Computing Center, IHEP, CAS
National HEP Data Center

通用大模型与领域大模型的发展研究

王虹

中国科学院高能物理研究所计算中心

ON BEHALF OF HEPAI GROUP

第二十一届全国科学计算与信息化会议暨第十二届科研信息化联盟会议

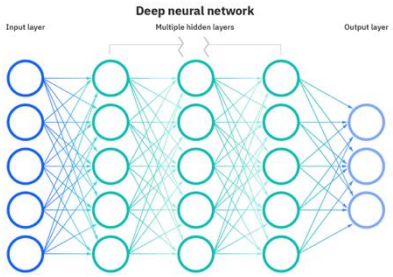
2025.08.27 中国 长春

目录

- 背景介绍
- Transformer与MoE框架
- 通用大模型
- 领域大模型
- HepAI模型进展
- 总结

背景介绍

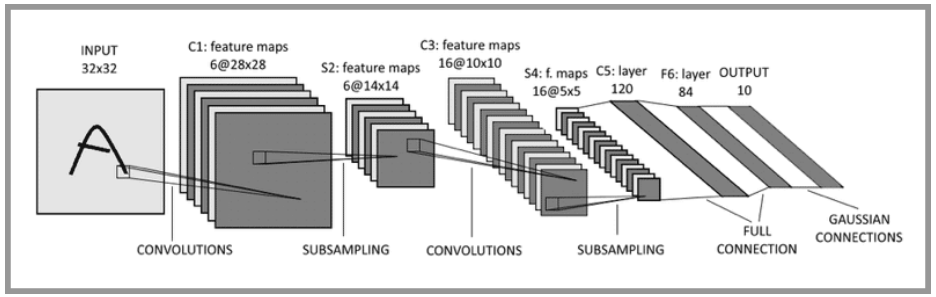
神经网络的架构对比与数据适用性



全连接神经网络（Fully Connected Network, FCN）

针对定长向量形式数据

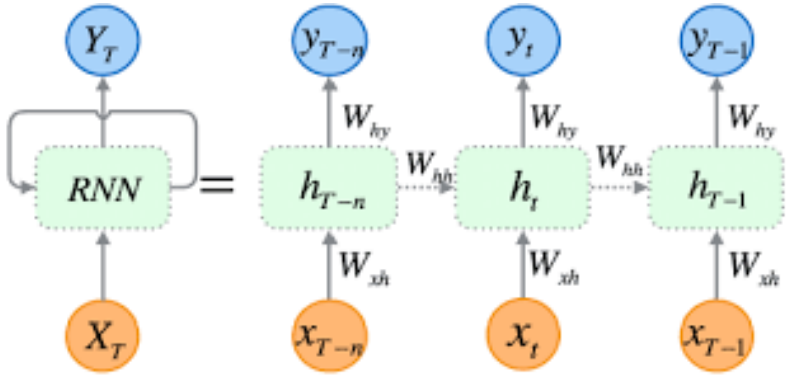
无数据结构信息（局部空间相关性、文本序列依赖）



卷积神经网络（Convolution Neural Network, CNN）

针对具有局部相关性和空间结构的数据

用于图像分类、目标检测等



循环神经网络（Recurrent Neural Network, RNN）

处理序列数据的神经网络

在序列的每一个时刻t，不仅接受当前的输入，还会接受上一个时刻的隐藏状态，从而具有记忆能力

特别适合处理自然语言、时间序列

局限性：

顺序计算，无法并行

难以捕获长距离依赖

难以扩展到长序列

解释性差

LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (2002): 2278-2324.

Ahmed, M.A., et al. "Using machine learning for the classification of the modern arabic poetry." *TELKOMNIKA Telecommunication Computing Electronics and Control* 17.5 (2019): 2667-2674.

Leifer, E. "Image classification of stars and galaxies using different machine learning models (2023)."

