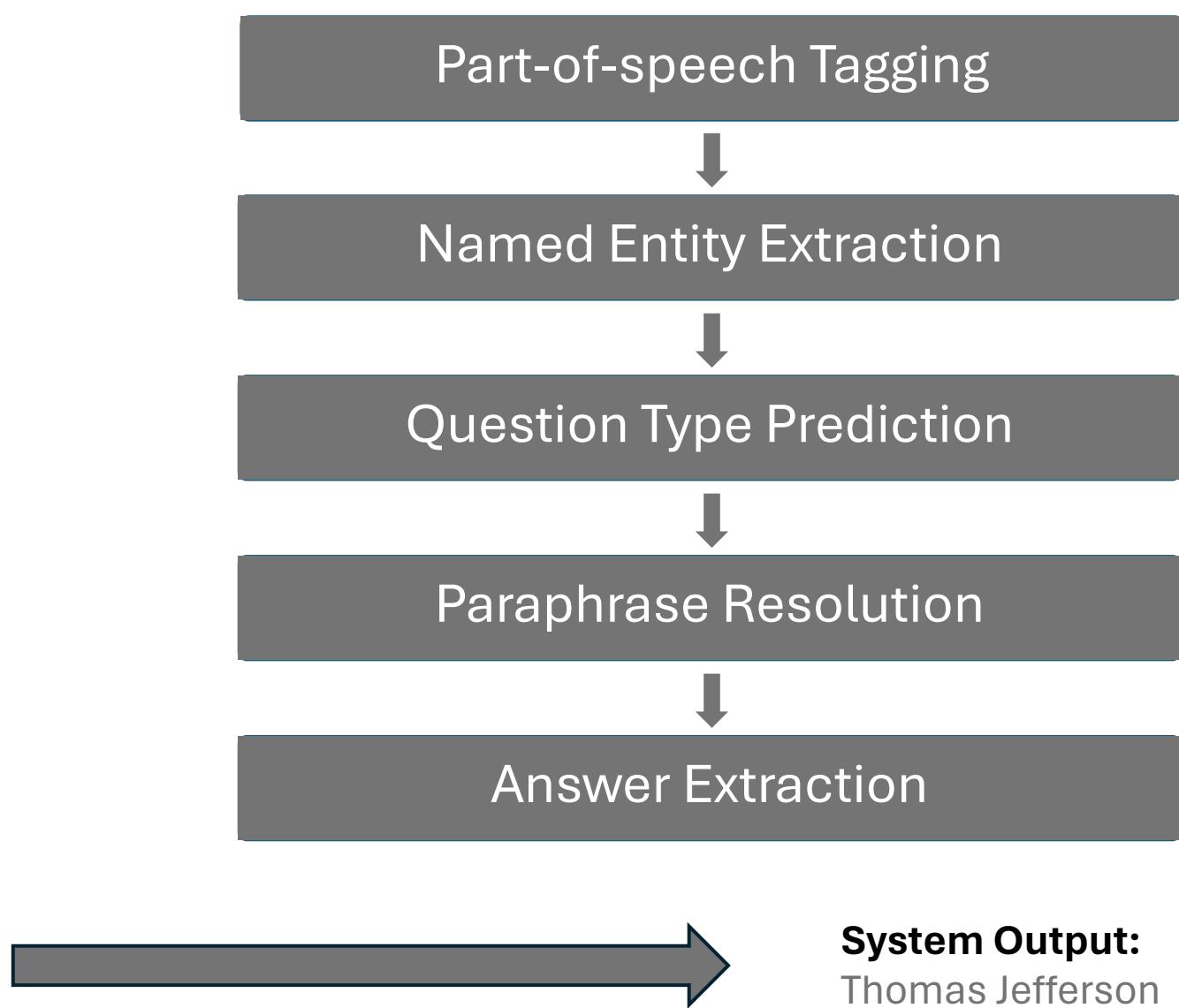
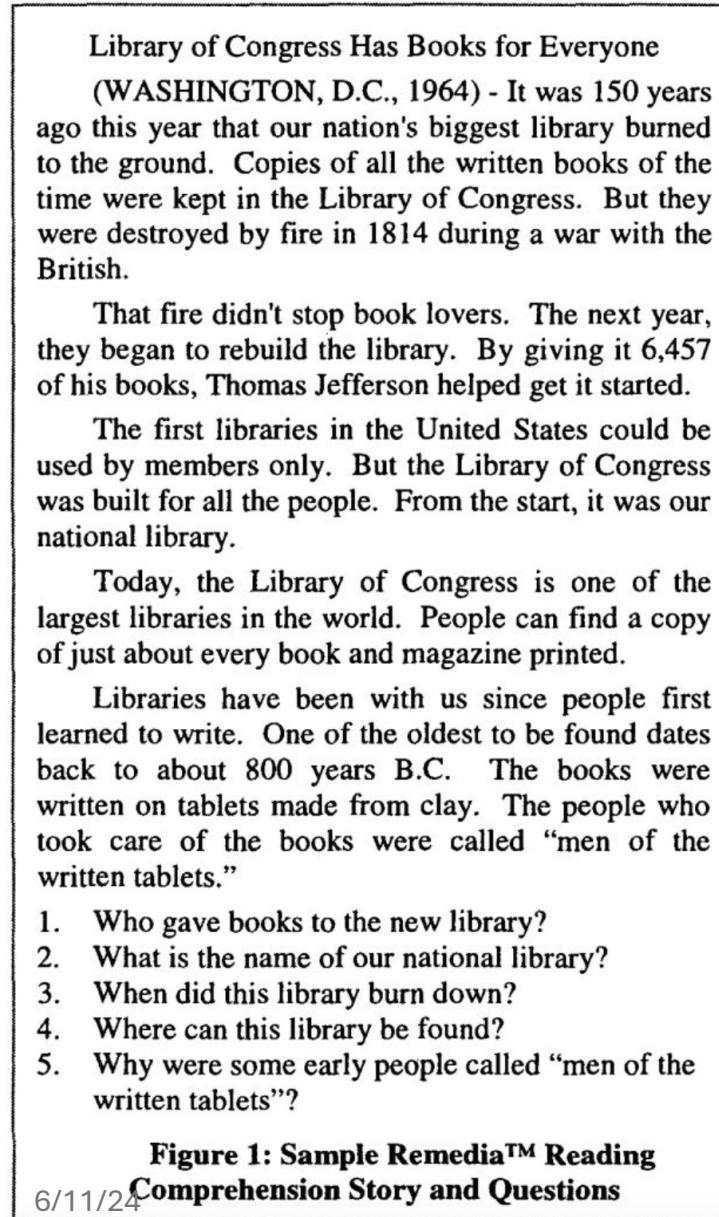
≡ EPOCH

Training compute (FLOP)

<https://epochai.org/blog/compute-trends>



# State-of-the-art in 2000s: Multiple Separate Models



### Library of Congress Has Books for Everyone

(WASHINGTON, D.C., 1964) - It was 150 years ago this year that our nation's biggest library burned to the ground. Copies of all the written books of the time were kept in the Library of Congress. But they were destroyed by fire in 1814 during a war with the British.

