LLM -Basic

MinXie

2024/05

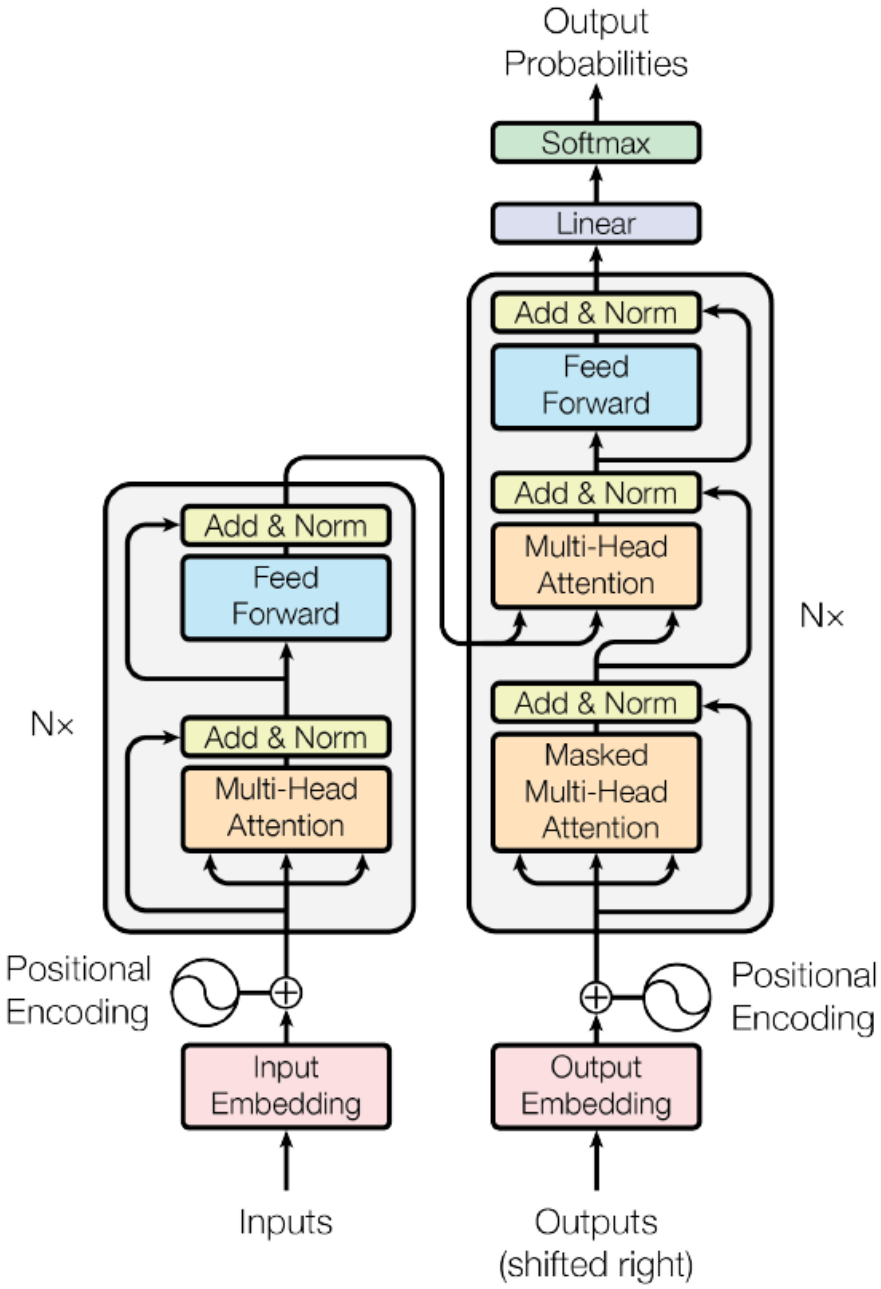
transformer 模型架构

Attention is all you need
https://arxiv.org/abs/1706.03762

transformer 模型的架构
Encoder and Decoder Stacks

sequence to sequence model with attention

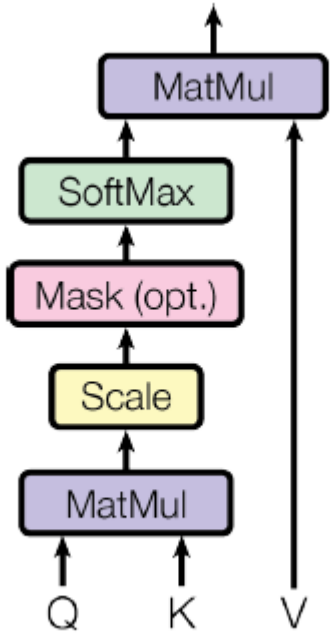
Transformer架构中attention是一个创新



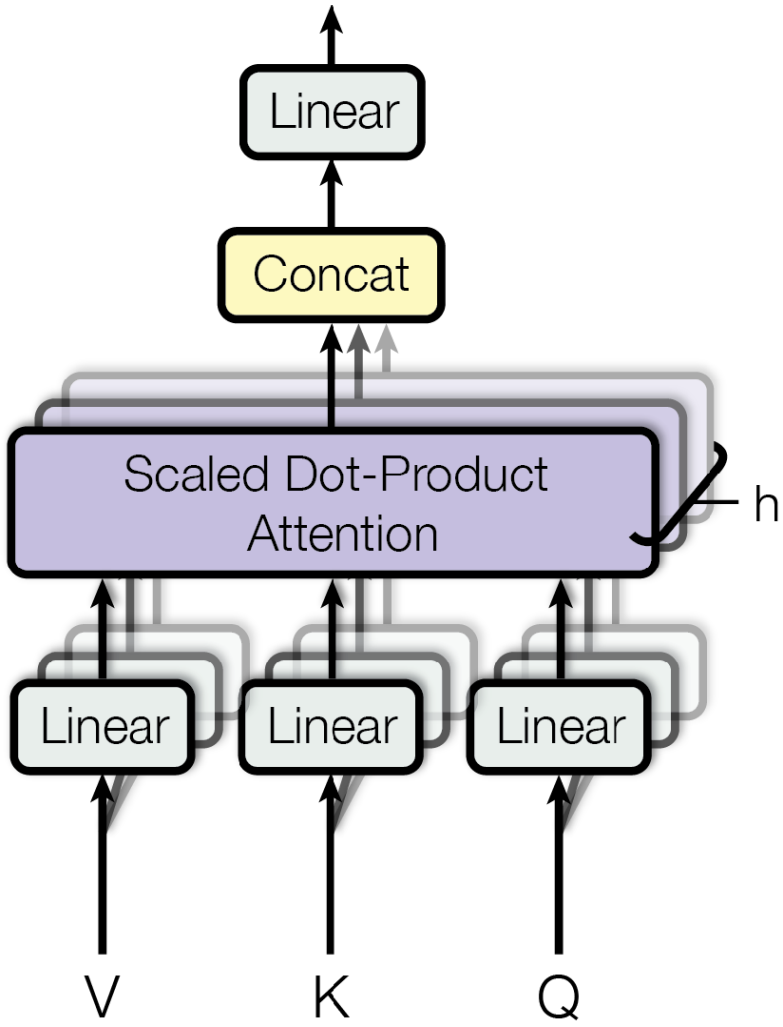
Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- attention
- self-attention
- Multi head attention
- Multi head self attention