- 背景介绍
- Transformer与MoE框架
- 通用大模型
- 领域大模型
- HepAI模型进展
- 总结

Transformer

自然语言处理模型Transformer

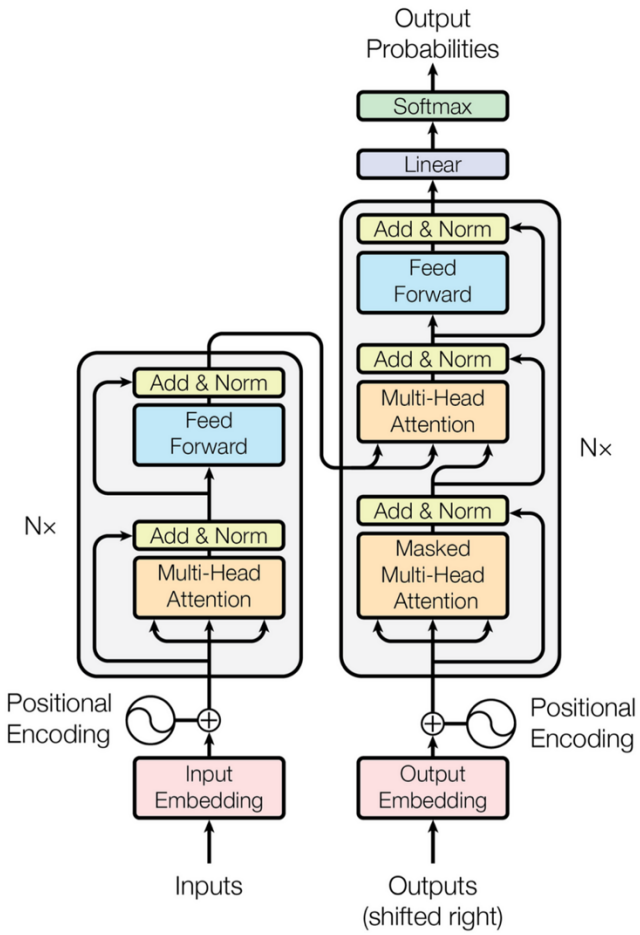
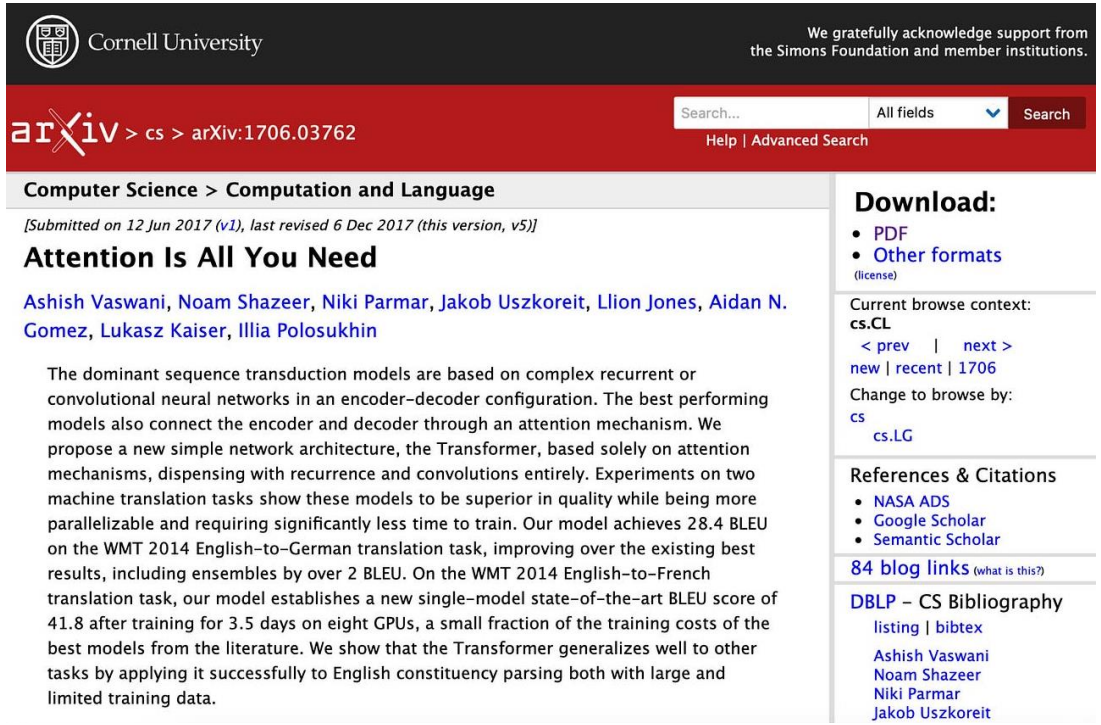


Figure 1: The Transformer - model architecture.

Transformer模型框架

2017 年，由论文 "Attention Is All You Need" 提出

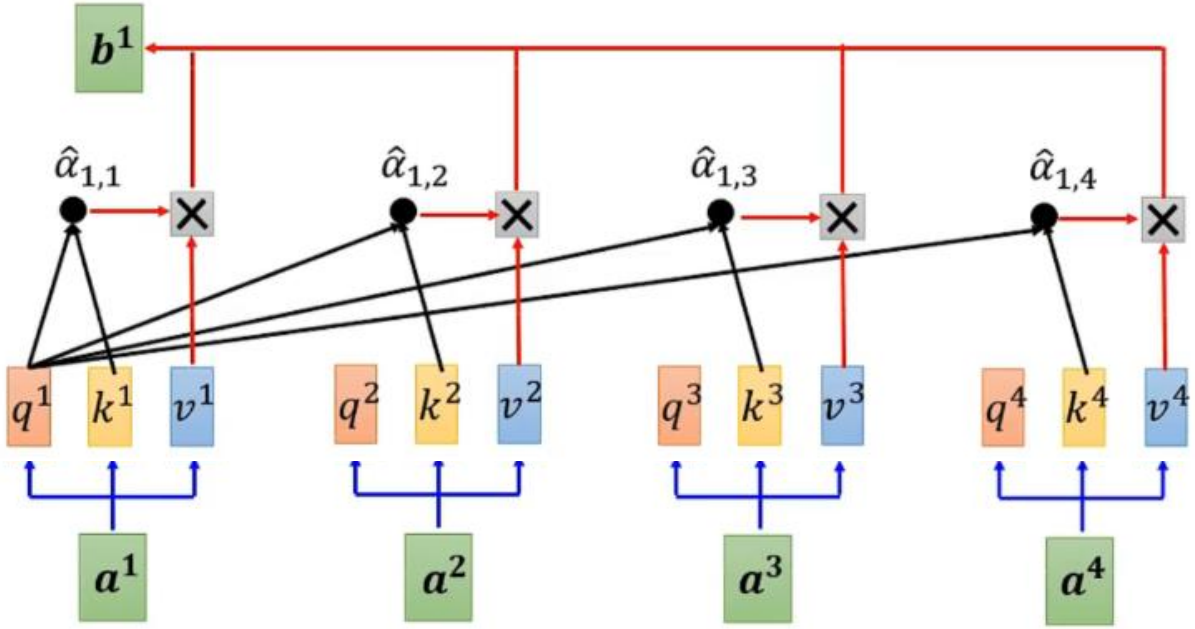
Attention：基于RNN架构被提出
Self-Attention：将注意力机制独立成模块



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

Self-Attention的实现方法——注意力模块




```
Many words map to one token, but some don't: indivisible.

Unicode characters like emojis may be split into many tokens containing
 the underlying bytes: ������

Sequences of characters commonly found next to each other may be grouped
 together: 1234567890
```

- 线性映射：词嵌入与三个权重矩阵W_q， W_k，W_v相乘
  - 生成Query、Key、Value 向量

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

- 通过 Q与 K的点积计算注意力分数（Scaled Dot-Product）
  - 对第一个词计算自注意力向量，用输入句子中的每个单词对第一个词打分。这些分数决定了每个单词对编码当下位置的贡献。
- 不需要等前序向量的结果，可以实现并行

