

高阶大语言模型课程

Huajun Zeng

12/13/2024 - 1/31/2025
(12月27日和1月3日放假, 共计6次课)
每周五 5pm-7pm PT / 8pm-10pm ET

第一课: Retrieval Augmented Generation (RAG) for LLM

课程安排

Week	Date	Content	Week	Date	Content
1	2024-12-13	Retrieval Augmented Generation (RAG) for LLM <ul style="list-style-type: none">• Why augmenting LLMs?• Methods for LLM augmentation• Augmenting LLMs with retrieval• Augmenting LLMs with fine tuning	4	2025-01-17	Pipeline for LLM Applications: From Code to Products <ul style="list-style-type: none">• Full stack LLM: tools needed for an LLM application• Case study: build an LLM app from ground
2	2024-12-20	Chatbot Building with LLM APIs <ul style="list-style-type: none">• Environment setup• Introduction to LLM APIs• Using chat completion APIs• Using fine tuning APIs	5	2025-01-24	More LLM Applications and Course Project <ul style="list-style-type: none">• Showcase of potential LLM applications for productivity, creativity and more• More advanced LLM applications: AI Agent, Multi-modality, etc.• Introduction to the course project: requirement and discussion
3	2025-01-10	Chatbot Building with LLM Frameworks and Vector Database <ul style="list-style-type: none">• Introduction to Langchain and LlamaIndex• Case study: a chatbot from Langchain and vector database	6	2025-01-31	Project Presentation <ul style="list-style-type: none">• Student presentation on the course project

所需的基础知识

- 大语言模型的基础知识
 - ChatGPT
 - 提示词工程
- Python编程
 - Python语法: 变量、数据类型、条件语句、循环、函数
 - 数据处理: 字符串、列表、字典等操作及相关库函数

大语言模型基础知识

大模型的发展现状

- ChatGPT
 - Launched in 11/2022
 - 100 million users in 2 months
- Major player
 - OpenAI
 - Anthropic
 - Meta
 - Google
 - ...
- Trends

Multimodal

Reasoning

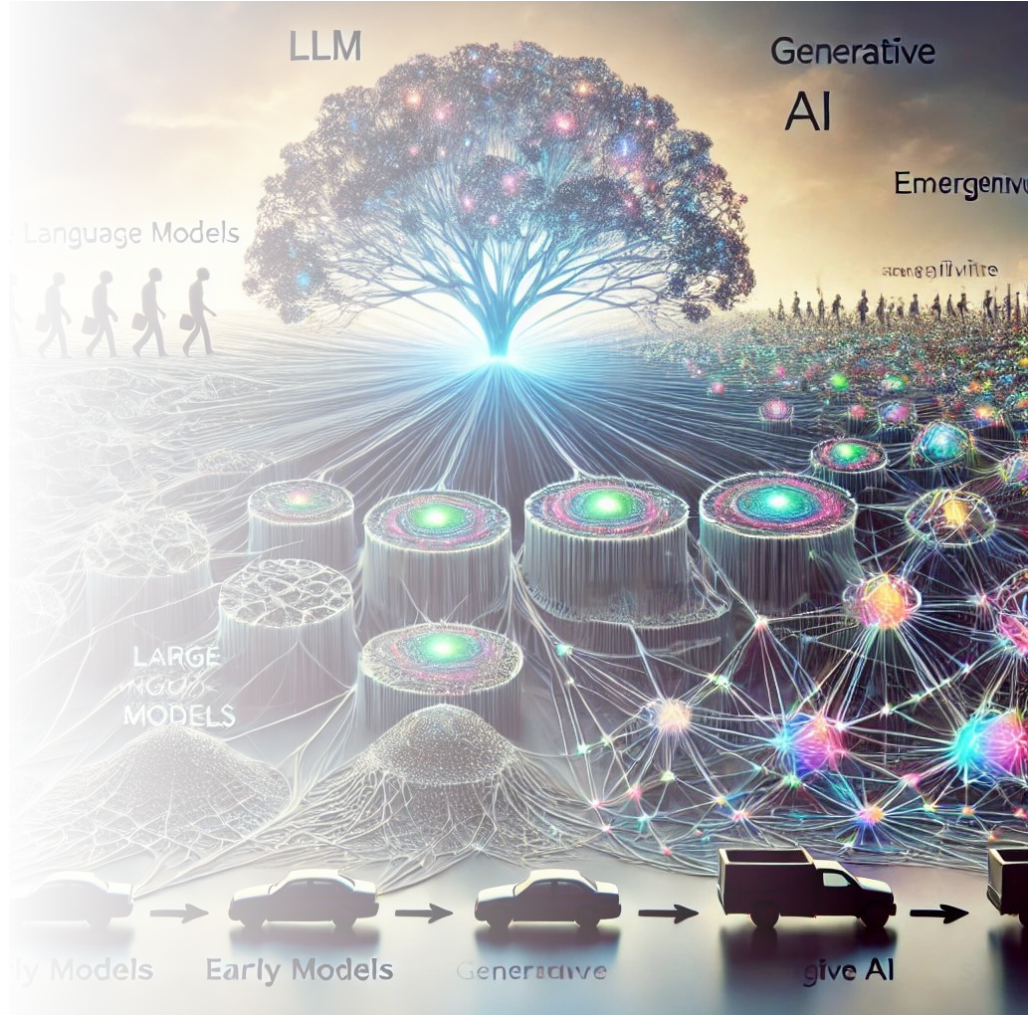
App: RAG,
Agent, ...