Scaled Dot-Product Attention



Multi head Attention

Tokenization & Embedding

```python
class InputEmbedding(nn.Module):
    """Implementation of the input embeddings"""

    def __init__(self, d_model: int, vocab_size: int) -> None:
        super().__init__()
        self.d_model = d_model
        self.vocab_size = vocab_size
        self.embedding = nn.Embedding(vocab_size, d_model)

    def forward(self, x):
        return self.embedding(x) * math.sqrt(self.d_model)
```

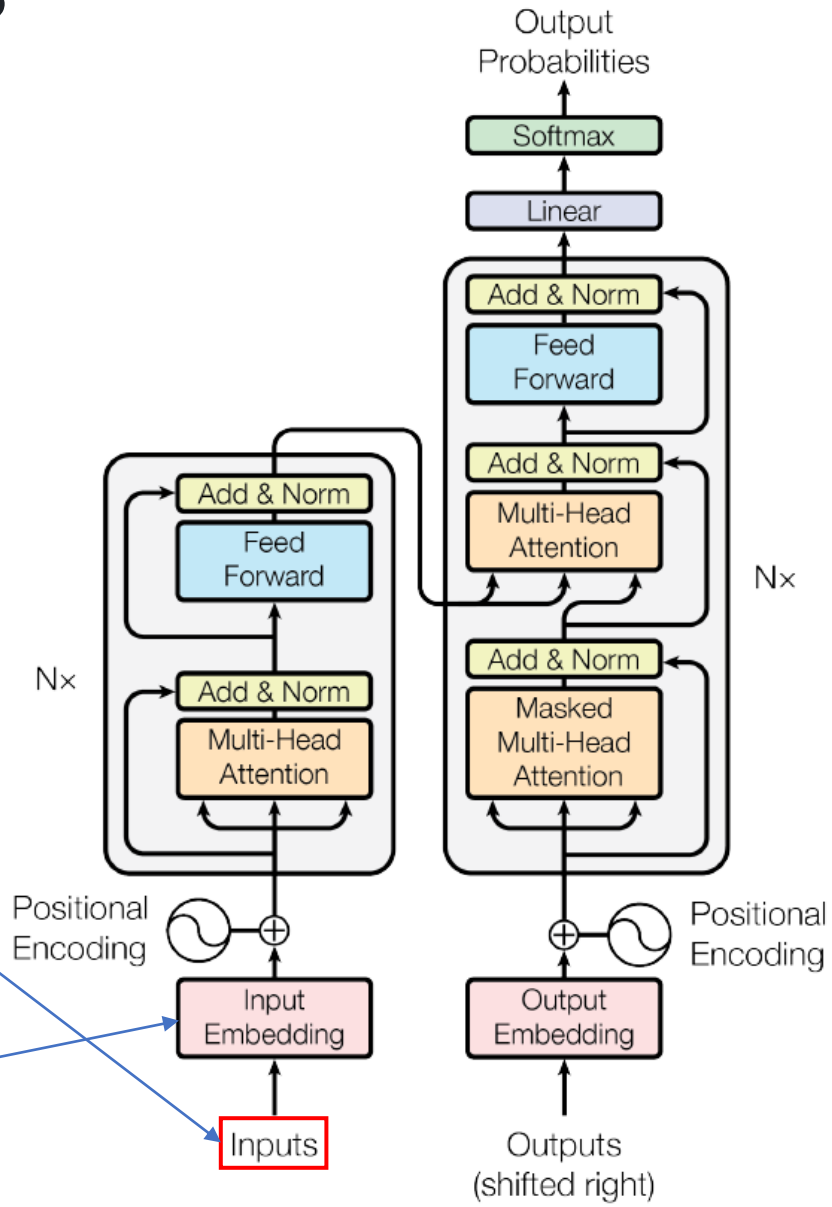
"This is a input text."

Tokenization

[CLS]	This	is	a	input	.	[SEP]
101	2023	2003	1037	7953	1012	102

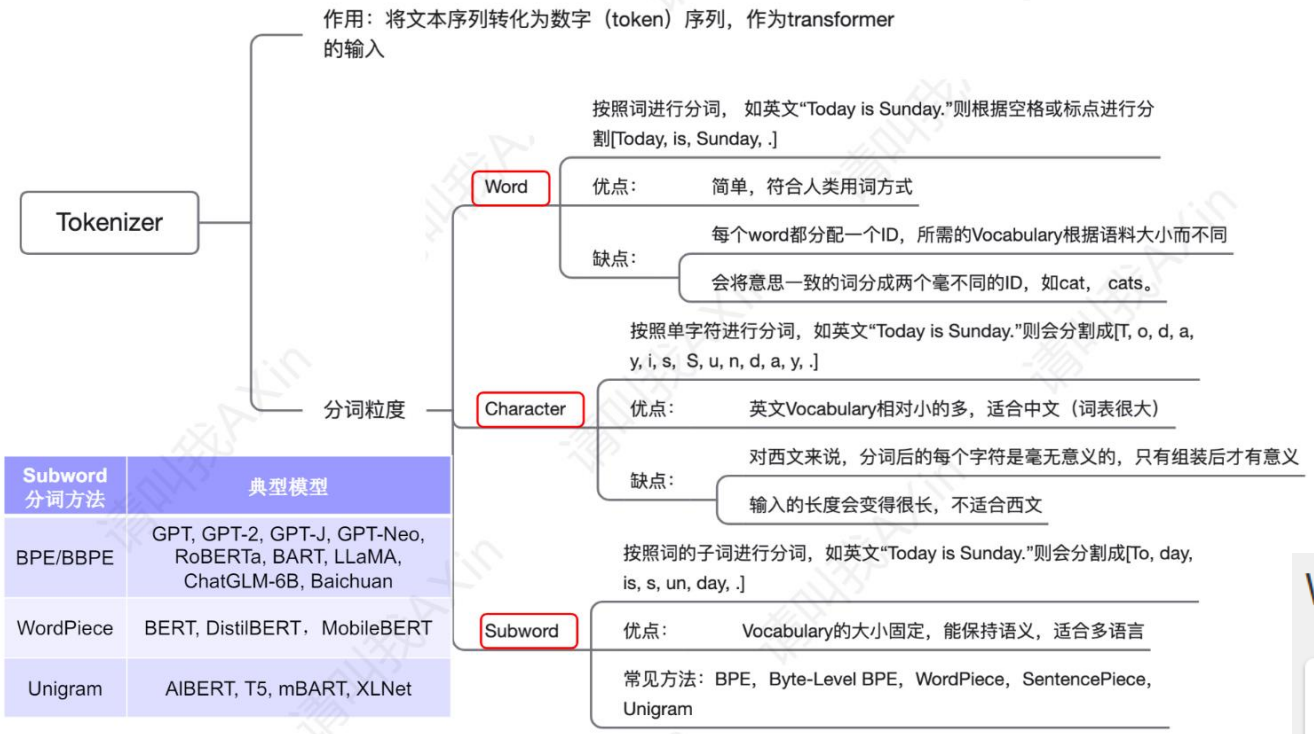
Embeddings

0.0390, -0.0123, -0.0208, ...	-0.0558, 0.0151, 0.0031, ...	-0.0440, -0.0236, -0.0283, ...	0.0119, -0.0037, -0.0402, ...	0069, 0.0057, -0.0016, ...	0.0199, -0.0095, -0.0099, ...	-0.0788, 0.0202, -0.0352, ...