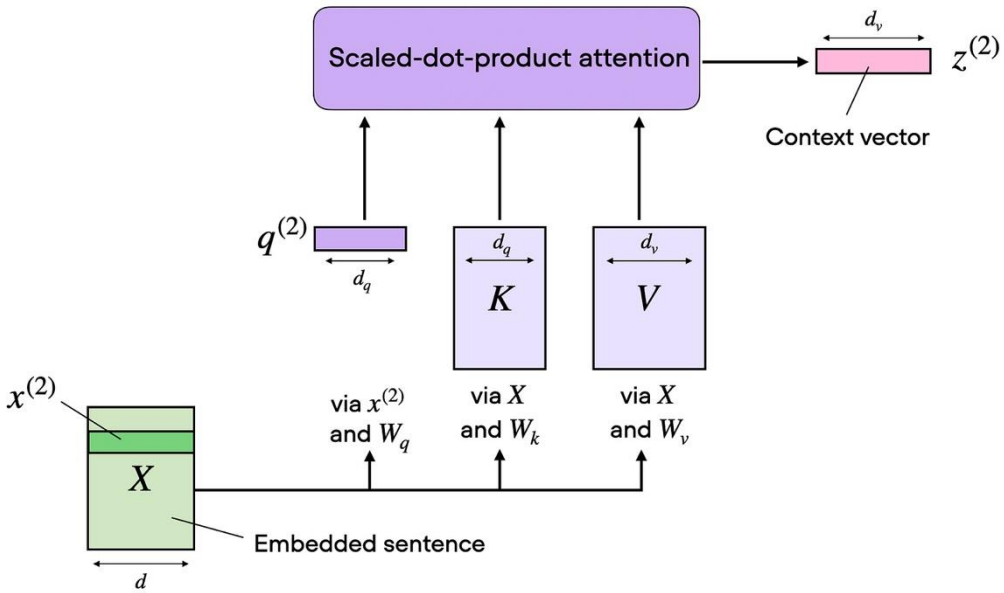
从大型语言模型（LLM）的角度来看，一段话是由多个tokens组成的。

Wang, Haiyao, et al. "A novel hybrid model of CNN-SA-NGU for silver closing price prediction." Processes 11.3 (2023): 862.

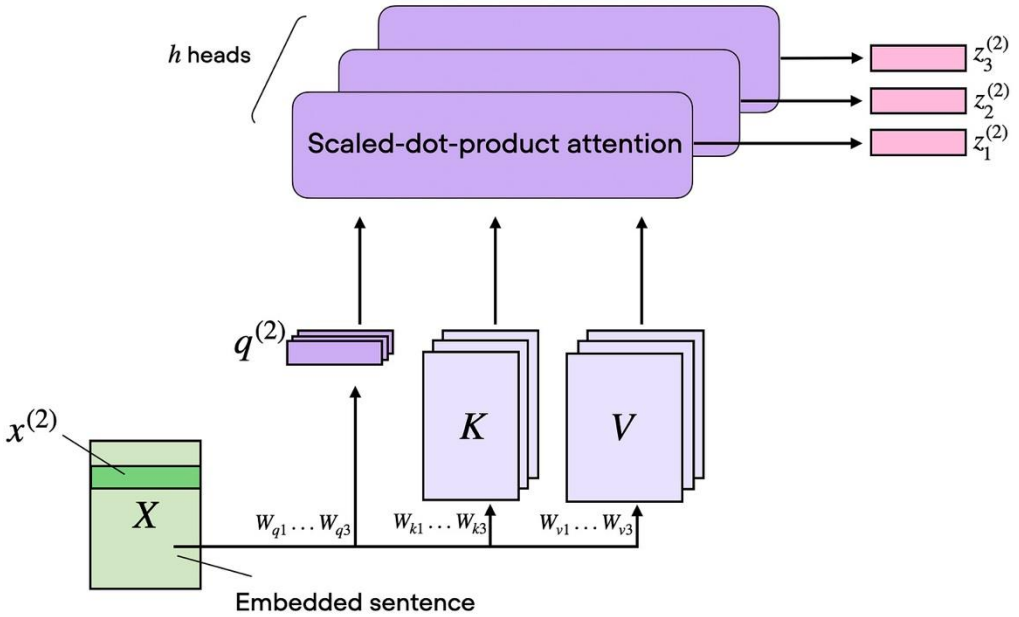
Transformer

Multi-head attention：多头注意力机制

- 每个head使用不同的W_q， W_k，W_v，将所有头的输出拼接并再线性映射输出
- 不同heads同时关注不同语义关系，捕获更丰富的上下文



self-attention mechanism



Multi-head attention mechanism

捕获多样语义视角

多头注意力使得模型能够从不同的角度并行关注输入的不同部分，从而更全面地理解上下文关系

https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Transformer

并行处理自然语言的Transformer模型框架

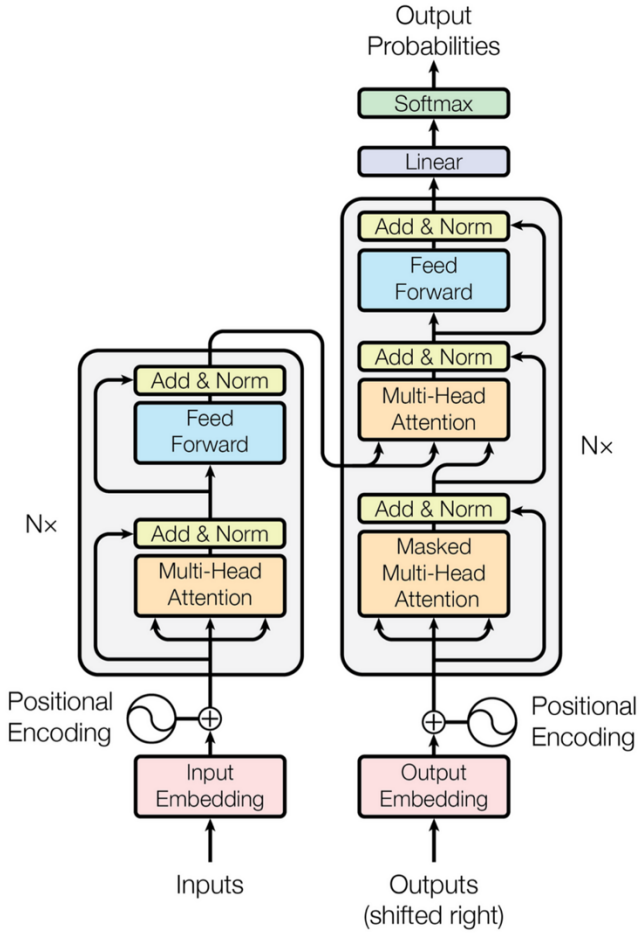


Figure 1: The Transformer - model architecture.