大模型的能力

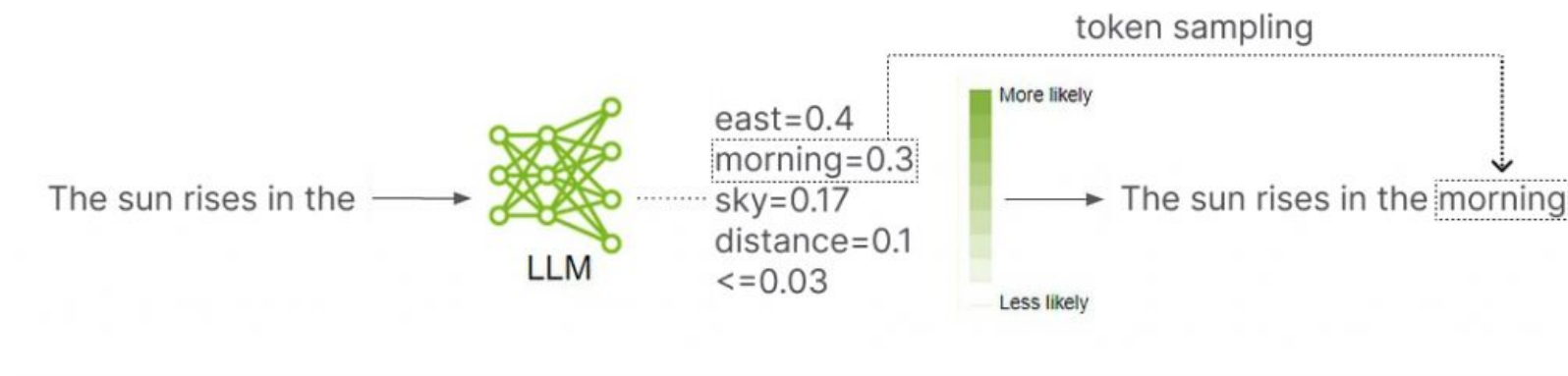
- 使用海量数据预训练
 - Web
 - Book
 - ...
- 参数多
 - GPT-3.5: **175B**
 - GPT-4: **8 x 222B**
 - Llama 3.x: **8B, 70B, 405B**
- 具有超强能力
 - 文本处理
 - 写作
 - 知识问答
 - 数学分析
 - 推理