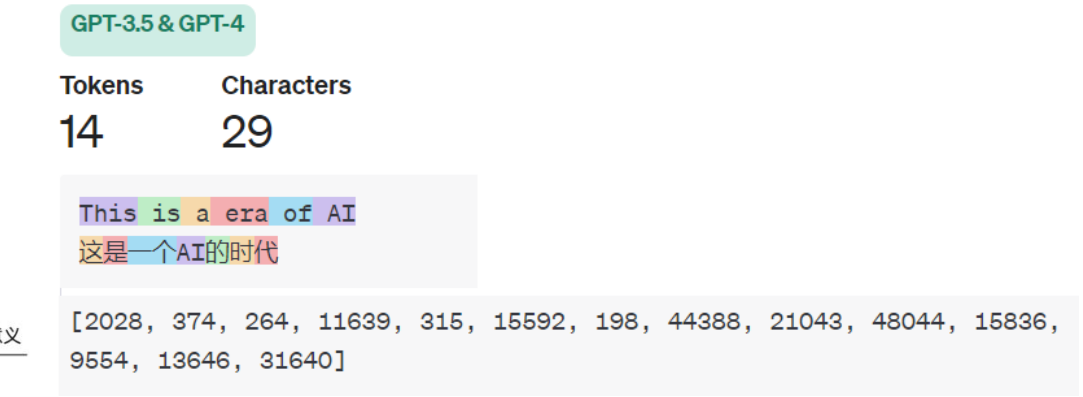


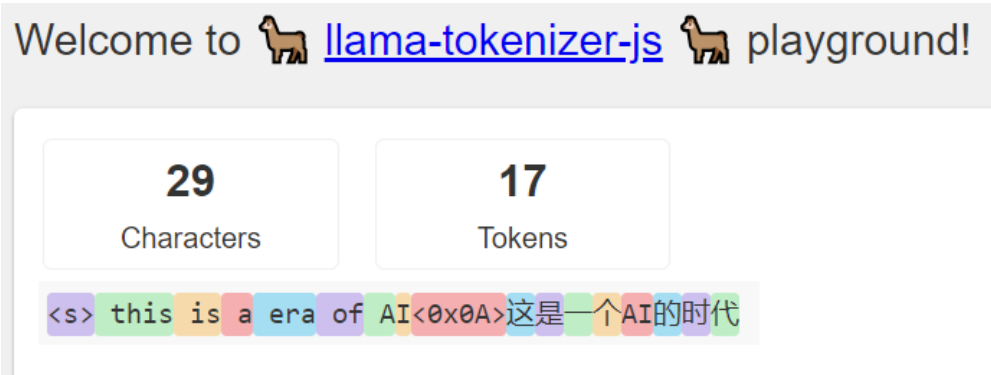
tokenization and embedding layer for transformer

Tokenization



Subword 分词方法	典型模型
BPE/BBPE	GPT, GPT-2, GPT-J, GPT-Neo, RoBERTa, BART, LLaMA, ChatGLM-6B, Baichuan
WordPiece	BERT, DistilBERT，MobileBERT
Unigram	AlBERT, T5, mBART, XLNet





GPT tokenizer

https://platform.openai.com/tokenizer

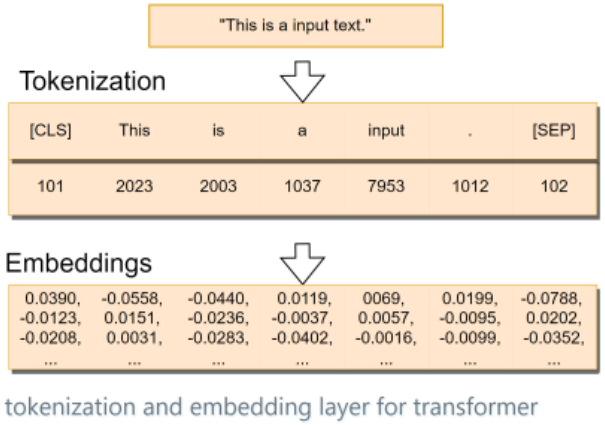
llama2 tokenizer

https://belladoreai.github.io/llama-tokenizer-js/example-demo/build/

Embedding

Why

- 将token映射到向量空间
- 在不同维度表示字/词的位置
- Attention根据每个token的vector计算token和token之间的距离（相关性）