Transformer模型框架

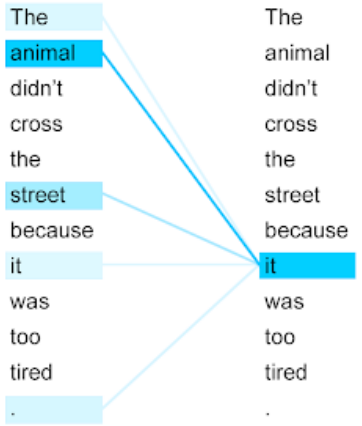
Encoder–Decoder 架构

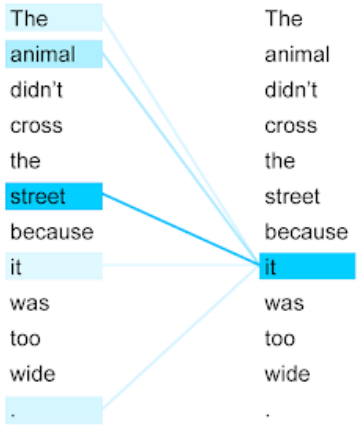
- 文本 → tokens → 词向量
- 位置编码：transformer没有循环结构和卷积操作，让模型能够利用序列的顺序信息
- 词向量与位置编码组成输入向量
- 编码器Encoder：6层，每层有两个子层，第一个子层为multi-head attention，第二个子层为前馈神经网络
- 解码器Decoder：6层，每层有三个子层，其中一个子层对编码器的输出执行多头注意力

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

Transformer的上下文理解能力





编码器对单词“it”从第5层到第6层进行英法翻译训练的Transformer的自注意分布

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

Transformer和Self-attention框架的优势

1. 选择性聚焦，提升信息处理效率
注意力机制能让模型聚焦在输入的关键部分

2. 更好地捕捉长距离依赖关系
注意力机制可以让模型直接访问序列中任意位置的内容

3. 并行计算能力强
注意力机制可以在单步中处理整个序列，支持GPU并行运算

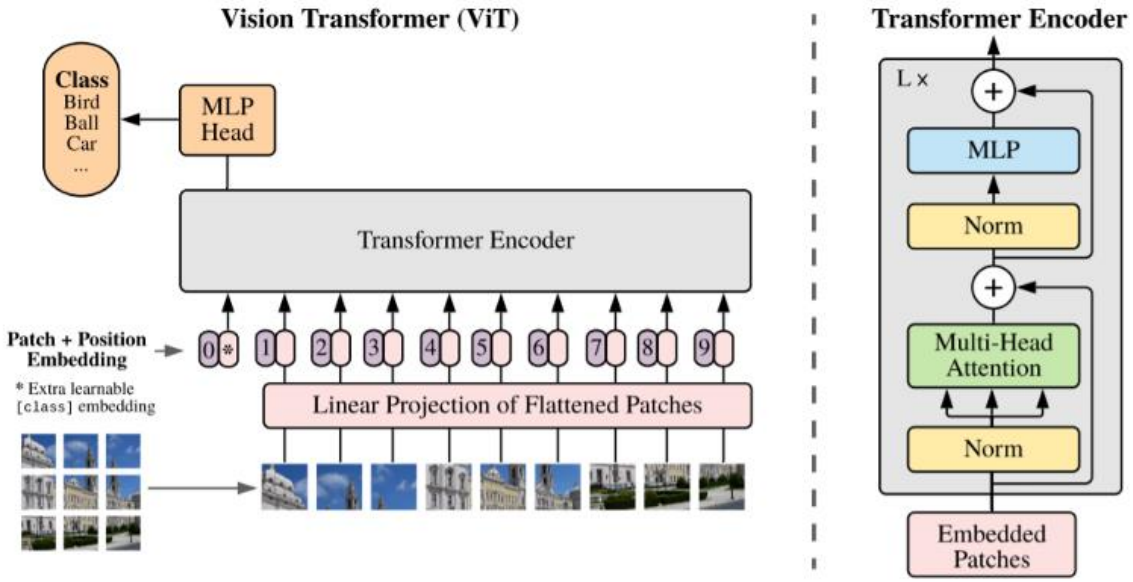
4. 可解释性强
通过可视化attention权重，直观了解模型关注了哪些部分

5.模块化，易扩展
由多个编码器和解码器层堆叠而成

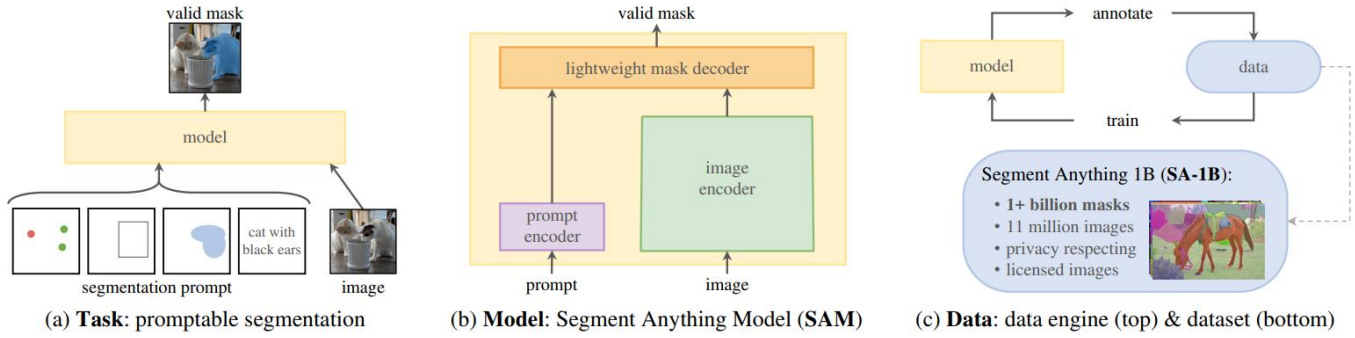
Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