涌现



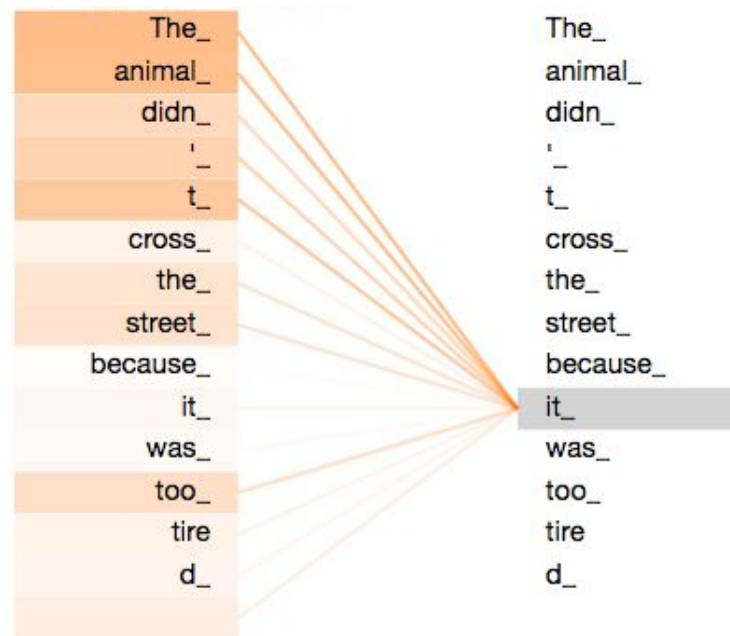
大语言模型是怎么工作的

- Probabilistic model of a natural language
- Predict next token one at a time



大语言模型的基本单元: Attention

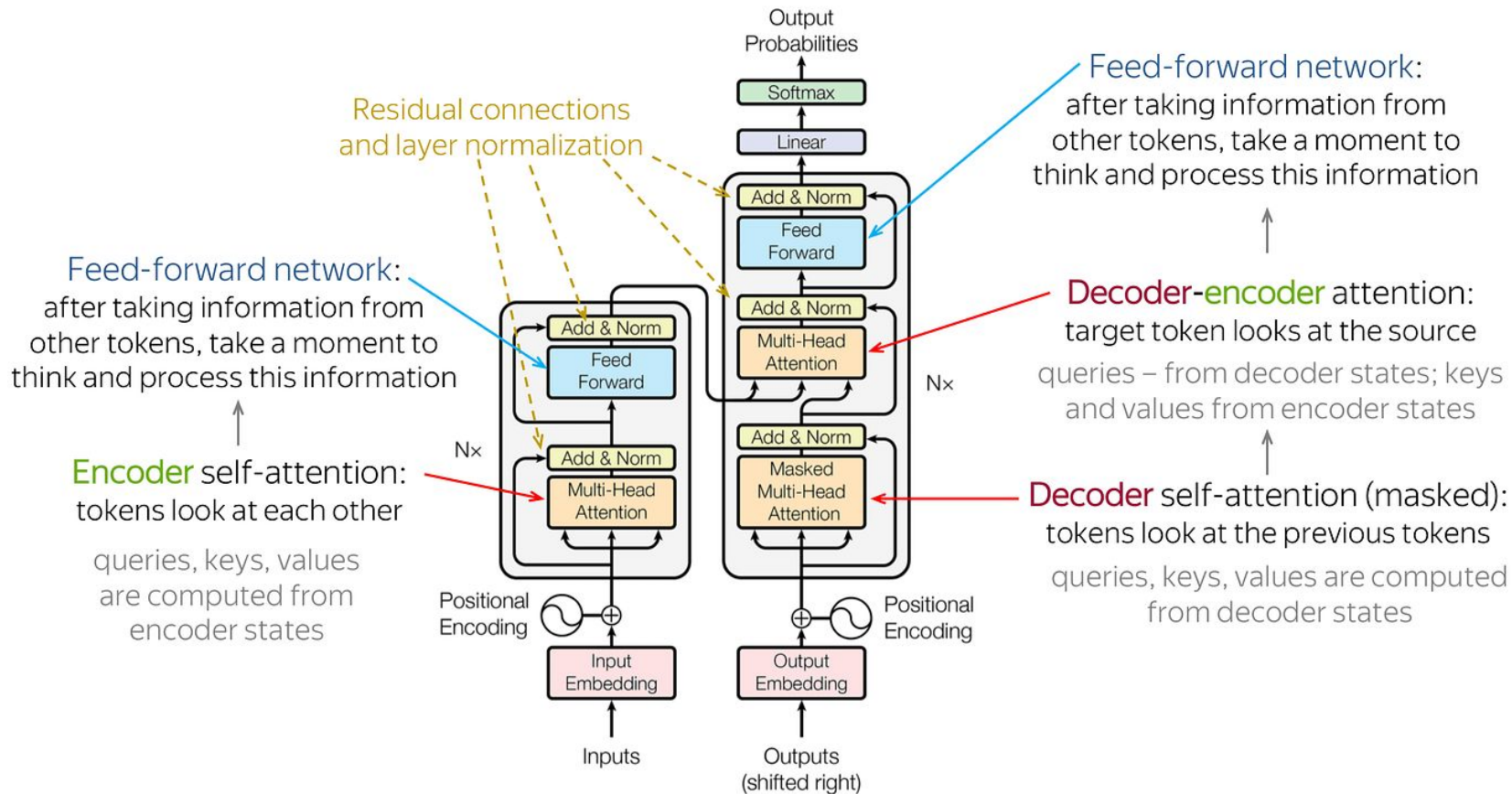
- A mechanism that enhances the information by including its context
- Good at long-term dependencies and parallelism



Example Weights in a Self Attention Structure

<https://arxiv.org/abs/1706.03762>

Transformer架构



大模型是如何训练出来的

- 预训练 (Pre-Training)
 - The models learn to predict the next word in a sequence of text, using the previous words as context.
 - The training data consists of huge corpuses of natural text from web pages, books, Wikipedia, etc. consisting hundreds of billions of words.
- 有监督微调 (Supervised Fine-tuning)
 - The model is further trained on specific tasks or domains using smaller, specialized datasets
- 强化学习 (RLHF)
 - Optimize the model's performance in generating outputs that align with human preferences and values

提示词工程 (Prompt Engineering)

- 设计和优化提示词的过程
 - 帮助AI模型理解和执行特定任务
 - 提高AI响应的有效性和准确性
- Zero-Shot or Few-Shot Learning



简单的提示词

我生产了一款电饭煲, 请为它撰写一份营销文案。

优化的提示词

请撰写一份吸引人的 营销文案, 推广我们公司最新推出的智能 电饭煲。这款电饭煲的主要特点如下:

1. 采用AI智能控温技术, 精确控制煮饭温度
2. 具备10种烹饪模式, 包括煮饭、煲汤、蒸菜等
3. 内胆采用纳米级钛金涂层, 不粘易清洁
4. 大容量 5L, 适合3-6人家庭使用
5. 智能预约功能, 最长可提前24小时预约

请按照以下要求撰写文案:

1. 突出目标用户, 并且使用目标用户语言
例如: "作为一名爱美的女性, 你是否经常为了护肤而熬夜? 第二天起来皮肤状态却不尽如人意..." (化妆品)
2. 突出产品的特点和优势, 在用户心目中超越竞争对手
例如: "市面上的手机拍照都是千篇一律? 只有我们的AI智能手机, 才能拍出媲美单反的高质量照片。" (手机)
3. 说出并解决用户痛点。
例如: "担心发胖, 担心没时间锻炼? 有了这款跑步机, 在家也能获得好身材" (健身器材)
4. 语言风格: 现代、活泼、有亲和力, 多使用口语化表达
例如: "嘿, 追求时尚的你! 你的妆容是不是缺一支百搭的口红?" (化妆品)
5. 长度: 600字左右, 适合在社交媒体上传播

请提供一个完整的 营销文案示例, 并展示你的思考过程和链式推理, 解释你是如何根据以上要求和特点一步步构思 这个文案的。

提示词的要素

- 指令 (Instruction)
- 背景信息 (Context)
- 角色 (Role)
- 示例 (Examples)
- 思维链 (Chain of Thoughts)



怎样扩展大语言模型

大语言模型不 包含最新信息

User

Who will be the next US president?

Assistant

I'm unable to predict future events, including the outcome of political elections. The next U.S. presidential election is scheduled for November 5, 2024, and the candidates and results will be determined by the voters.