That fire didn't stop book lovers. The next year, they began to rebuild the library. By giving it 6,457 of his books, Thomas Jefferson helped get it started.

The first libraries in the United States could be used by members only. But the Library of Congress was built for all the people. From the start, it was our national library.

Today, the Library of Congress is one of the largest libraries in the world. People can find a copy of just about every book and magazine printed.

Libraries have been with us since people first learned to write. One of the oldest to be found dates back to about 800 years B.C. The books were written on tablets made from clay. The people who took care of the books were called "men of the written tablets."

1. Who gave books to the new library?

Mask

## AI Foundation Models

**OpenAI:** GPT-3 (2020), ChatGPT (2022)

**META:** LLaMA Models (2023)

**Google:** GEMINI (2023)

...

**Large Language Models - LLMs**  
**(Only One Module)**

### System Output:

2. The library was built for \_\_\_\_\_.
  3. One of the oldest libraries was found \_\_\_\_\_.
  4. The people who took care of the books were called \_\_\_\_\_.
- Answers: 1. Thomas Jefferson 2. all the people 3. 800 years B.C. 4. "men of the written tablets."

# LLMs are:

---

- **General:** A single model can handle many tasks without additional training
- **Broad:** It has as much (or more?) “knowledge” of the world as an adult
- **Accessible:** You interact with the model directly in language, not in code



# Some tasks where LLMs are effective:

- **Text Summarization**
- **Information Extraction:** Reading Comprehension, Question Answering
- **Document Understanding:** Sentiment Analysis, Insight Extraction
- **Text Generation:** Poetry Writing, Style Transfer
- **Code Generation:** Copilot (Auto-complete), Toolformer

# Some tasks where LLMs are effective:

- **Text Summarization**
- **Information Extraction:** Reading Comprehension, Question Answering
- **Document Understanding:** Sentiment Analysis, Insight Extraction
- **Text Generation:** Poetry Writing, Style Transfer
- **Code Generation:** Copilot (Auto-complete), Toolformer



However, can LLMs still accomplish  
these tasks in every domain?

# Why build a domain-specific LLM?

- Foundational models cannot understand the specific context beyond the massive datasets it was trained with.

c You

Can you give me three different references of submodular maximization using the multilinear extension?



ChatGPT

Certainly! Here are three references on submodular maximization using the multilinear extension:

1. Krause, Andreas, et al. "Near-optimal nonmyopic value of information in graphical models."

*Proceedings of the 22nd national conference on Artificial intelligence - Volume 1.* AAAI Press, 2007. This paper introduces a general framework for near-optimal nonmyopic information gathering in graphical models using submodular functions and the multilinear extension.

2. Buchbinder, Niv, et al. "A tight linear time  $(1/2)$ -approximation for unconstrained submodular maximization"

*Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms.* 2014. This paper presents a linear-time algorithm for

Hallucination

unconstrained submodular maximization.

3. Vondrák, Jan. "Submodularity and its applications in combinatorial optimization."

*Lecture notes from the International Congress of Mathematicians (ICM), Madrid, Spain. 2006.* Vondrák's lecture at the International Congress of Mathematicians discusses various aspects of submodularity in combinatorial optimization, including its applications and the use of the multilinear extension for submodular maximization problems.

c You

Sentence: Technopolis plans to develop in stages an area of no less than 100,000 square meters in order to host companies working in computer technologies and telecommunications, the statement said.

Label: Neutral

Sentence: The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility; contrary to earlier layoffs the company contracted the ranks of its office workers, the daily Postimees reported.

Label: Negative

Sentence: With the new production plant, the company would increase its capacity to meet the expected increase in demand and would improve the use of raw materials and therefore increase the production profitability.

Label: Positive

Can you predict the label of the following sentence?

Sentence: When this investment is in place , Atria plans to expand into the Moscow market .

Label:

< 3 / 3 >

ChatGPT

Positive

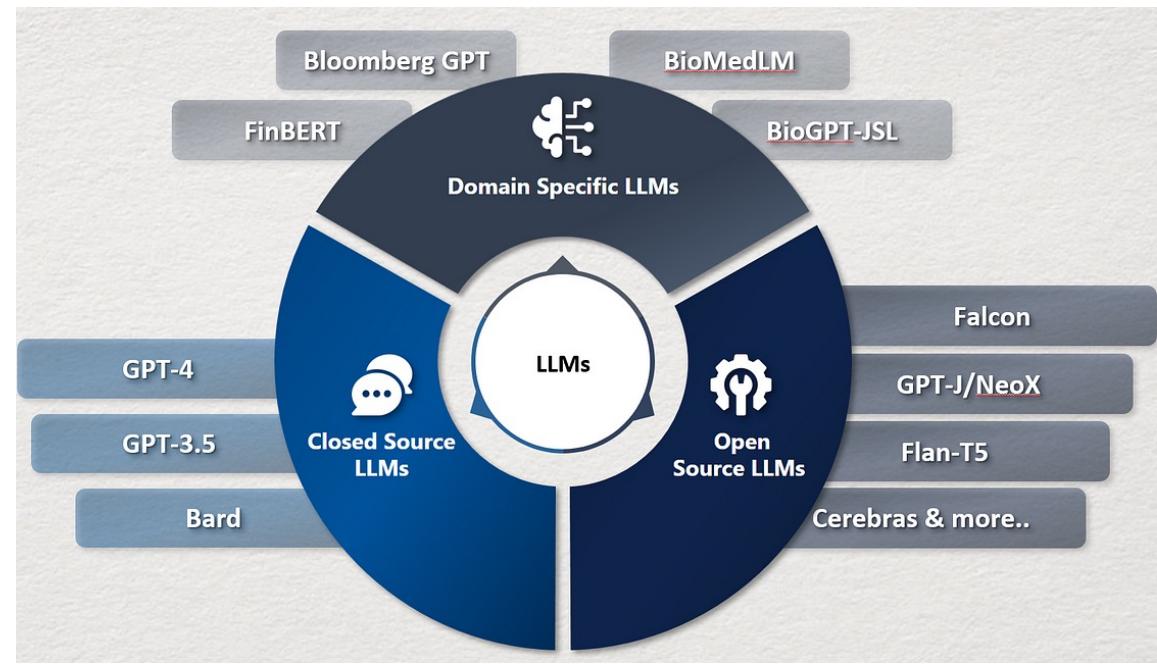


Wrong Prediction

# What is a domain-specific LLM

---

- Domain-specific LLM is a general model trained or fine-tuned to perform well-defined tasks dictated by organizational guidelines.
- Domain-specific LLMs serve a clearly-defined purpose in real-world applications.