Transformer用于图像数据



Vision Transformer模型框架



Segment Anything Model框架

Vision Transformer：

Segment Anything Model的backbone模型

Vision Transformer流程：

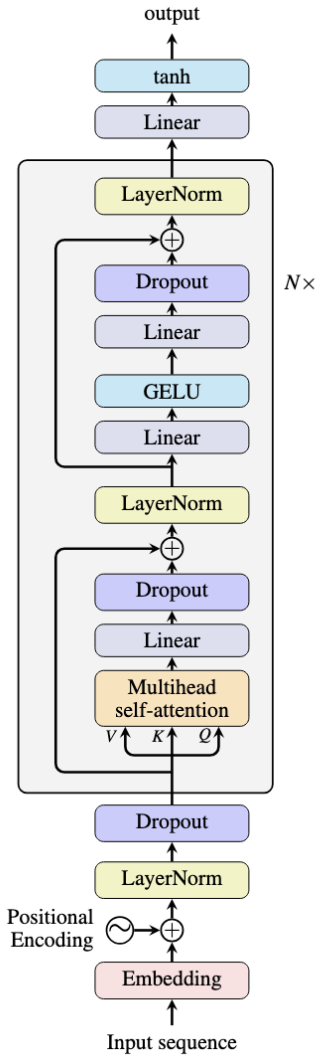
- 将图像切分为固定大小的小patch，并flatten成向量
- 每个向量通过线性层映射成嵌入，再加上位置编码表示空间位置信息
- 经编码器处理，最终得到全局图像表示
- 在分类任务中，通常加一个专用的[CLS] token来代表整个图像，接一个MLP输出分类结果。

SAM（Segment Anything Model）

- 图像分割模型
- 图像编码器使用ViT
- 支持零样本泛化

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
Kirillov, Alexander, et al. "Segment anything." *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.

基于Transformer的语言模型



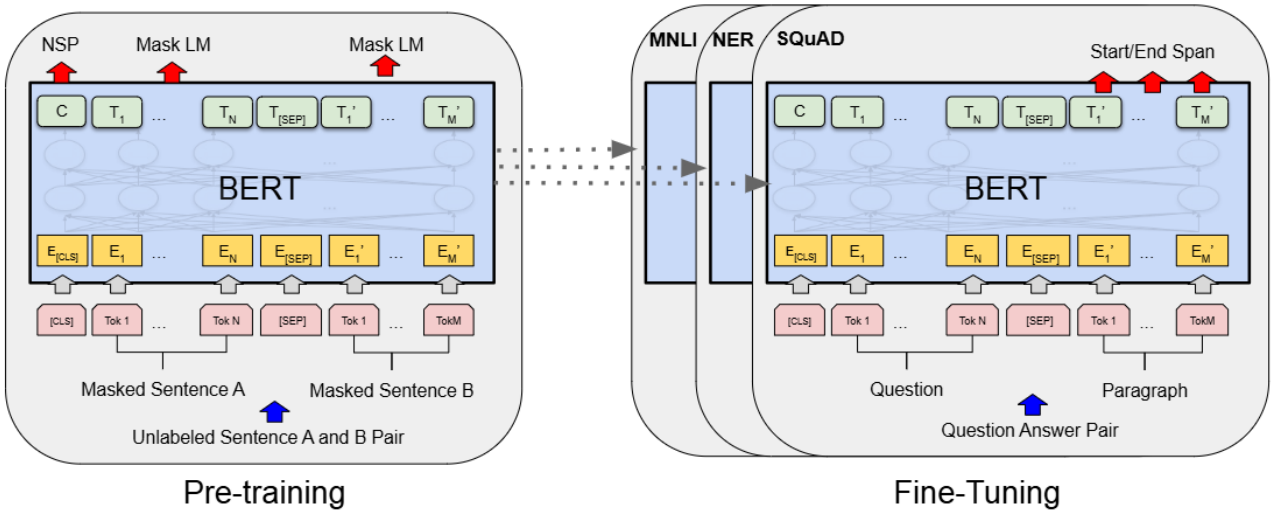
Bert模型框架

基于Transformer的语言模型架构：

- Encoder-decoder
- Encoder-only
- Decoder-only

Encoder-only

BERT: Bidirectional Encoder Representations from Transformers



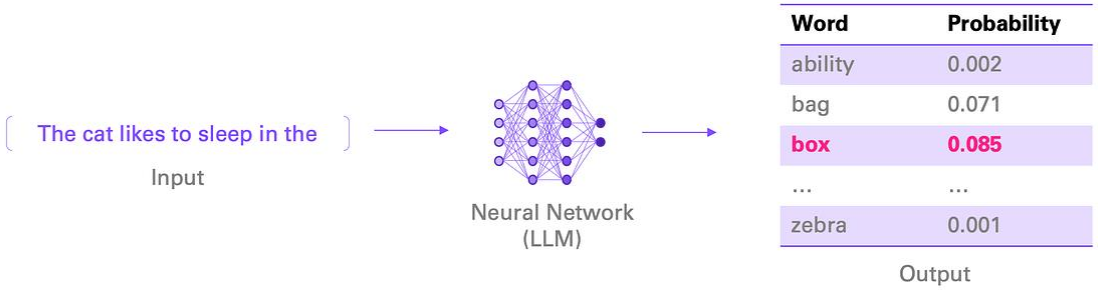
- 双向编码：同时考虑上下文信息，提升语义理解能力。
- 通用性强：适用于理解任务，如文本分类、命名实体识别等。
- 生成能力有限：对于文本生成任务的表现不如 GPT 等模型。