🗑️ 🗣️ {=}

- LLM doesn't automatically update with up-to-date information
- LLM providers usually just periodically retrain new models

大语言模型不 包含私有领域 信息

- Private data is not available in LLM's training data

USER	Create a project plan for Otter.ai in the next quarter.
ASSISTANT	<p>Project Plan for Otter.ai</p> <p>Objective: To enhance user experience and increase user engagement on the Otter.ai platform in the next quarter.</p> <p>Key Deliverables:</p> <ol style="list-style-type: none">1. Implement a new user feedback system to gather insights and suggestions for improvements.2. Develop and launch new features to improve usability and functionality.3. Conduct a marketing campaign to increase awareness and acquisition of new users.4. Collaborate with key partners to expand integration opportunities and reach new audiences. <p>Timeline:</p> <p>Week 1-2:</p> <ul style="list-style-type: none">- Set up the new user feedback system and begin collecting insights from existing users.- Analyze feedback data to identify key areas for improvement and

大语言模型不能准确理解应用场景

User

作为一个客服代表，请回答一个用户要求退货的请求。

Assistant

当然，我很乐意为您处理退货请求。请您按照以下步骤操作：

1. ****提供订单信息****：请您提供订单号和购买日期，这将有助于我们更快地找到您的订单信息。
2. ****确认产品状态****：请确认产品是否未开封，未使用，并附带所有原始包装、配件和标签。如果产品有任何质量问题，请详细说明。
3. ****查看退货政策****：根据我们的退货政策，请确认您是在购买后的一定时限内申请退货。

4. ****寄回产品****：在我们确认你的退货请求后，你将收到一个退货授权号码和详细的寄回说明，请按照指示操作。

- Every use case is unique
- LLM needs to be specialized

使用提示词扩展LLM的例子

System message

Today is Sunday, 12/1/2024. You are a personal scheduling chatbot. You have access to a simulated Google calendar with the following appointments:

Last Week:

Monday (11/25):

- 9:30am: Team Stand-up Meeting
- 2pm: Client Presentation

Tuesday (11/26):

- 11am: Performance Review
- 4pm: Department Budget Meeting

Wednesday (11/27):

- 1pm: Lunch with Colleagues
- 3:30pm: Website Planning Session

Upcoming Week:

Monday (12/2):

- 9:30am: Team Stand-up Meeting
- 1:30pm Project Review Meeting

Tuesday (12/3):

- 2pm: Conference Call

User

Hi, What is my availability tomorrow?

Assistant

Tomorrow, on Monday, December 2nd, you have the following appointments:

- 9:30am: Team Stand-up Meeting
- 1:30pm: Project Review Meeting

Your availability tomorrow is:

- Before 9:30am
- Between 10:00am and 1:30pm
- After 2:30pm

🗑️ 🗒️ ☰

大语言模型上下文窗口有限制

- Context windows of typical LLMs
 - GPT 4o: 128k tokens (roughly 380 pages)
 - Claude 3.5: 200k tokens (roughly 600 pages)
- There are more than 10 million words in the news everyday
- There are trillions of text data in a typical enterprise

扩展大语言模型的步骤

- Use simple prompt with out-of-the-box LLM
- If not good enough, optimize prompt via prompt engineering
- If not good enough, add few-shot examples
- If not good enough, collect more examples and
 - a. Build a RAG system, or
 - b. Fine tune a model, or
 - c. Both
- If still not good enough, spend \$\$\$ and train a new LLM

<https://ai.meta.com/blog/adapting-large-language-models-llms/>

是否要做Fine-Tuning

- Benefit of fine-tuning
 - Tone, style, and format customization
 - Increasing accuracy and handling edge cases
 - Addressing underrepresented domains
 - Cost reduction
 - Certain new tasks/abilities
- Potential problem with fine-tuning
 - Catastrophic forgetting
 - Newer large language models, developed monthly or more often, can defeat fine-tuning improvements.

<https://ai.meta.com/blog/when-to-fine-tune-llms-vs-other-techniques/>

Fine-Tuning

- 通过少量Fine-Tuning提升效果
 - 当提示词不足以应对更复杂或更细微的任务时，可以通过微调进行调整。
 - 例如在法律文本理解或医疗报告生成中，使用少量专门领域的示例对模型进行微调，可以大幅提升其在特定领域的准确性。
- 微调的实践步骤
 - 收集特定领域数据：例如医疗报告、法律案例等。
 - 训练并微调模型：使用这些数据对大语言模型进行训练，使其在该领域任务中的表现更加精准。
 - 结果评估：对比微调前后的模型表现，验证微调是否有效提升了模型的任务执行能力。

Fine-Tuning的方法

- Full parameter fine tuning
 - Updating all the parameters of a pre-trained model to adapt it to a specific task
- Parameter efficient fine tuning, for example LoRA
 - An LLM is like an already constructed building with a complete and powerful infrastructure.
 - LoRA is equivalent to adding some detachable furniture and decoration (low-rank matrices) within the building, rather than altering the building's structure.
 - These furniture and equipment allow the building (model) to quickly adapt to different uses and needs (tasks).

OpenAI's Fine Tuning

<https://platform.openai.com/finetune>

检索增强生成 (RAG)

检索增强生成的例子

User

Given the following context:

The Guardian

US election 2024 results live: Donald Trump wins presidency, defeating Kamala Harris

US election 2024 results live: Donald Trump wins presidency, defeating Kamala Harris ... Find your result ...