tokenization and embedding layer for transformer

How

- 基于统计的方法

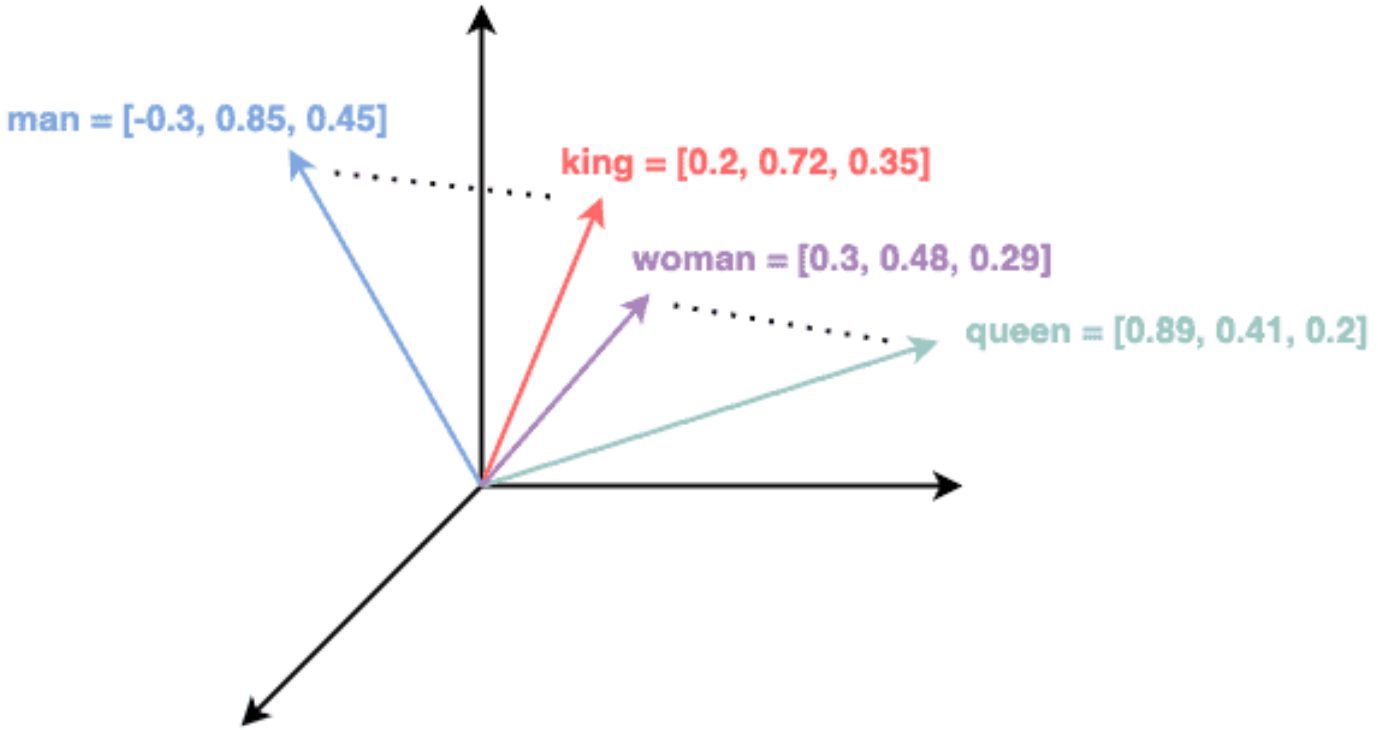
  TF-IDF, N-gram

- 基于神经网络的方法

  词嵌入(word3Vec, Glove)

  句子嵌入(RNN—LSTM GRU)

  文档嵌入(Doc2Vec, BERT)