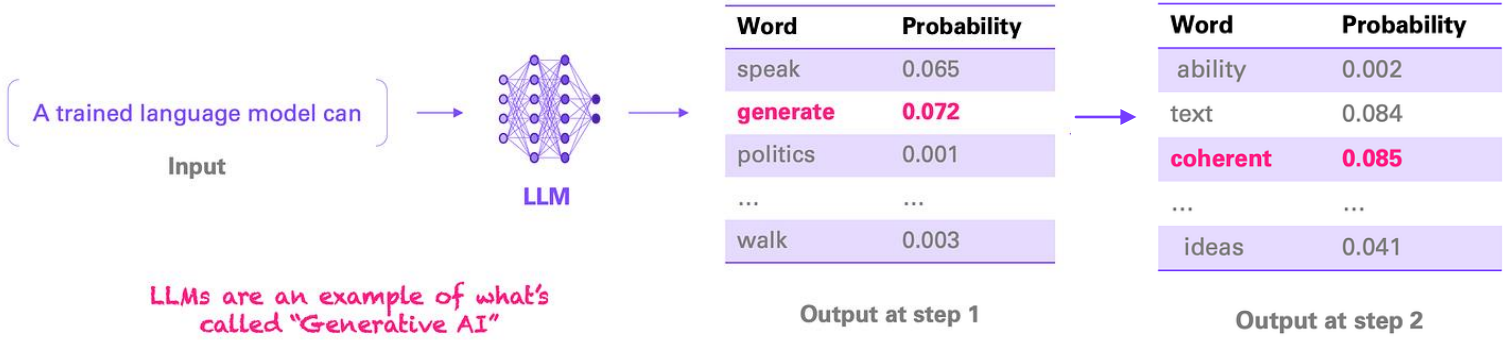
https://machinelearningmastery.com/encoders-and-decoders-in-transformer-models/

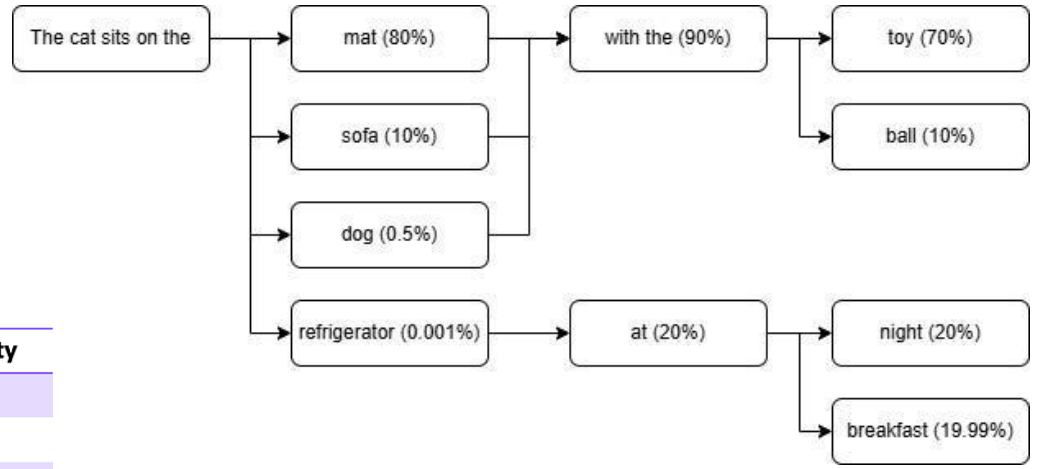
基于Transformer的语言模型

自回归语言模型：预测下一个词

- 给定一段文字前缀，让模型延续接下来的内容。



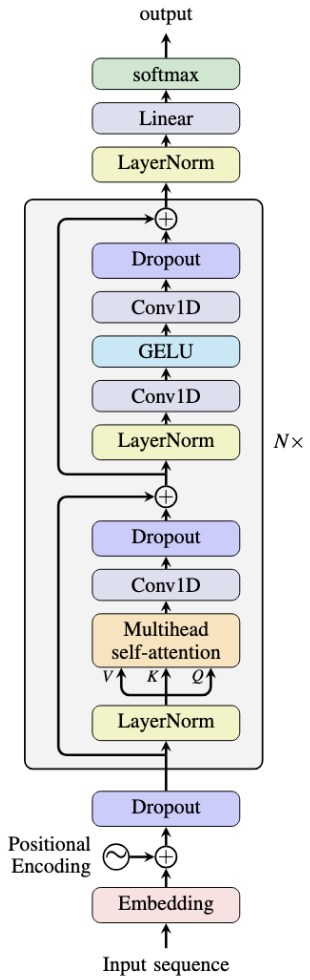




模型从前几个选项中抽取样本，增加多样性

https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f
https://medium.com/@sybrandwildeboer/how-large-language-models-predict-the-next-word-and-why-thats-powerful-b1ac78995e74

基于Transformer的语言模型



Decoder-only模型

Decoder-only

1. 嵌入向量+位置编码
2. 多层Decoder-only Transformer模块
3. 输出层预测下一个token

Decoder-only模型的优势：

- 简化了模型结构，更易扩展到超大规模
- 多任务的统一处理
- 通过prompt+自回归生成来完成不同任务
- 生成可以是任意长度，不必预设输出长度
- 适合开放式生成