Mass. NJ. Conn. Del. DC. Md. RI. Calif...

1 month ago

Al Jazeera

US election result: Where did Harris and her campaign go wrong?

The pre-election predictions were clear: The contest between former President Donald Trump and Vice President Kamala Harris would be decided...

1 month ago

Answer this question:

Who will be the next US president?



Assistant

According to the provided context, Donald Trump won the 2024 US presidential election, defeating Kamala Harris, and will be the next US president.



检索增强生成

- Definition
 - Retrieve relevant content to the query
 - Compose a prompt using those information
 - Send to LLMs to generate a response
- Benefit
 - Get up-to-date information
 - Get private domain information
 - Make LLMs to adapt to certain application domain
 - Solve the hallucination problem

信息检索 (Information Retrieval)

- Information retrieval is the process of finding and accessing information from huge amounts of data.
- Typical application is search engine
- Major components
 - Inverted index
 - Ranking

使用Search Engine 实现RAG

- Query word generation
- Inverted index lookup
- Ranking
- Result extraction
- Respond by LLM



What are latest iPhone models?

🔗 Sources

List of iPhone models -
Wikipedia

 en.wikipedia

Best iPhones 2024:
Which iPhone model
should you buy? -...

 zdnet

iPhone comparison
chart: Every iPhone
compared - Macworld

 macworld



Show all

✳️ Perplexity

🛒 5 Products

The latest iPhone models as of 2024 are from the iPhone 16 series, released in September 2024:

- **iPhone 16:** Base model with A18 chip and improved cameras.
- **iPhone 16 Plus:** Larger version with extended battery life.
- **iPhone 16 Pro:** Advanced features like A18 Pro chip, dual 48MP cameras, and refined titanium design.
- **iPhone 16 Pro Max:** Largest display, top-tier performance, and longest battery life. 1

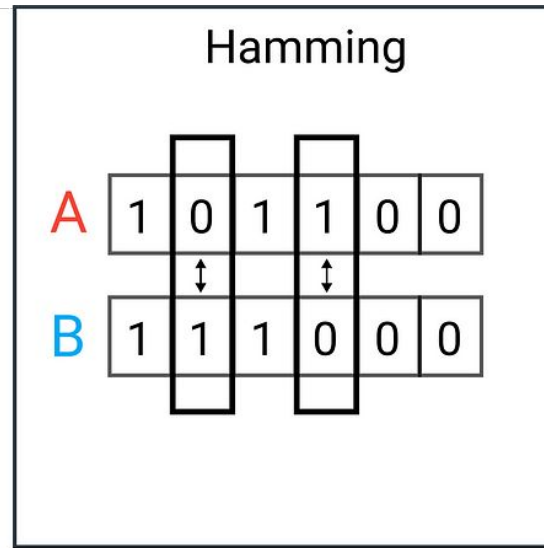
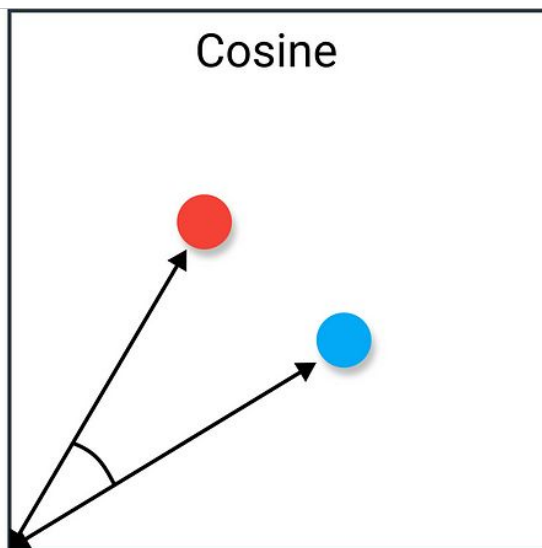
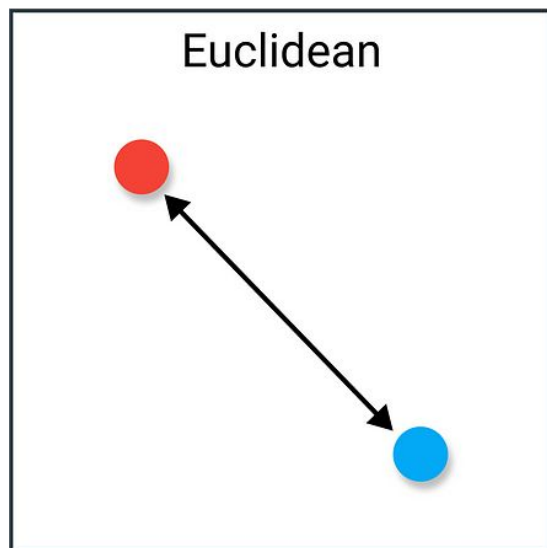
3

使用关键词搜索的优缺点

- Pros:
 - Widely available prebuilt search tools like ElasticSearch
 - Fast and scalable with inverted indexing
 - Tuning search experience with weights and boosts
- Cons:
 - No semantic relevance

What are latest iPhone models?

相似度计算



Cosine similarity (dot product):

$$\mathbf{a} = [a_1, a_2, \dots, a_n] \quad \mathbf{b} = [b_1, b_2, \dots, b_n],$$

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$



文本Embedding算法

- OpenAI
- Voyage
- BGE
- Cohere
- Sentence Transformers
- ...