# What are techniques of building Domain-Specific Large Language Models?

---

**Methods**

# Three Fundamental Ways of Building Domain-Specific LLM...

## Black-box LLMs



- More accessible but hard to interpret.
- Their decision-making processes (algorithm) are not directly observable.

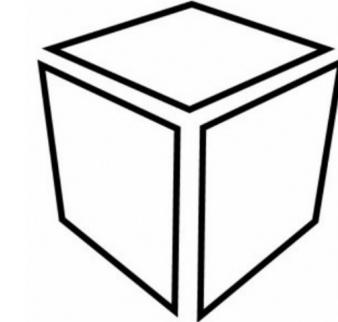


## External Augmentation

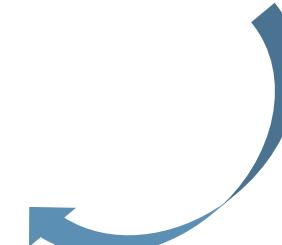
## Model Fine-tuning

## Prompt Crafting

## White-box LLMs



- Less accessible but easy to interpret.
- We have full access to the model architecture and data (maybe?)



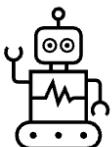
## Grey-box LLMs

# External Augmentation – Domain Knowledge

- Without modifying/touching the model, some tasks cannot be achieved if there are no background information provided.



Instruction I



LLM

Can we provide LLMs  
with necessary  
Background Info?



You

Can you give me a brief summary of "Silicon Valley Bank and Signature Bank collapsed in March 2023"?



ChatGPT

I'm sorry, but I couldn't find any information about Silicon Valley Bank and Signature Bank collapsing in March 2023. It's possible that it might be a hypothetical scenario or a piece of fictional content. If you have any other questions or need information on a different topic, feel free to ask!

Missing  
Domain Data



WIKIPEDIA  
The Free Encyclopedia



Collapse [edit]

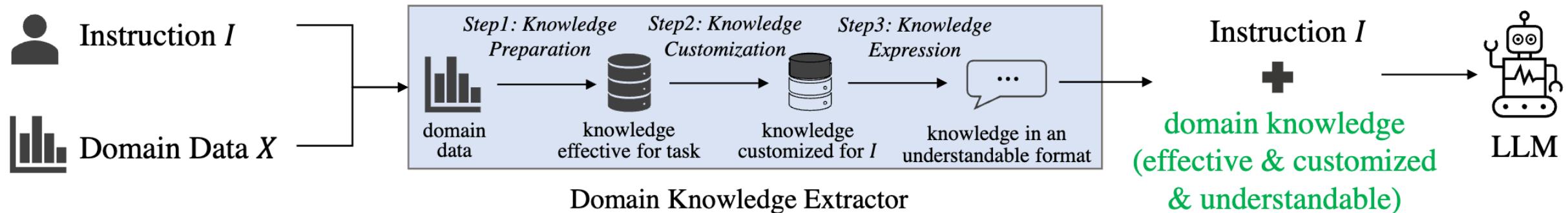
Main article: [Collapse of Silicon Valley Bank](#)

In 2022, SVB began to incur steep losses following increased interest rates and a major downturn in growth in the tech industry, with the bank heavily concentrated in long-term Treasury bonds.<sup>[68][69]</sup> As of December 31, 2022, SVB had mark-to-market accounting unrealized losses in excess of \$15 billion for securities held to maturity.<sup>[68]</sup> In early March of 2023, a combination of factors – including poor risk management and a bank run driven by tech industry investors – caused the bank to collapse.<sup>[70][71]</sup> Use of social media was reported to be a factor in both the initial bank run and its aftermath, with those affected by the potential loss of deposits calling for regulators to ensure that uninsured accounts were made whole.<sup>[72][73][74][75]</sup>

Early in the morning of March 10, examiners from the Federal Reserve and the FDIC arrived at the offices of SVB to assess the company's finances.<sup>[76]</sup> Several hours later, the California Department of Financial Protection and Innovation (DFPI) seized SVB<sup>[77]</sup> citing inadequate liquidity and insolvency,<sup>[78]</sup> and appointed the FDIC as receiver.<sup>[79][80]</sup> The FDIC then established a deposit insurance national bank, the Deposit Insurance National Bank of Santa Clara, to re-open the bank's branches the following Monday and enable access to insured deposits.<sup>[18][81][82][83]</sup> The CEO of Silicon Valley Bank, Greg Becker, was previously on the board of directors at the Federal Reserve Bank of San Francisco, but exited that position.<sup>[84]</sup> An initial auction of Silicon Valley Bank assets on March 12 attracted a single bid from an undisclosed suitor,<sup>[85]</sup> after PNC Financial Services and RBC Bank backed away from making offers.<sup>[86][87]</sup> The FDIC rejected this offer and planned to hold a second auction to attract bids from major banks, now that the bank's systemic risk designation allowed the FDIC to insure all deposits.<sup>[85]</sup>



# External Augmentation – Domain Knowledge



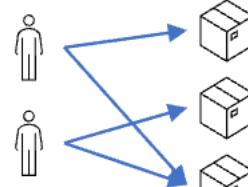
**Instruction  $I$ :**

Can you give me a brief introduction of Silicon Valley Bank and Signature Bank collapse happened in March 2023?

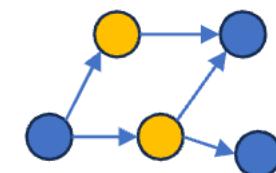
**Domain Data  $X$  Formats:**



Metadata  
6/11/24



Interaction Data



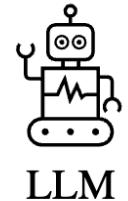
Knowledge Graph

**Augmented Instruction  $I$ :**

*[Background]: In 2022, SVB began to incur steep losses following increased interest rates, ...*

*Happened in March 2023, ..., caused the bank to collapse.*

Can you give me a brief introduction of Silicon Valley Bank and Signature Bank collapse happened in March 2023?



In 2022,  
SVB was ...

In 2023, ...

Summary

# External Augmentation – Domain Tools

- Domain tools refer to specialized software, libraries, or frameworks that are developed specifically for a particular domain or field.
- We can do Calculus problems (linked to calculator), make travel plannings (links to Tripadvisor), or even trading analysis (Yahoo Finance).