Position embedding

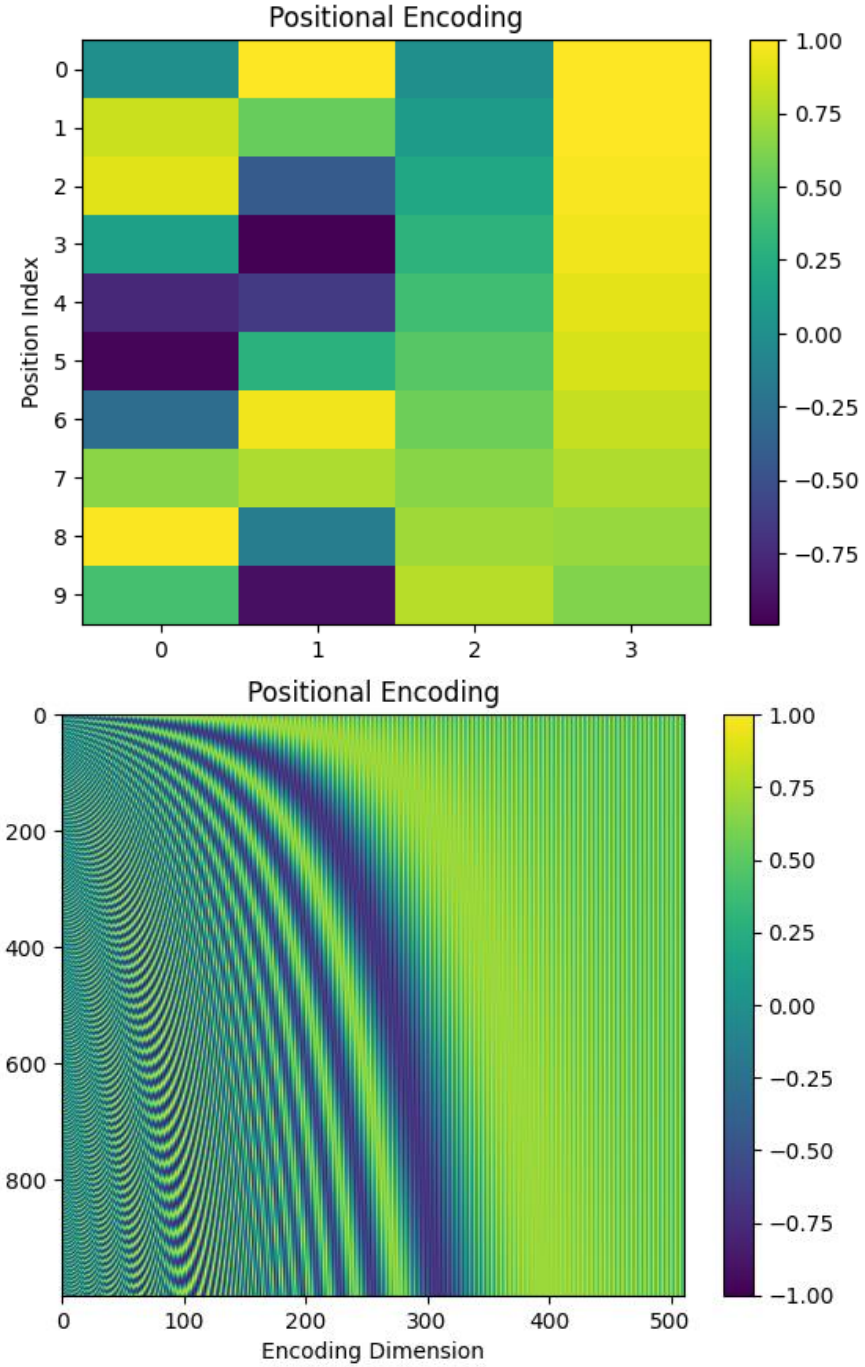
why

- 为序列中的每个token加入位置编码，向模型提供序列中元素的位置信息.
- 给模型提供序列的长度信息
- 模型可以看到文字之间的"距离“

How

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$



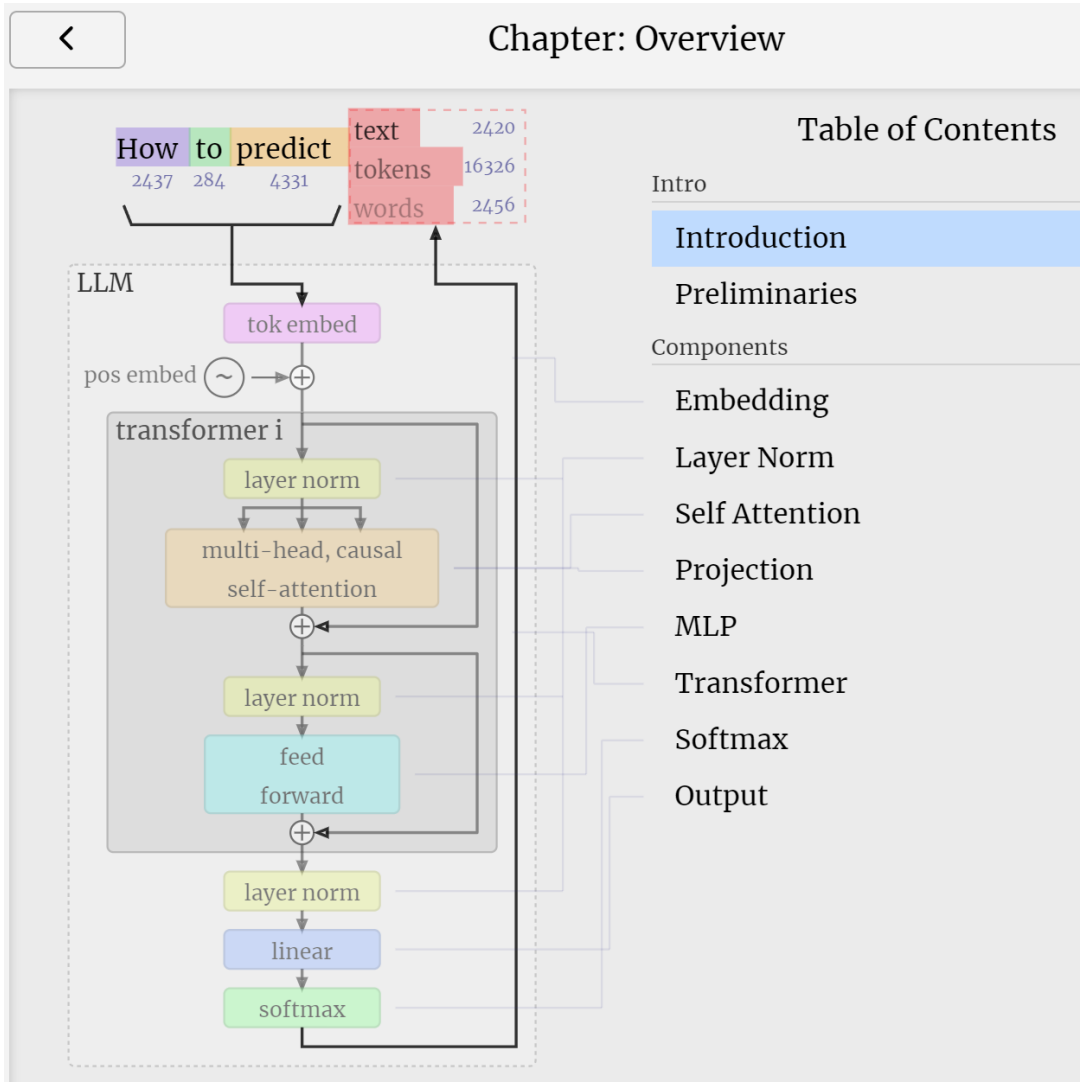
Feed Forward（neutral network)

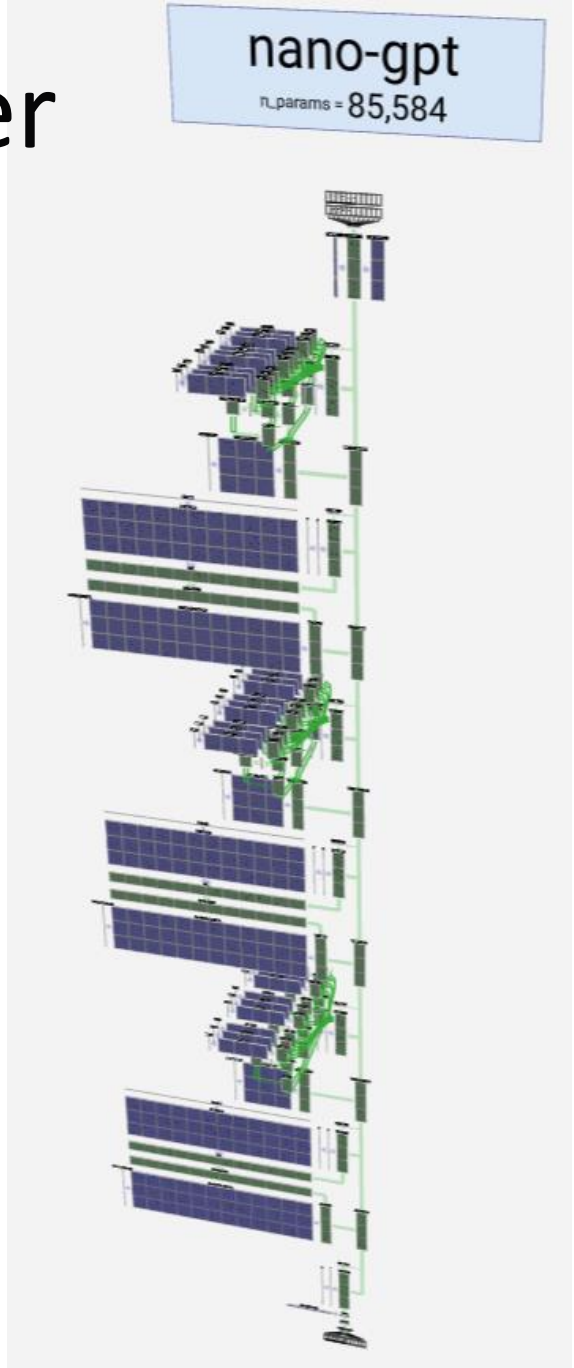
- 全连接网络，两层线性变换和一个 ReLU 激活函数
- 负责对输入序列的每个位置的隐藏表示进行非线性变换，这些隐藏表示不仅包含了当前位置的词嵌入信息，还蕴含了该位置周围的上下文信息
- 使模型能够捕捉输入序列中的非线性关系，提高模型对复杂模式的理解和预测能力
- 帮助助模型在处大量数据时保持高效和准确

$$\mathrm{FFN}(x) = \mathrm{ReLU}(xW_1 + b_1)W_2 + b_2$$

GPT-Generative Pre-trained Transformer

- GPT实现，使用的transformer的解码部分
- Scaling law—transformer架构支持扩展， 同时增加模型的规模和训练的数据量能够快速提升模型的能力