Embedding效果排行榜

<https://huggingface.co/spaces/mteb/leaderboard>

基于Embedding的搜索算法

- Bruteforce (e.g. numpy.dot)
- HNSW
- Faiss (Meta)
- ScaNN (Google)
- ...

K近邻 (KNN) 算法

- Given a query vector $[x_1, x_2, x_3, \dots]$
- And a list of document vector
 - $[y_{11}, y_{12}, y_{13}, \dots]$
 - $[y_{21}, y_{22}, y_{23}, \dots]$
 - ...
- Output the k nearest neighbors from the documents for the input query

使用numpy来寻找K近邻

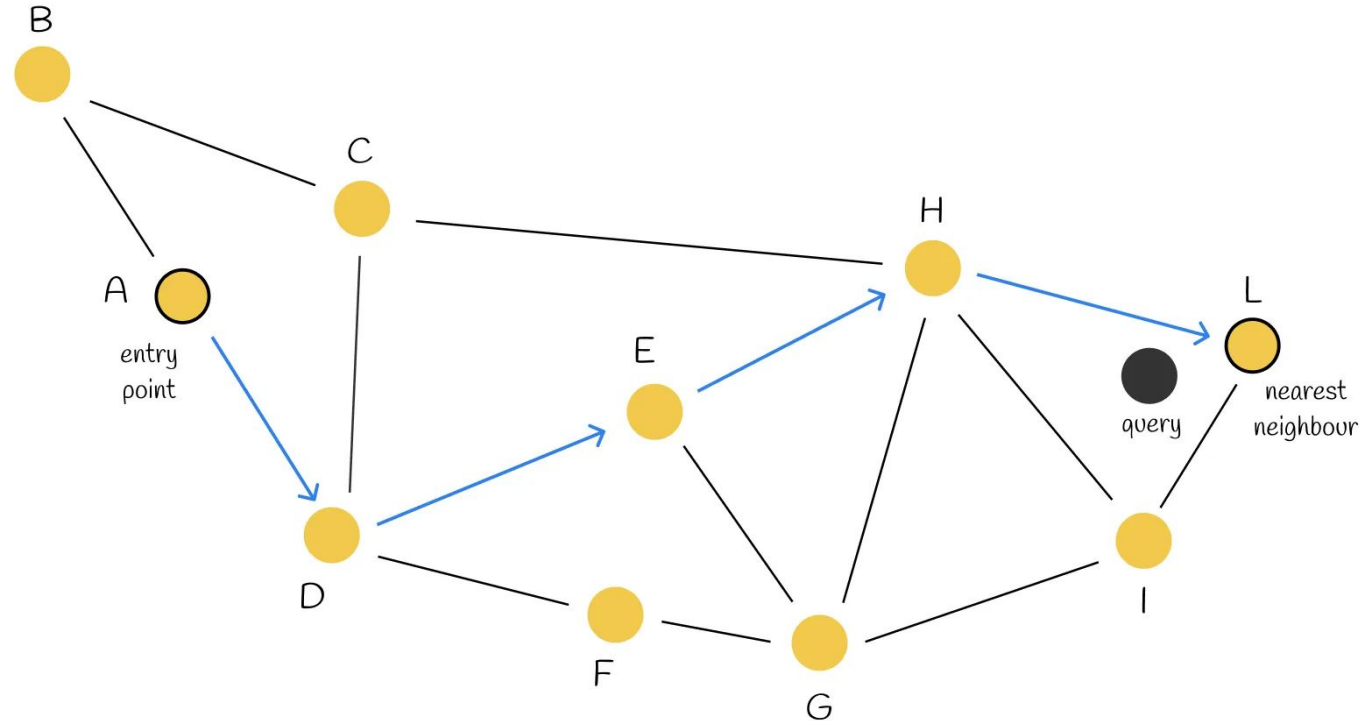
```
import numpy as np

def find_nearest_neighbor(query_vec, list_of_vectors):
    similarities = np.dot(query_vec, list_of_vectors)
    nearest_neighbor_index = np.argmax(similarities)
    return nearest_neighbor_index
```

The numpy implementation of matrix multiplication is very fast because of vectorization and hardware optimization (e.g. BLAS, MKL)