**Trending**  
Most popular GPTs by our community

- 1  **image generator**  
A GPT specialized in generating and refining images with a mix of professional and friendly tone.image generator  
By NAIF J ALOTAIBI
- 2  **Write For Me**  
Write tailored, engaging content with a focus on quality, relevance and precise word count.  
By puzzle.today
- 3  **Consensus**  
Your AI Research Assistant. Search 200M academic papers from Consensus, get science-based answers, and draft content...  
By consensus.app
- 4  **Scholar GPT**  
Enhance research with 200M+ resources and built-in critical reading skills. Access Google Scholar, PubMed, JSTOR, Arxiv, an...  
By awesomegpts.ai
- 5  **Canva**  
Effortlessly design anything: presentations, logos, social media posts and more.  
By canva.com
- 6  **Logo Creator**  
Use me to generate professional logo designs and app icons!  
By community builder
- 7  **Humanizer Pro**  
#1 Humanizer in the market! This tool humanizes AI-generated content, maintaining content meaning and quality...  
By charlyaisolutions.com
- 8  **AskYourPDF Research Assistant**  
Free Chat Unlimited PDFs, Access 400M+ Papers (PubMed, Nature, Arxiv, etc), Analyse PDF (Unlimited PDFs), Generate...  
By askyourpdf.com

Connect human request with finetuned image generator ...

Avoid hallucination by directly searching for references in external database ...

6/11/24

19

# Prompt Crafting

Large Language Models are  
Few-shot Learners  
(Brown, et al.)



- **Prompt** is a cue given to the pre-trained language model to allow it better understand human's questions.
- **With good prompts, LMs can already achieve decent zero-shot performance** on tasks require domain knowledge as LLMs are trained on them!

**Prompt:**

The sky is

**Output:**

Blue

**Prompt:**

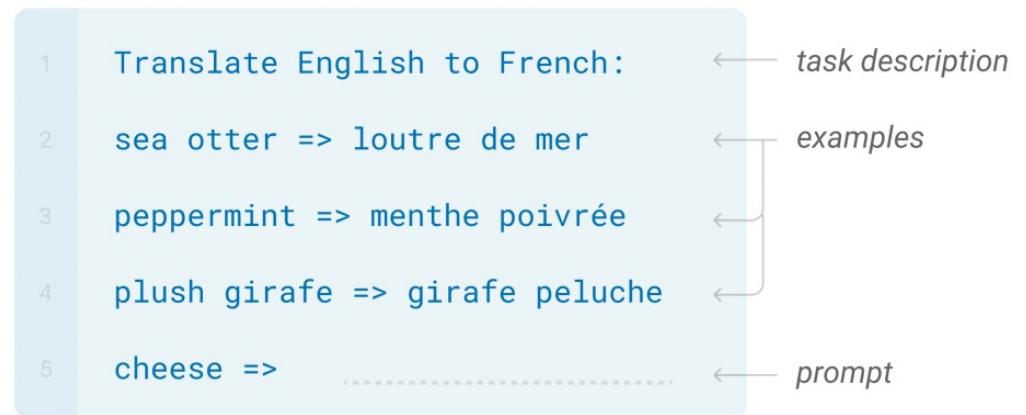
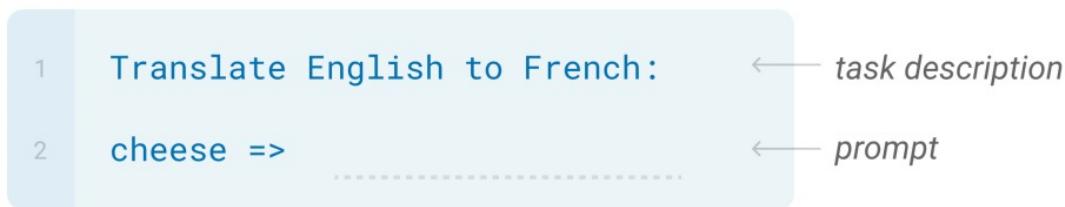
Complete the sentence:  
The sky is

**Output:**

blue during the day and  
dark at night.

# Prompt Crafting – Discrete Prompt

- **Zero-shot:** The model predicts the answer given only a natural language description of the task. No gradient updates are performed.
- **Few-shot:** In addition to the natural language description of the task, the model sees a few examples of the task.