https://bbycroft.net/llm

LLM three stage

1 模型预训练

2 微调 fine-tune

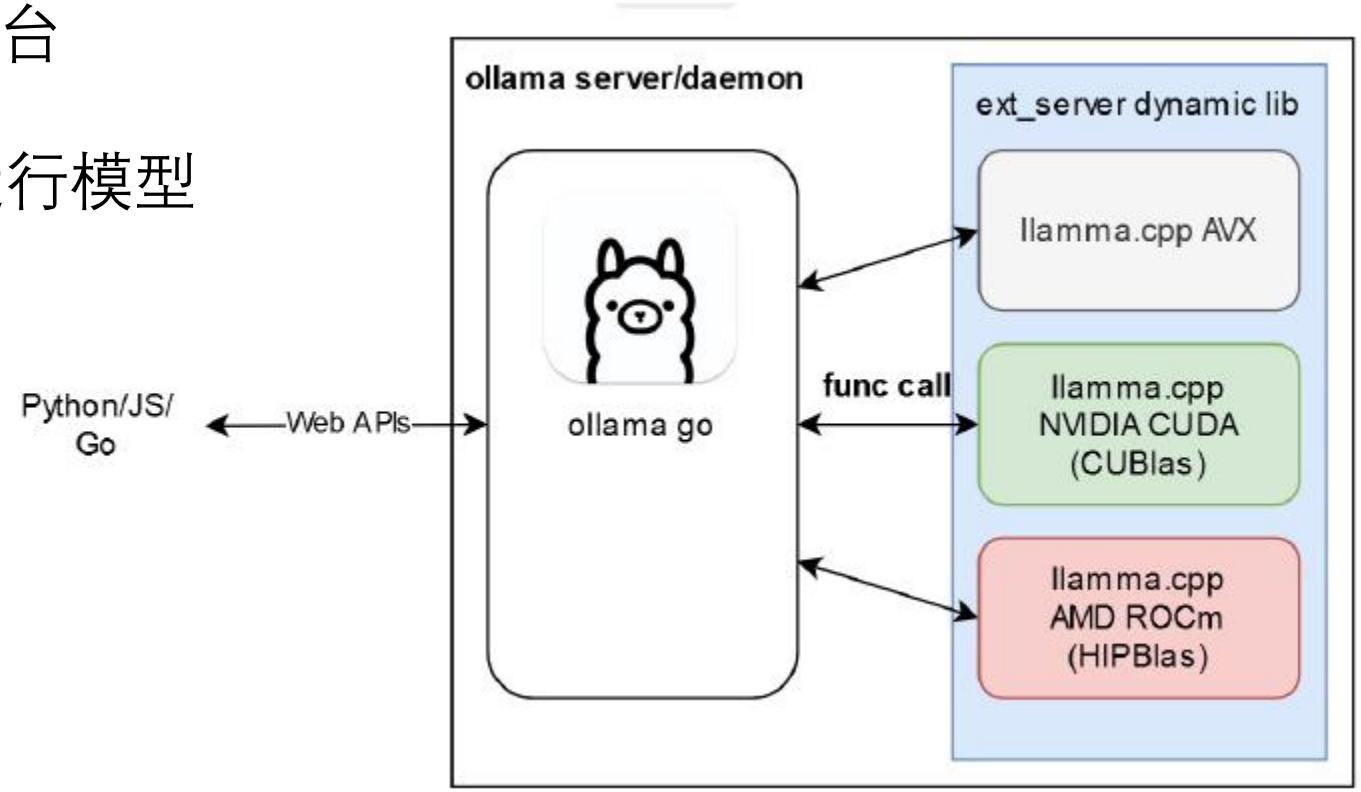
3 RLHF, reinforcement learning human feedback