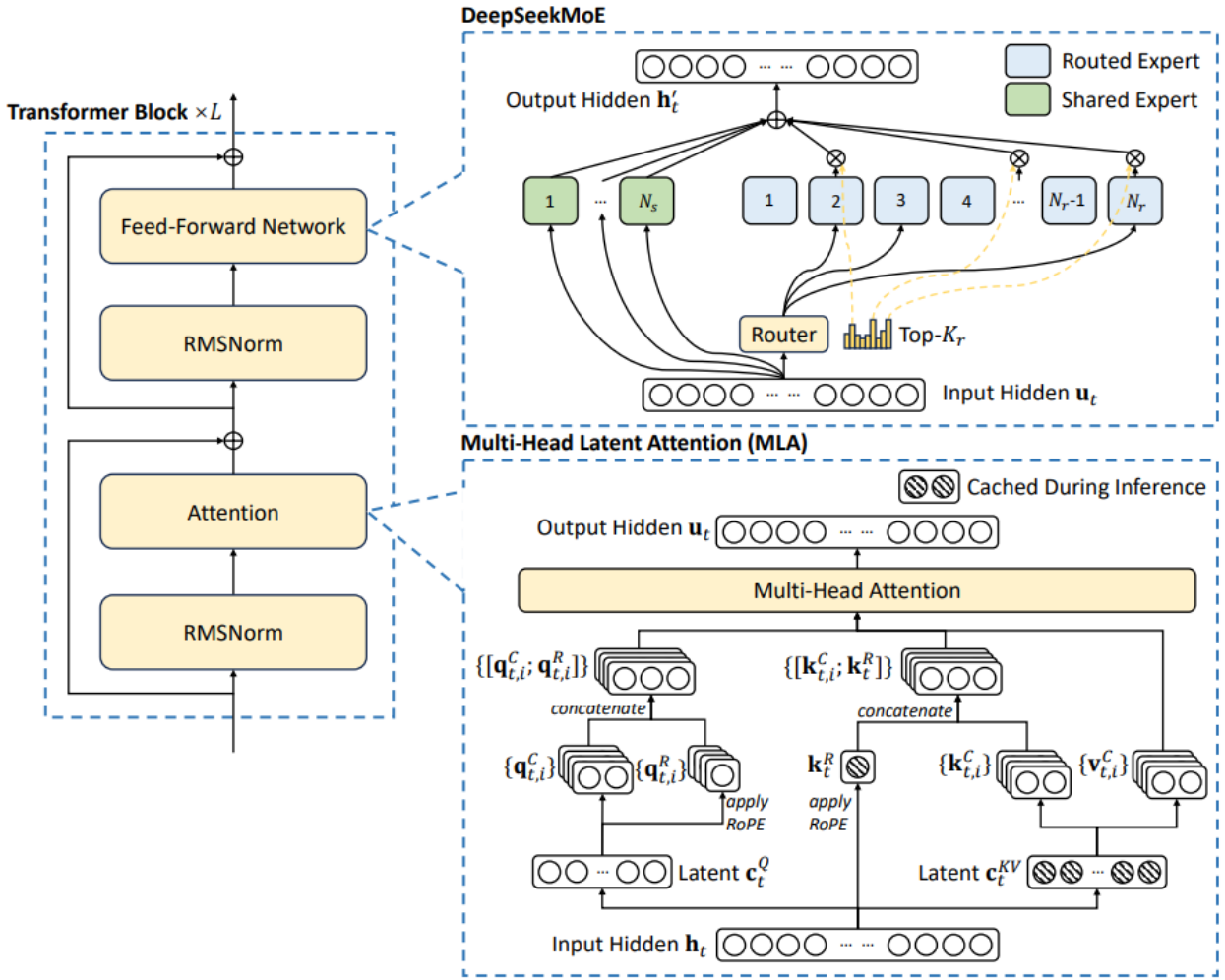
GPT-1 (2018年)：
12个transformer模块，每层12个attention heads
参数量：0.117B
训练数据：BookCorpus (约7000本书)
预训练+微调模式

GPT-2 (2019年)：
参数量：1.5B
训练数据：WebText (大量网页)
零样本，无需特定任务训练

GPT-3 (2020年)：
参数量：175B
训练数据：CommonCrawl、WebText、维基百科、书籍语料等
参数大量增加，能生成高度复杂和上下文感知的文本

https://machinelearningmastery.com/encoders-and-decoders-in-transformer-models/

混合专家模型（Mixture of Experts, MoE）



混合专家模型框架

混合专家模型将整个网络拆分为多个专门化的子网络，并通过一个门控网络来动态决定哪些专家参与每次推理

- 每个输入由门控网络计算一组专家亲和分数
- 只激活得分最高的若干个专家来处理该输入
- 未被选中的专家则不参与计算
- 形成了稀疏激活，大幅降低了计算成本

DeepSeekMoE将专家分为共享专家和路由专家

$$\mathbf{h}'_t = \mathbf{u}_t + \sum_{i=1}^{N_s} \mathrm{FFN}_i^{(s)}(\mathbf{u}_t) + \sum_{i=1}^{N_r} g_{i,t}\,\mathrm{FFN}_i^{(r)}(\mathbf{u}_t),$$

$$g_{i,t} = \frac{g'_{i,t}}{\sum_{j=1}^{N_r} g'_{j,t}},$$

$$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \mathrm{Topk}(\{s_{j,t} | 1 \leqslant j \leqslant N_r\}, K_r), \\ 0, & \text{otherwise}, \end{cases}$$

$$s_{i,t} = \mathrm{Sigmoid}\left(\mathbf{u}_t{}^T \mathbf{e}_i\right),$$

Liu, Aixin, et al. "Deepseek-v3 technical report." *arXiv preprint arXiv:2412.19437* (2024).