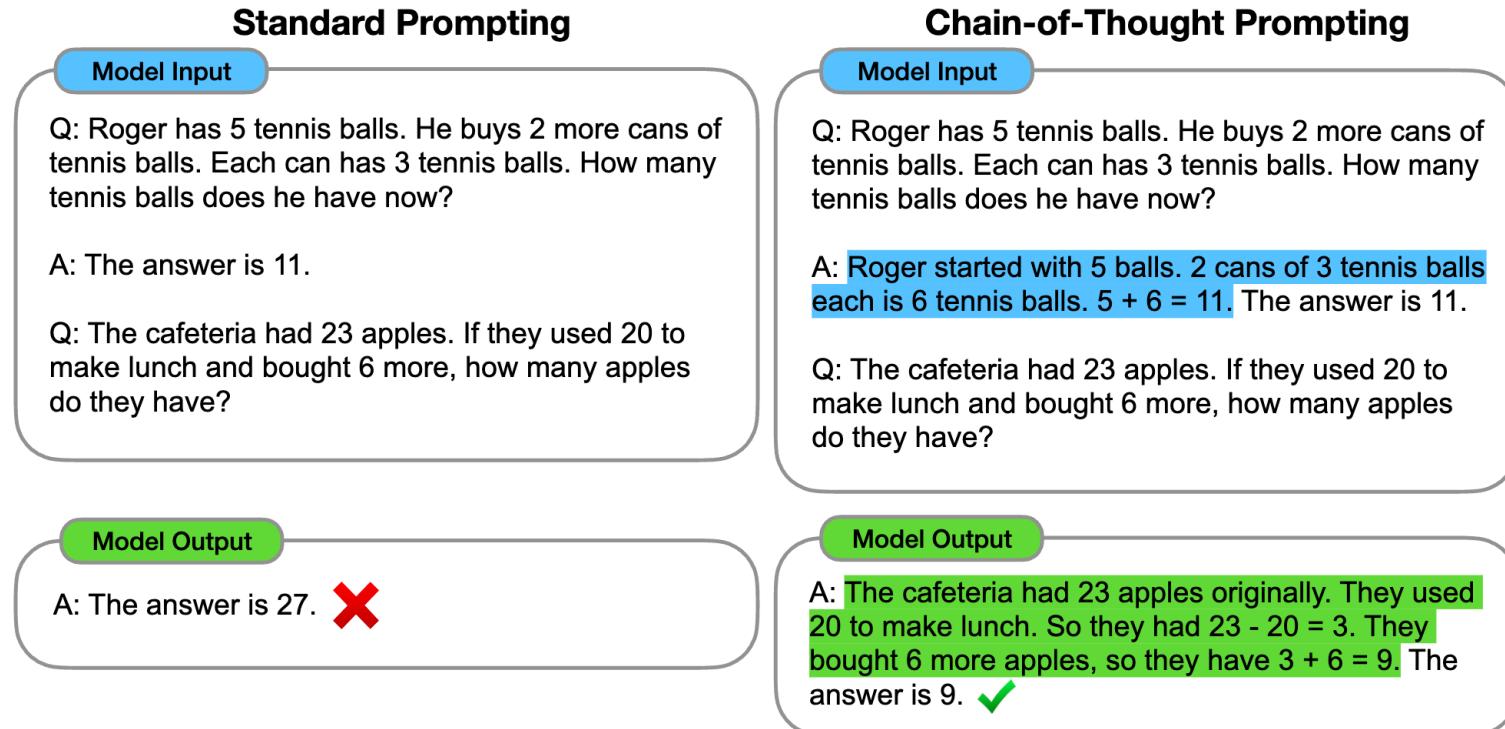


## General Paradigm of a good Prompt:

1. Detailed task description
2. A few examples with desired output
3. The final prompt of the output

# Advanced Discrete Prompt Techniques

- Chain-of-Thought (CoT) Prompting
  - Instead of directly showing the desired answer, show all thinking steps!
  - Perhaps adding a magical phrase: “**Let’s think step by step...**”



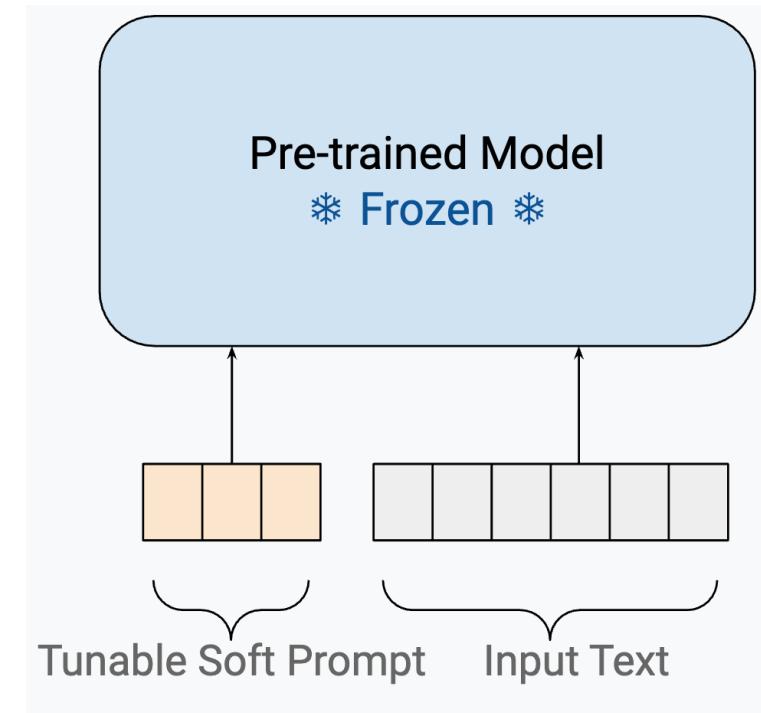
# Prompt Crafting – Continuous Prompt

- For domain-specific application, we need substantial **Human effort** to get good prompts
- You can only fit **a few examples** in the model's context, especially **NOT good** for things like reading comprehension with long inputs.
- **Searching in the text space** is hard...

Optimizing the prompt in  
the latent space with  
gradient descent ...



In-depth Domain  
Knowledge

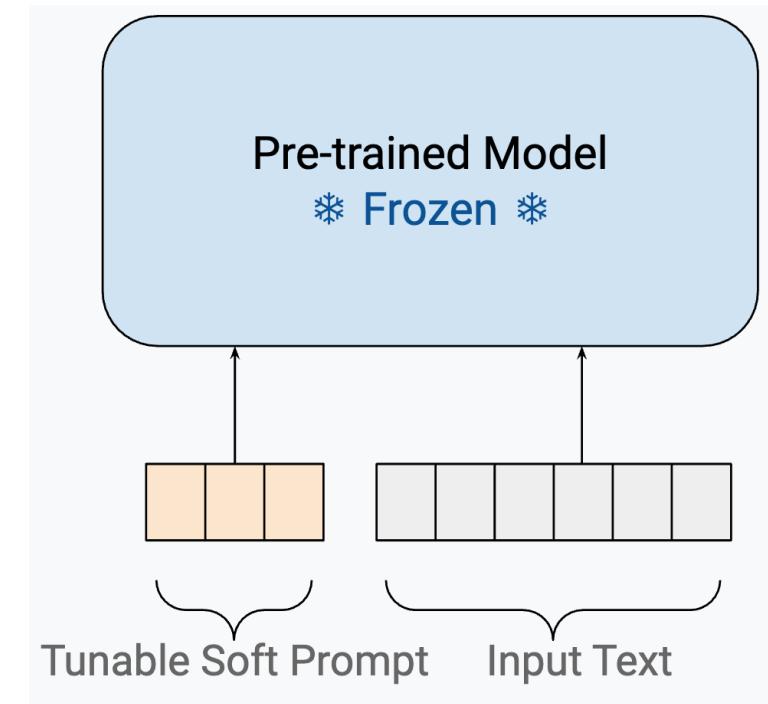


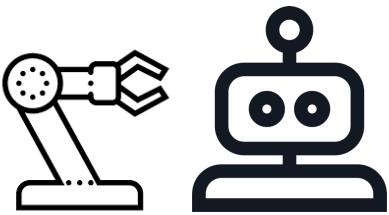
# Prompt Crafting – Continuous Prompt

- Prepend **virtual tokens** to input.
- Prompt and input representations flow through model like normal.
- **Learn embeddings of only these special tokens**, via backprop. Keep rest of model **fixed**.

## Advantages:

- Use whole training dataset.
- Automatically learn a new prompt for a new model.
- Keep the model frozen.
- Prompts are much smaller.