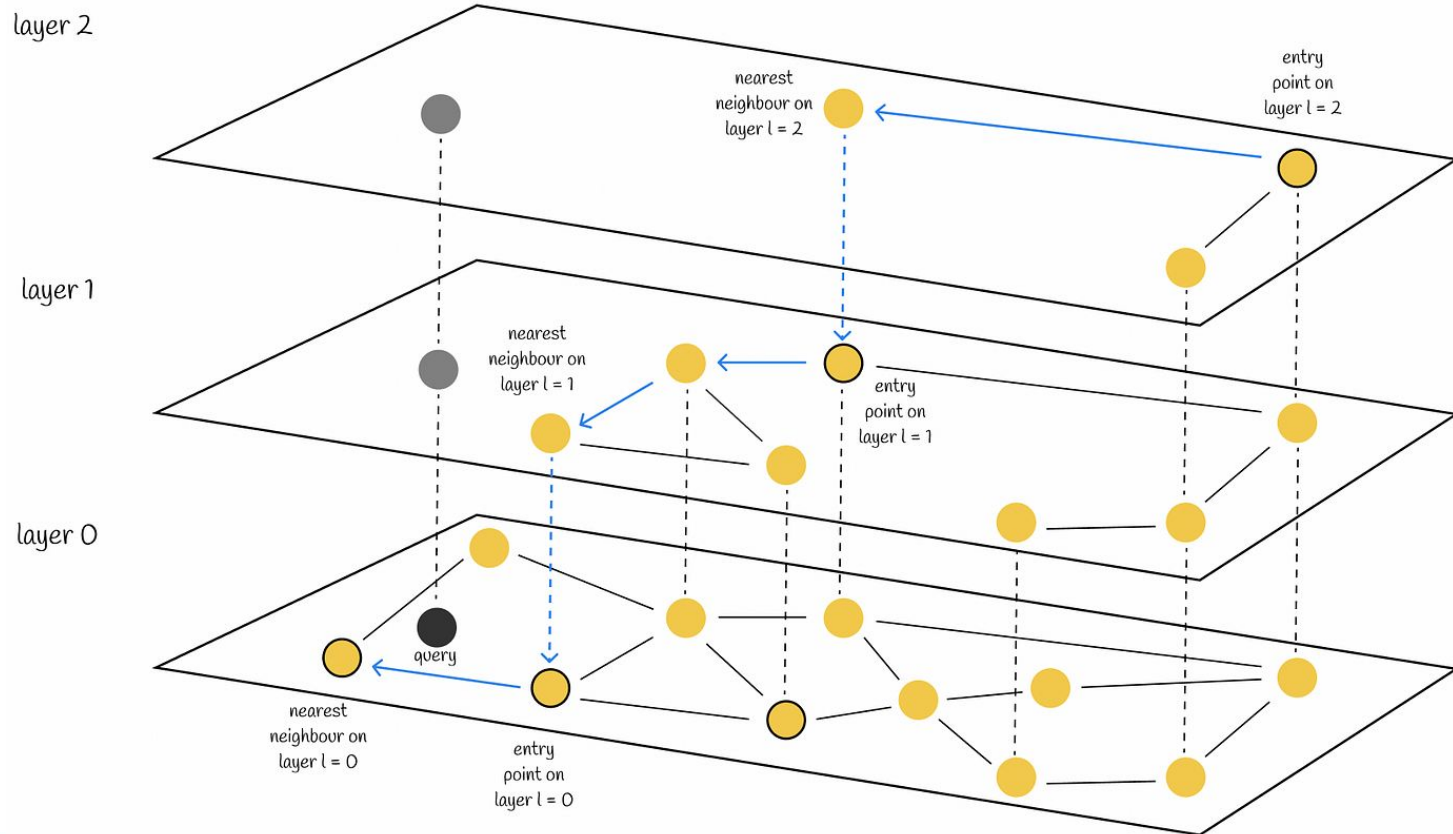
Hierarchical Navigable Small Worlds (HNSW)

Main idea: a graph where a path between any pair of vertices could be traversed in a small number of steps.