LLM related software framework & tools

- Ollama
- Llama-index & Langchain
- Huggingface

Ollama

- 一个大模型的管理/运行/部署平台
- 使用类似于docker的方法管理/运行模型
- 支持 python/JS/Go语言开发
- 支持本地和远程调用

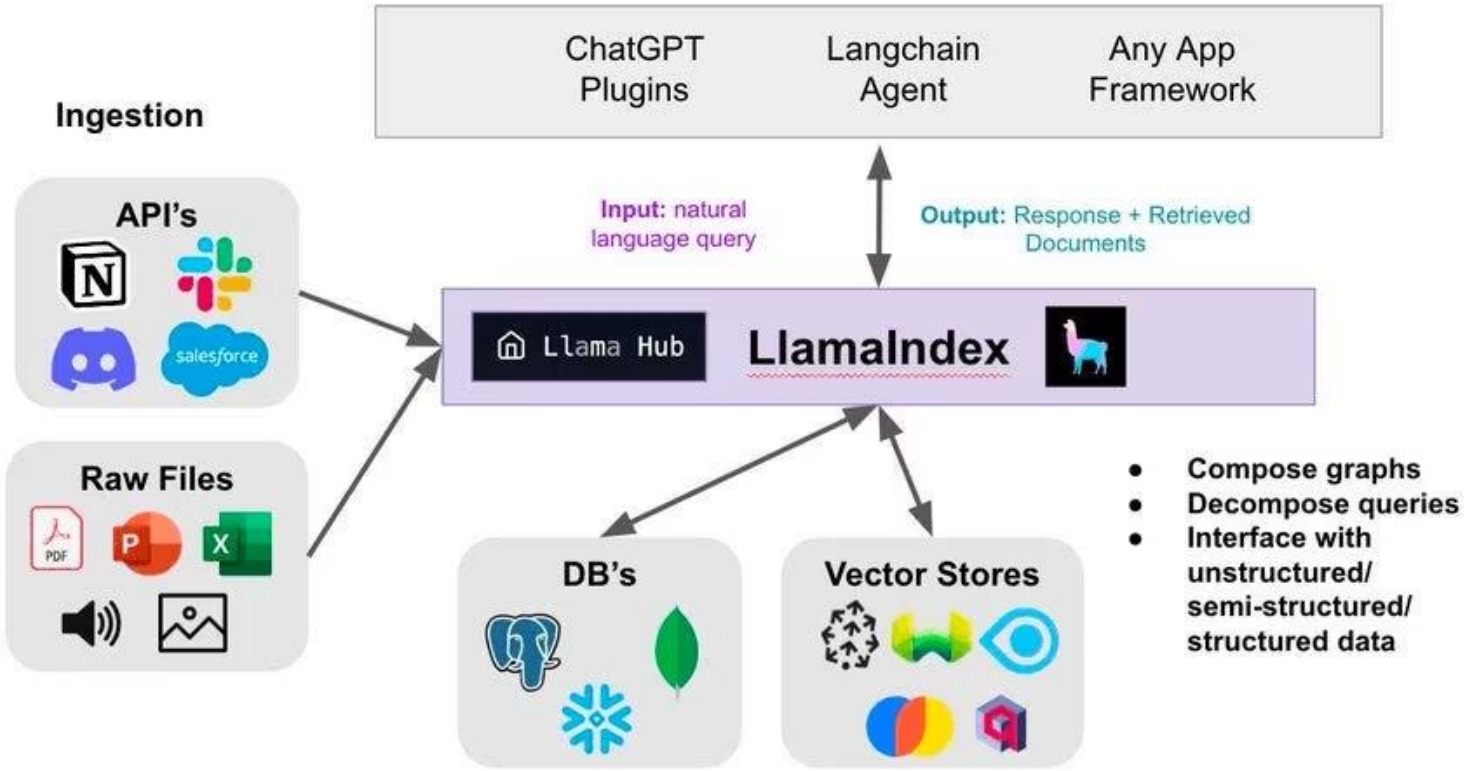


https://ollama.com/

https://github.com/ollama/ollama

Llamaindex

- 开源数据框架
- 为大模型提供数据服务
- 提供应用的数据流控制



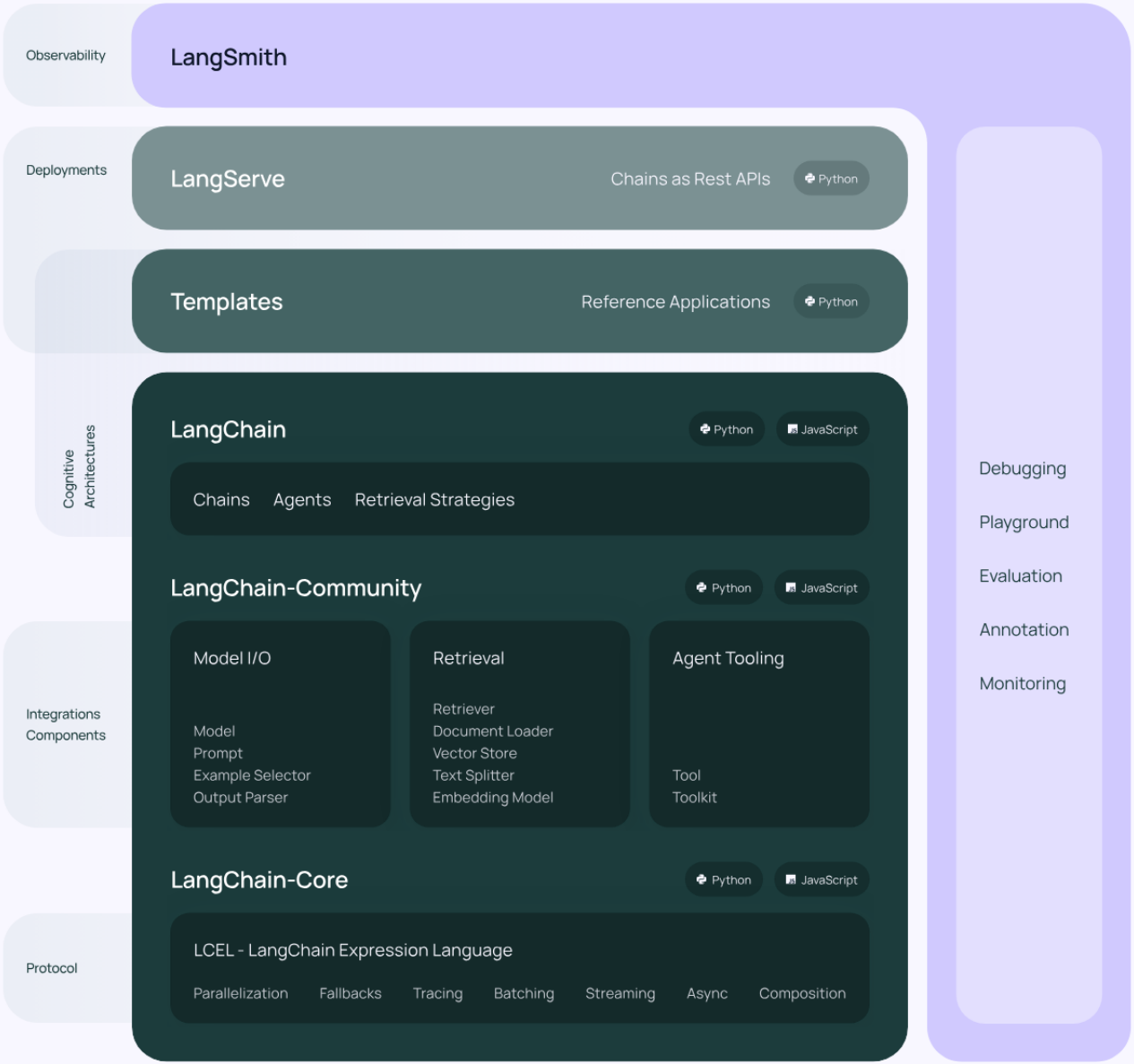
https://www.llamaindex.ai/

Langchain

开源数据框架，为大模型提供数据服务

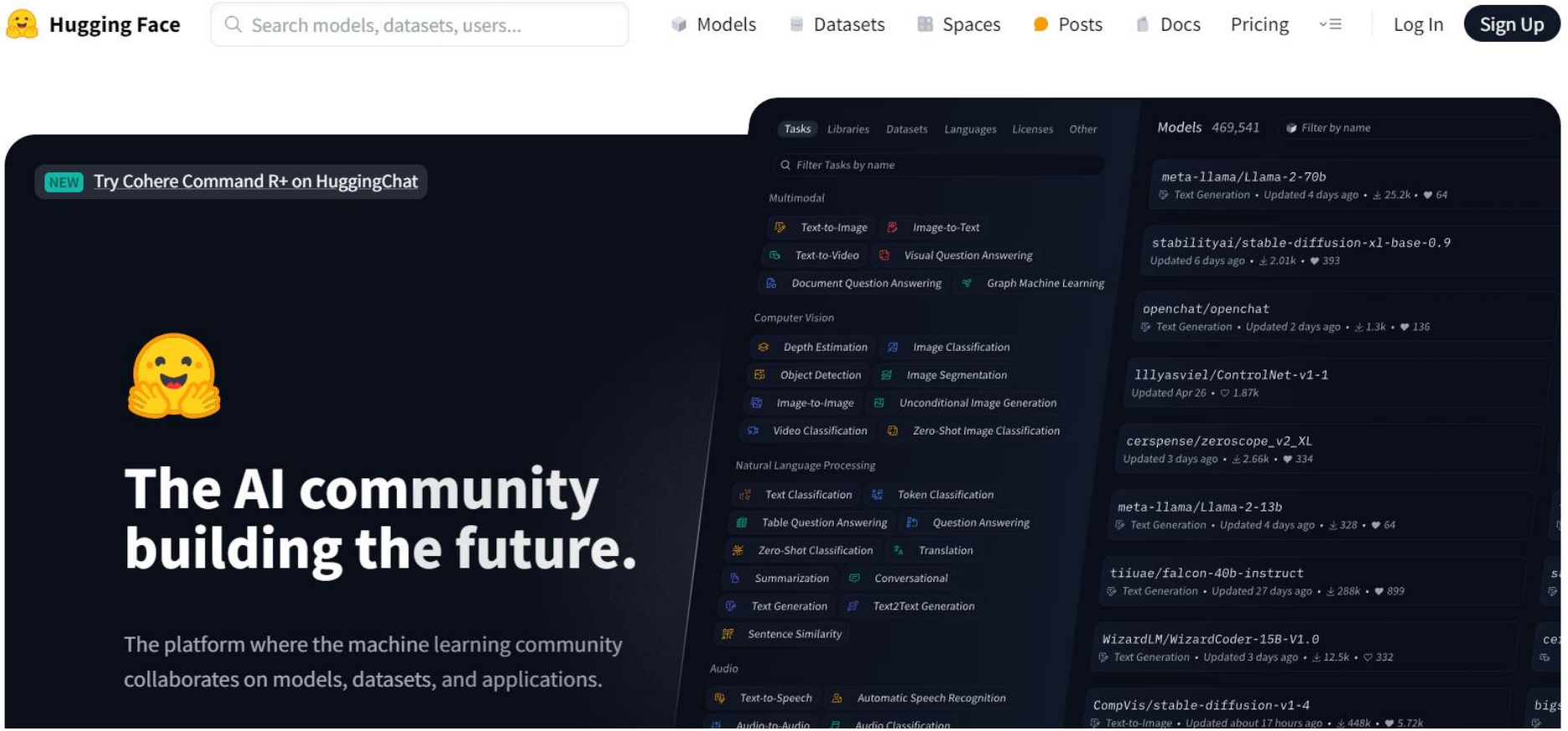
- langchain-core：基础抽象和 LangChain 表达式语言。
- langchain-community：第三方集成。
- langchain：链、代理和检索策略，构成应用程序的认知架构。
- langgraph：使用 LLMs 构建稳定且有状态的多因素应用程序，通过将步骤建模为图中的边和节点。
- langserve：将 LangChain 链部署为 REST API。 broader ecosystem 包括 LangSmith、LangGraph 和 LangServe。
- langSmith 可视化执行观察