# Model Fine-tuning

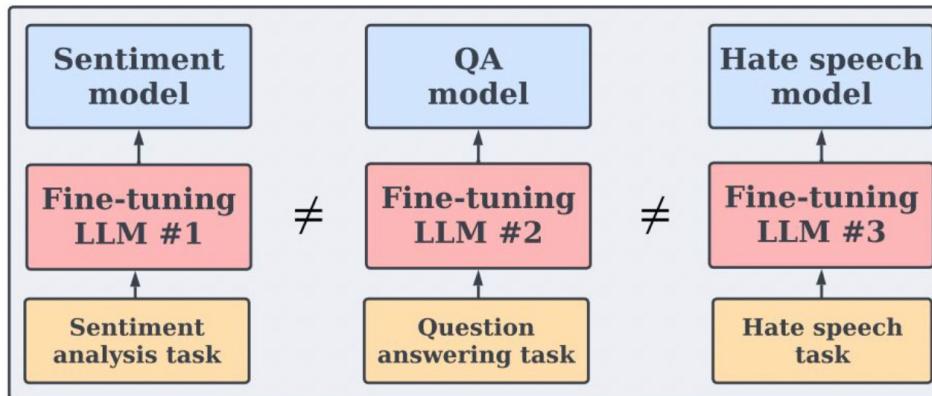
- **External Augmentation and Prompt Crafting** let LLMs be aware of the domain-specific tasks.
- **Fine-tuning** can directly **encode adequate knowledge** for specific tasks or domains.

Stage	Pretraining	Supervised Fine-tuning
Algorithm		Language modeling predict the next token
Dataset	Raw internet text ~trillions of words low-quality, large quantity	Carefully curated text ~10-100K (prompt, response) low quantity, high quality
Resource	<b>1000s of GPUs months of training</b> ex: GPT, LLaMA, PaLM	<b>1-100 GPUs days of training</b> ex: Vicuna-13B

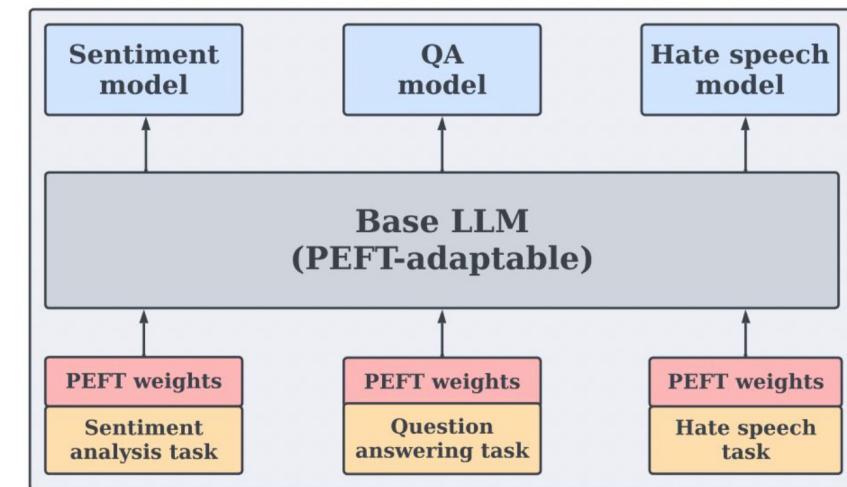
# Model Fine-tuning – Parameter Efficient

- **Parameter-Efficient Fine-tuning (PEFT)**: a class of methods that adapt LLMs by updating **only a small subset of model parameters**.
- Assumption: LLMs are smart enough to handle general tasks.