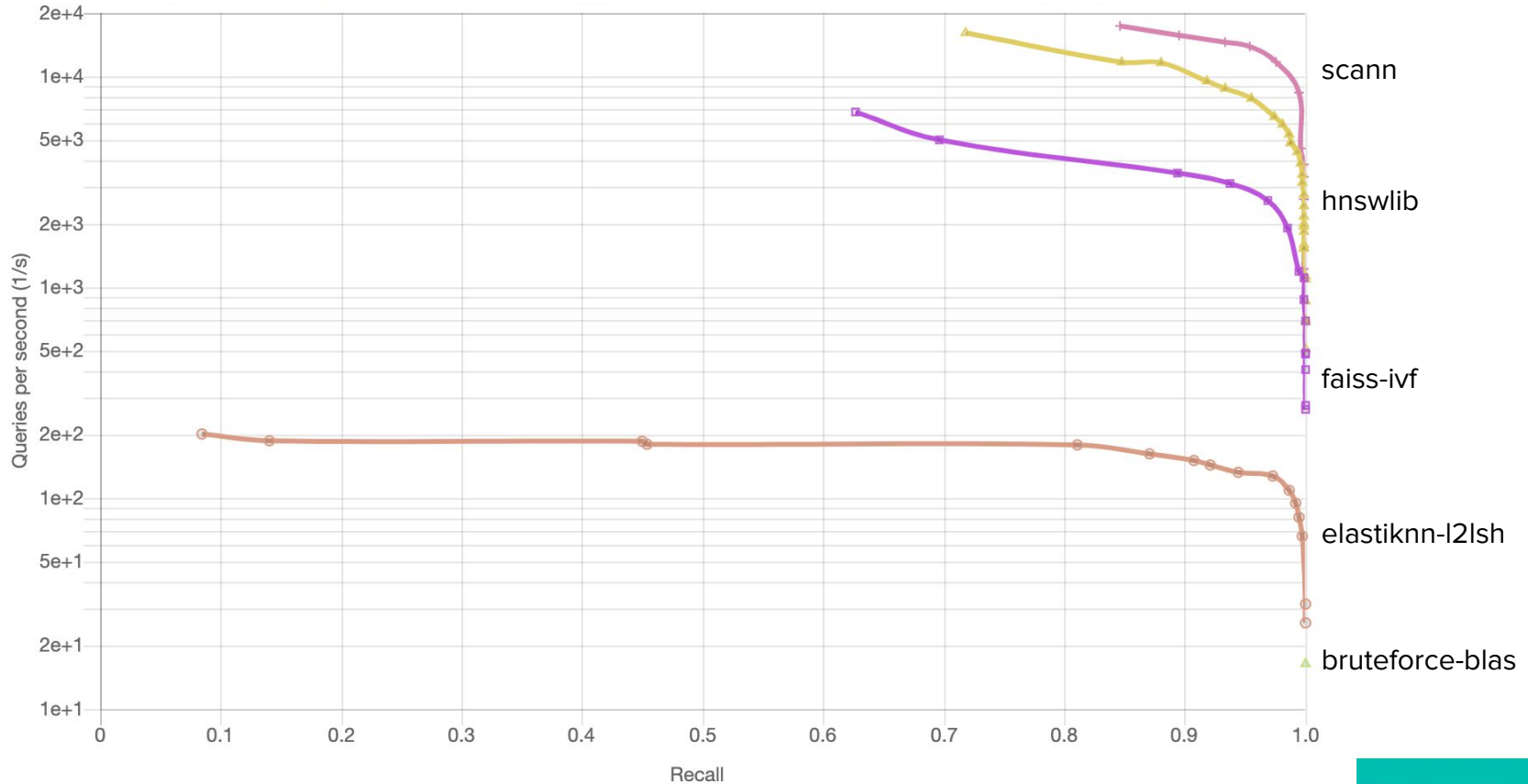


Hierarchical Navigable Small Worlds (HNSW)

Top layers with
sparse vertices
and do big
search steps
Low layer with
dense vertices
and do small
search steps



Benchmarks: ann-benchmarks.com

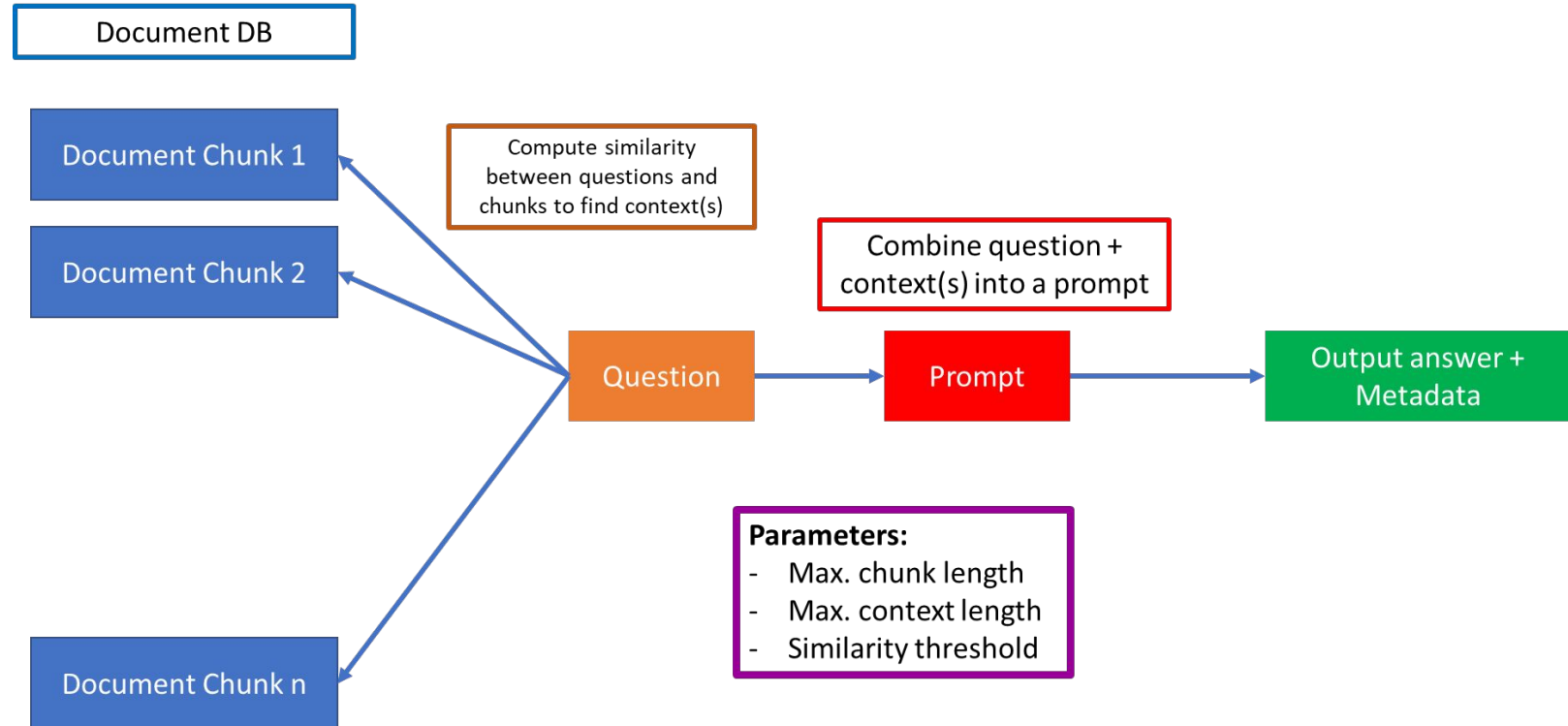


Vector Databases

- Pinecone
- Qdrant
- Milvus
- Chroma
- Weaviate
- AtlasDB
- Elasticsearch
- ...

not in any particular order

General Architecture of RAG



Pinecone Introduction

<https://www.pinecone.io/>

家庭作业

- Register an OpenAI account and start using the console (platform.openai.com)
- Register a Pinecone account and start experiencing the embeddings and index

Questions?