https://python.langchain.com



Huggingface

Huggingface 一个社区（发布信息，讨论），托管各种大语言模型，应用和库，应用展示，数据集和任务



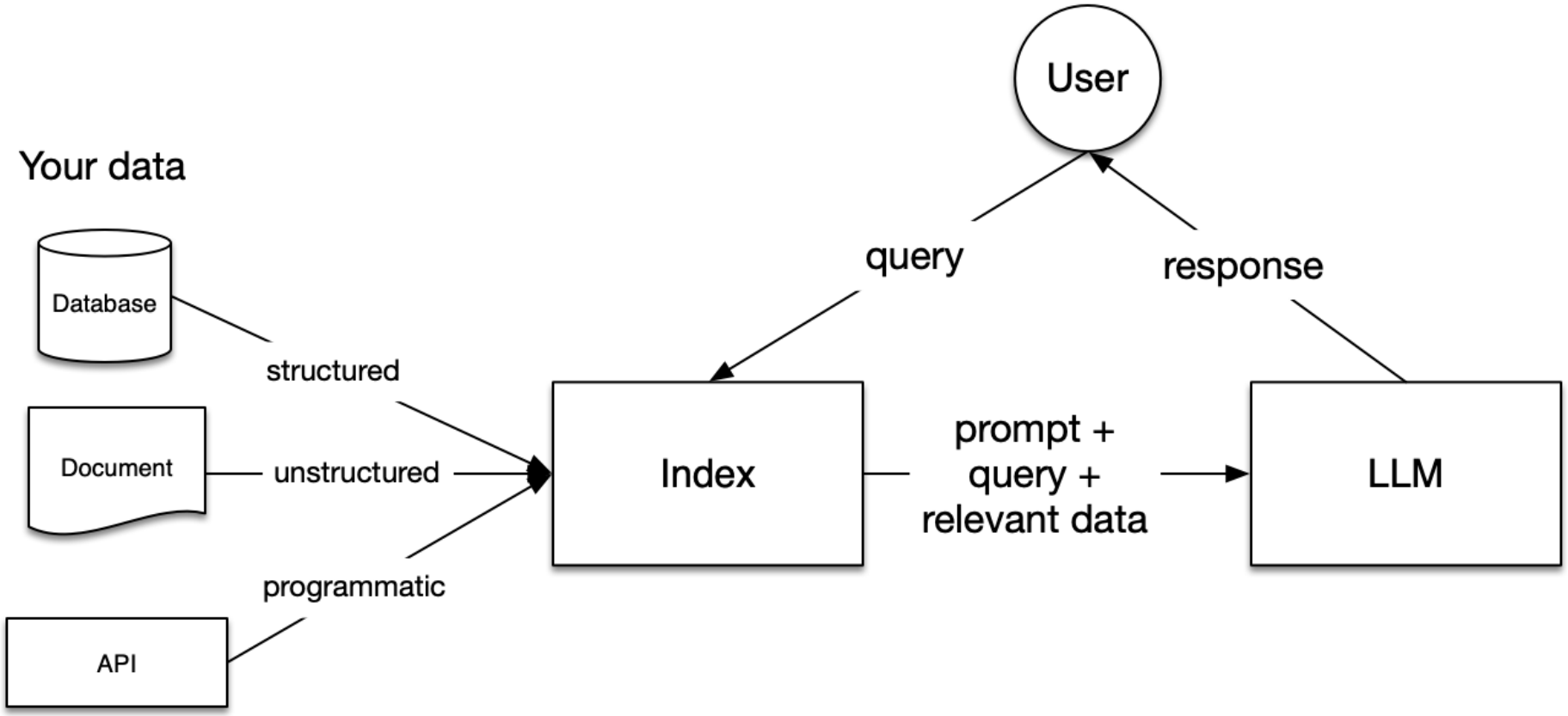
https://huggingface.co

LLM application

- Chat Q/A
- Search
- RAG
- AGENT

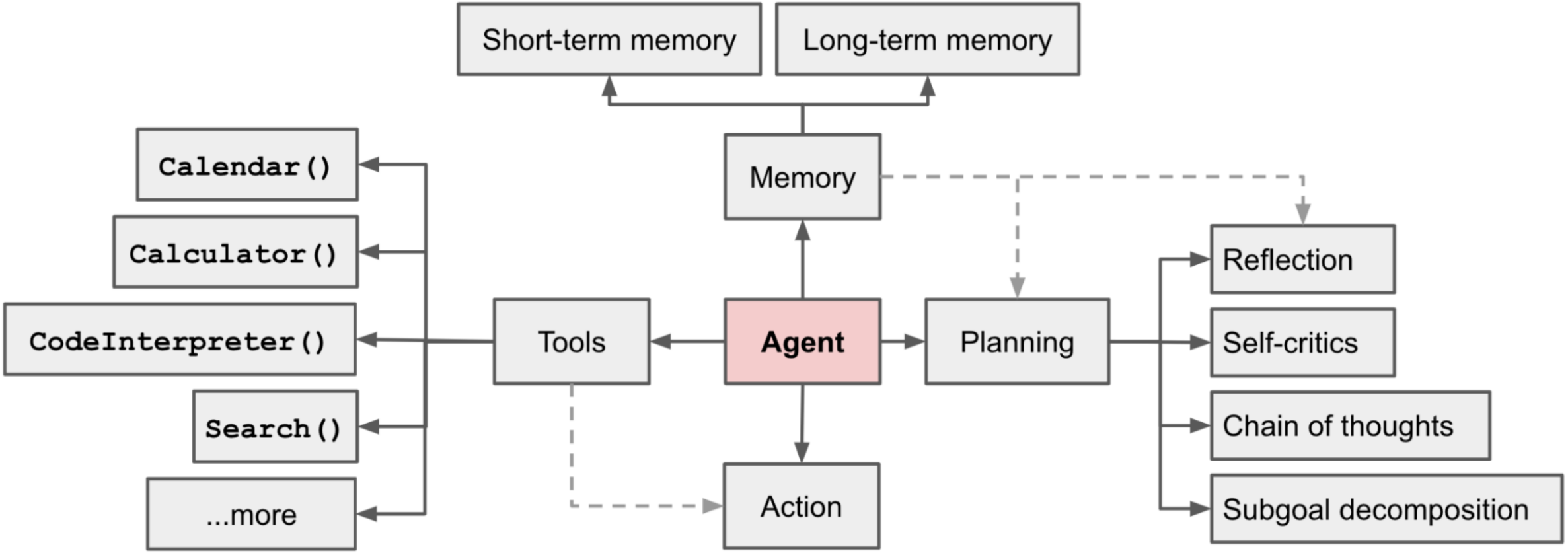
- ….

RAG-Retrieval-Augmented Generation



https://github.com/dibaotian/ma35_rag

AGENT



https://github.com/dibaotian/llm_agent/blob/master/agent.ipynb

THANKS