Full-parameter Fine-tuning

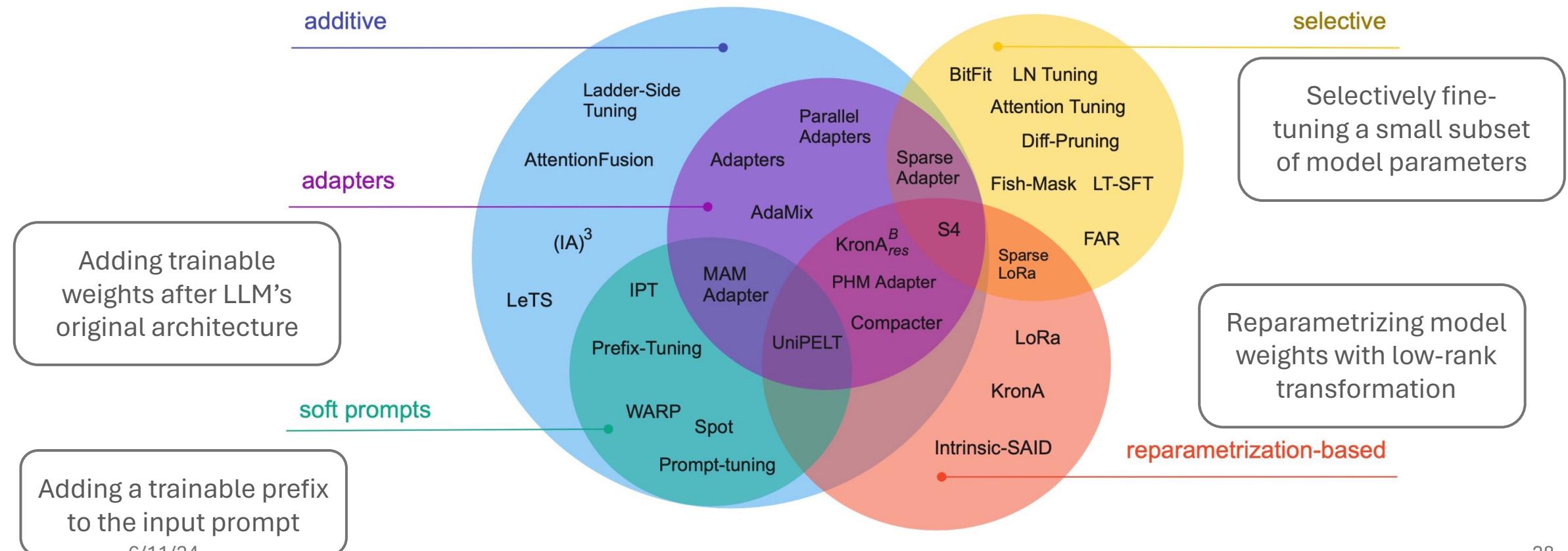


PEFT Fine-tuning



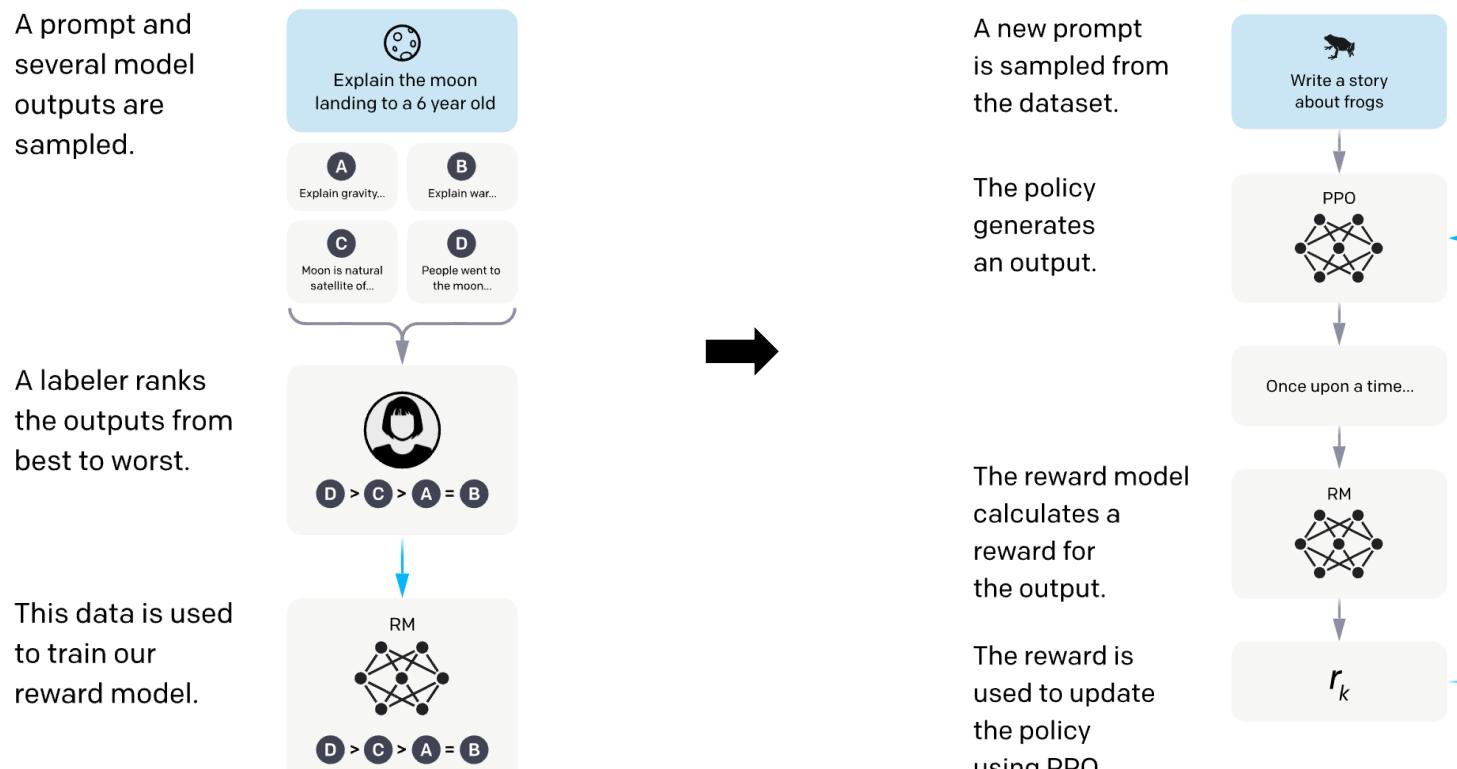
# Model Fine-tuning – Parameter Efficient

- **Parameter-Efficient Fine-tuning (PEFT)**: a class of methods that adapt LLMs by updating **only a small subset of model parameters**.
- Assumption: LLMs are smart enough to handle general tasks.



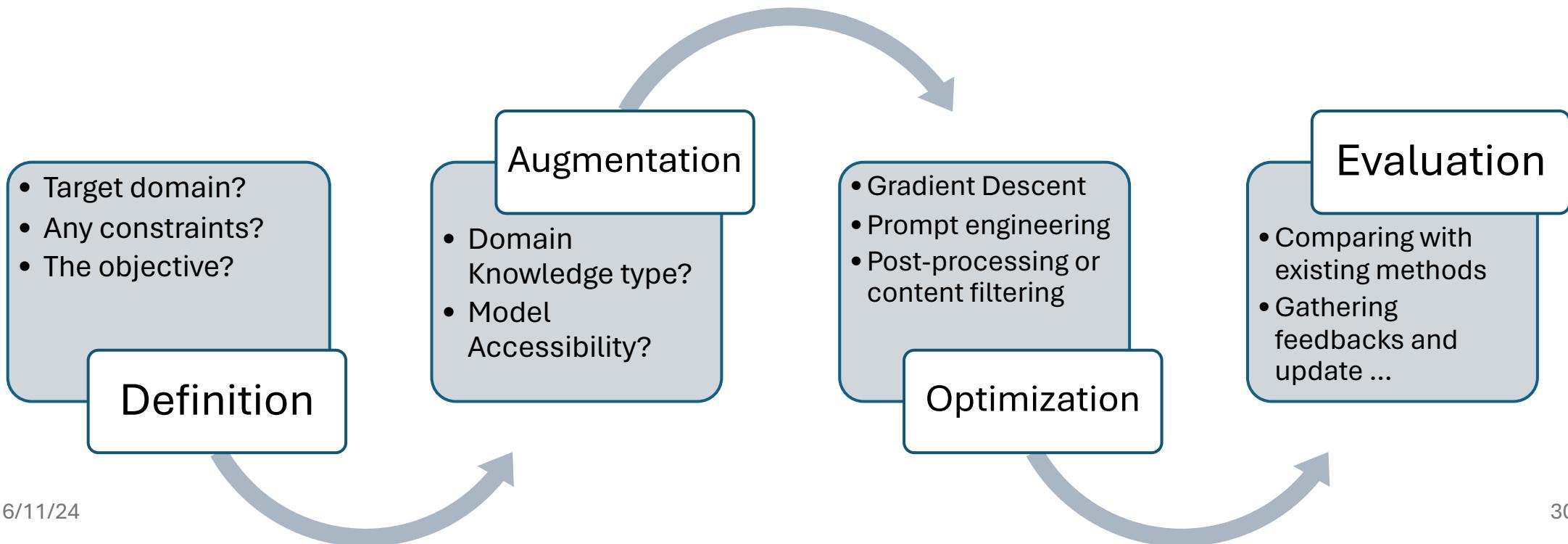
# Model Fine-tuning – Human Preference

- LLMs may produce
  - Harmful text – unparliamentary language, bias and discrimination
  - Text that can cause direct harm – allowing easy access to dangerous information
- Therefore, LLMs should be trained to produce outputs that **align with human preferences** and values



# Relations between Approaches in Different Categories

- Different levels of specialization (Black-box, Grey-box, and White-box):
  - **External Augmentation** focuses on providing domain-specific knowledge
  - **Prompt Crafting** works at input level
  - **Model Fine-tuning** injects knowledge to the internal parameters
- Three methods can work independently or collaboratively.



# Domain-specific Large Language Models for Artificial General Intelligence

---

**Conclusion**

# General vs. Specialized

- **[Initial Phase]** An infant can be viewed as a general, “untrained” model.
- **[Pretraining Phase]** From elementary school to high school, we take **general knowledge** from all subjects.
- **[Fine-tuning Phase]** In college, we take **Incremental Learning** only on certain subjects.
- **[Deployment Phase]** When we graduate and start job hunting, company hires you only if you have the **desired background**.



# Examples of Domain-Specific LLMs

- Industry leaders were aware of the limitations of general LLMs.
- So, they set forth to create custom LLMs for their respective industries:

**BloombergGPT:** 50-billion parameter large language model, purpose-built from scratch for finance

<i>Finance-Specific</i>	<b>BloombergGPT</b>	<b>GPT-NeoX</b>	<b>OPT-66B</b>	<b>BLOOM-176B</b>
<b>Financial Tasks</b>	<b>62.51</b>	51.90	53.01	54.35
<b>Bloomberg Tasks (Sentiment Analysis)</b>	<b>62.47</b>	29.23	35.76	33.39

<i>General-Purpose</i>	<b>BloombergGPT</b>	<b>GPT-NeoX</b>	<b>OPT-66B</b>	<b>BLOOM-176B</b>	<b>GPT-3</b>
<b>MMLU</b>	<b>39.18</b>	35.95	35.99	39.13	<b>43.9</b>
<b>Reading Comprehension</b>	<b>61.22</b>	42.81	50.21	49.37	<b>67.0</b>
<b>Linguistic Scenarios</b>	<b>60.63</b>	57.18	58.59	58.26	<b>63.4</b>

**Usage:**

1. **ChatBot:** 24/7 customer service
2. **Fraud Detection & Prevention:** Auto Anti-Money Laundering
3. **Credit Risk Management:** AI-enhanced risk management

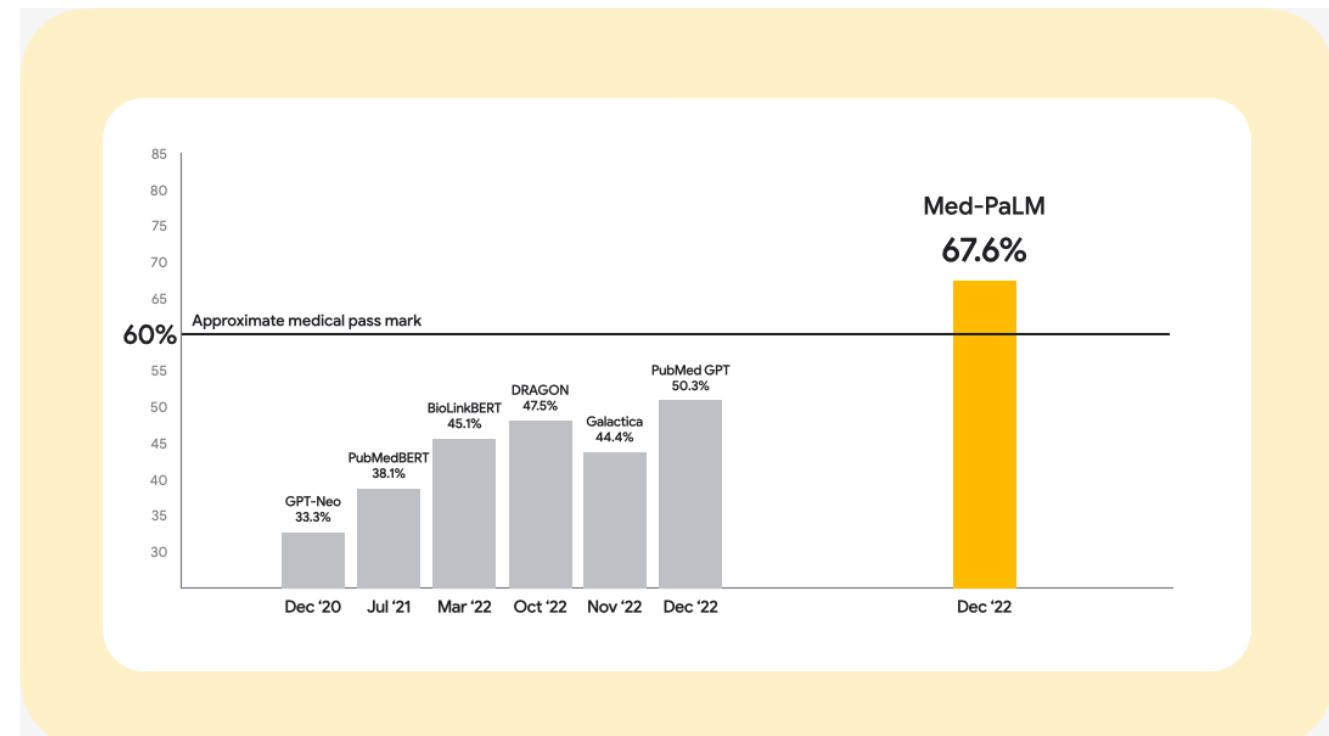
# Examples of Domain-Specific LLMs

- Industry leaders were aware of the limitations of general LLMs.
- So, they set forth to create custom LLMs for their respective industries:

**Med-PaLM:** Google built by training on carefully curated medical datasets.

## Usage:

1. Answer professional medical questions
2. Working on par with medical professionals in many use cases.
3. Passing US Medical Licensing Examination with **86.5% accuracy**.



# Examples of Domain-Specific LLMs

There are many more examples:



Customized GPT for legal services



AI Assistant for education



**GitHub Copilot**

Software :

Customized GPT for software development



rct AI

Analyze real-time player inputs and dynamically generate NPC responses

Customized GPT for game development

# Open Challenges

- **Domain Complexity**
  - Each domain has its unique intricacies and complexities
  - Legal and medical domain use many domain-specific jargons that follow certain rules.
- **Balancing General and Domain Knowledge**
  - Fine-tuning may hurt the generalization ability of a general LLMs
  - Overly specialized model → fail to understand or generate coherent responses to prompts.
- **Data Quality**
  - Domain-specific training data is a fair representation of the diversity of real-world data.
  - Banks need to train an AI Credit Scoring model based on data that can reflect customer demographics, otherwise bias may occur.

---

## Some Takeaways

- Domain-specific LLMs are a better fit for knowledge-specific tasks.
- Start small and understand the scaling law.
  - Larger ≠ Better
  - Model size and data quality are equally vital
- Domain specific models will positively transform industries:
  - Unlocking financial opportunities
  - Improving operational efficiency
  - Elevating customer experience.