- 背景介绍
- Transformer与MoE框架
- 通用大模型
- 领域大模型
- HepAI模型进展
- 总结

通用大模型

DeepSeek R1 系列（深度求索）

•DeepSeek-V3/R1

- 总参数量：671B，每次推理激活参数约37B
- 训练数据：约 14.8 T tokens
- 训练资源：使用2048 H800X两个月

•DeepSeek-V3.1

- 2025年8月25日发布

GPT系列（OpenAI）

GPT-4（估算）

•参数量：未公开

•训练数据：约 13T tokens+2T tokens

•训练资源：使用 25k A100 GPUX90–100 天

•GPT-5

•2025年8月7日发布

GLM 系列（清华、智谱AI）

•GLM-130B

- 参数量：130B

•GLM-4.5

- 参数量：355B，激活参数量：32B
- 训练数据：约23T tokens

LLaMA 系列（Meta）

•LLaMA 3/3.1

•参数量：8-405 B

•训练数据：约15T tokens

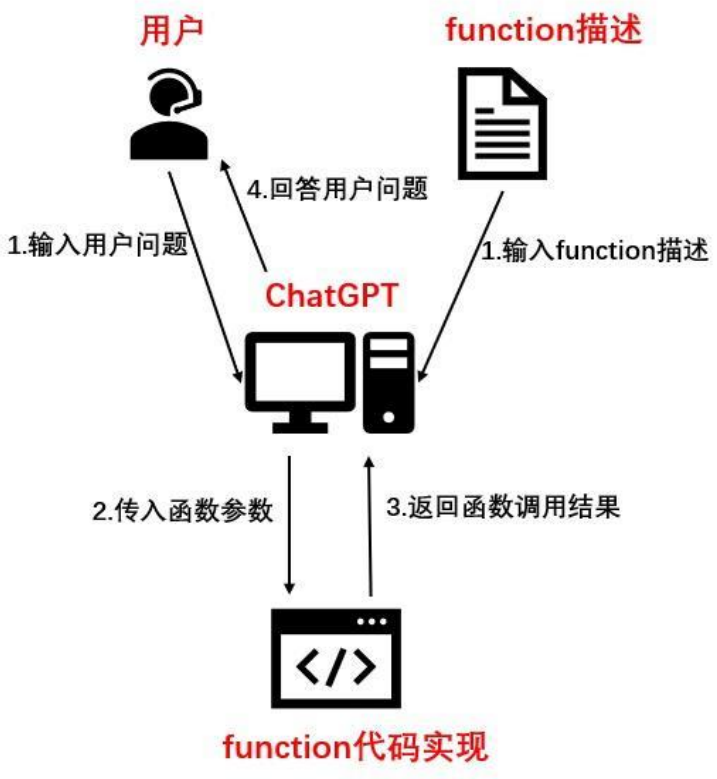
•LlaMA 4

•2025年4月5日发布

通用大模型

大模型的搜索功能

- Function call（工具调用）：当用户输入一个问题时，LLM通过文本分析是否需要调用某一个function，如果需要调用，那么LLM返回一个json，json包括需要调用的function名，需要输入到function的参数名，以及参数值。
- 输出调用指令 → 系统执行 → 返回结果 → LLM 整合。




```
func = {
        "name": "get_current_weather",
        "description": "获取今天的天气",
        "parameters": {
            "type": "object",
            "properties": {
                "location": {
                    "type": "string",
                    "description": "获取天气情况的城市或者国家，比如北京、东京、
                },
                "time": {
                    "type": "string",
                    "description": "时间信息"
                },

            },
            "required": ["location", "time"]
        }
    }
```

定义function

Q：8月27日长春天气怎么样?

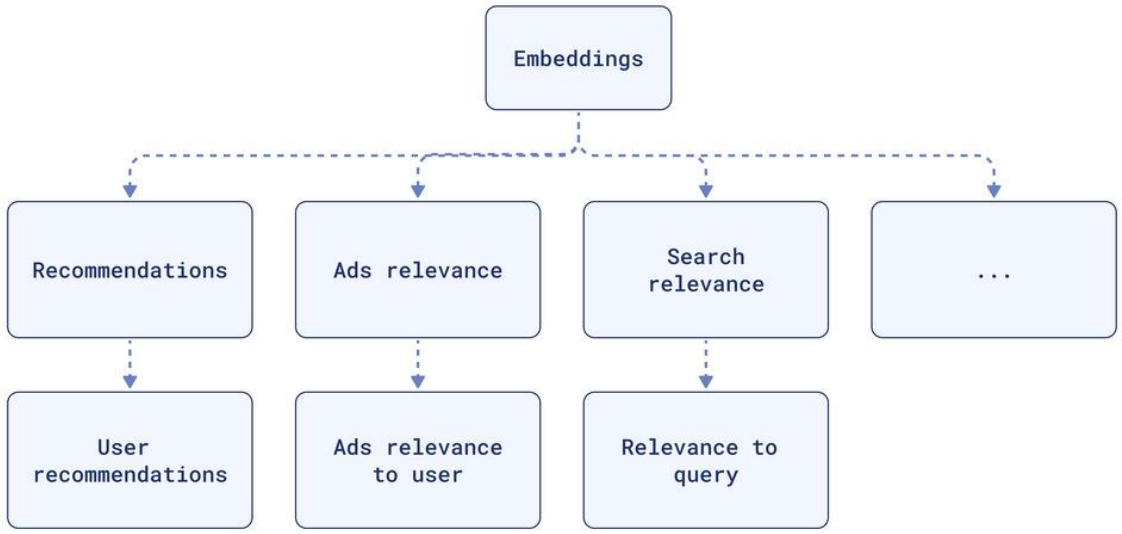
1. 大模型会判断是否需要调用function
2. 如果需要，会返回模型的名字和参数
3. 这些参数传给function，function产生结果
4. 大模型将结果处理成人类容易理解的文字

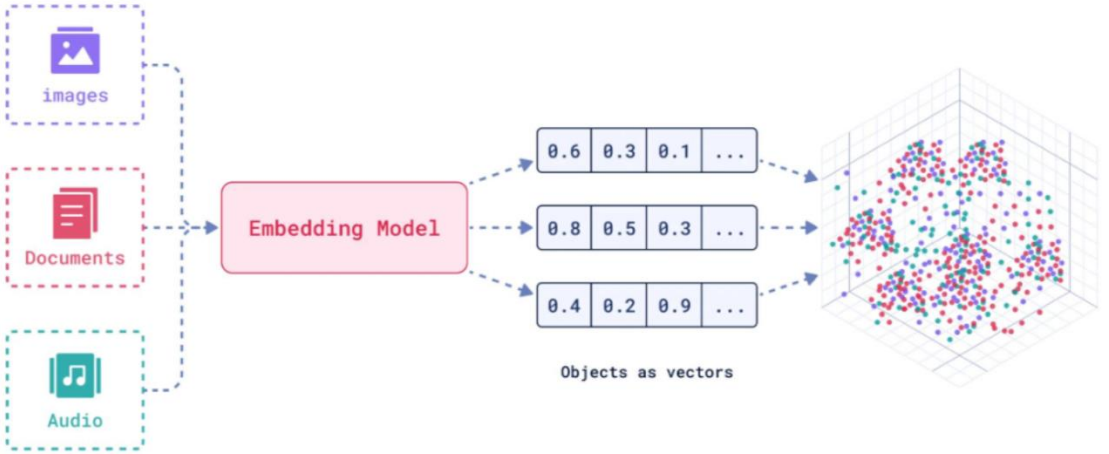
```
{'name':
'get_current_weather',
'arguments':
'{"location":"长春",
   "time": "2025-8-27"
'}
```

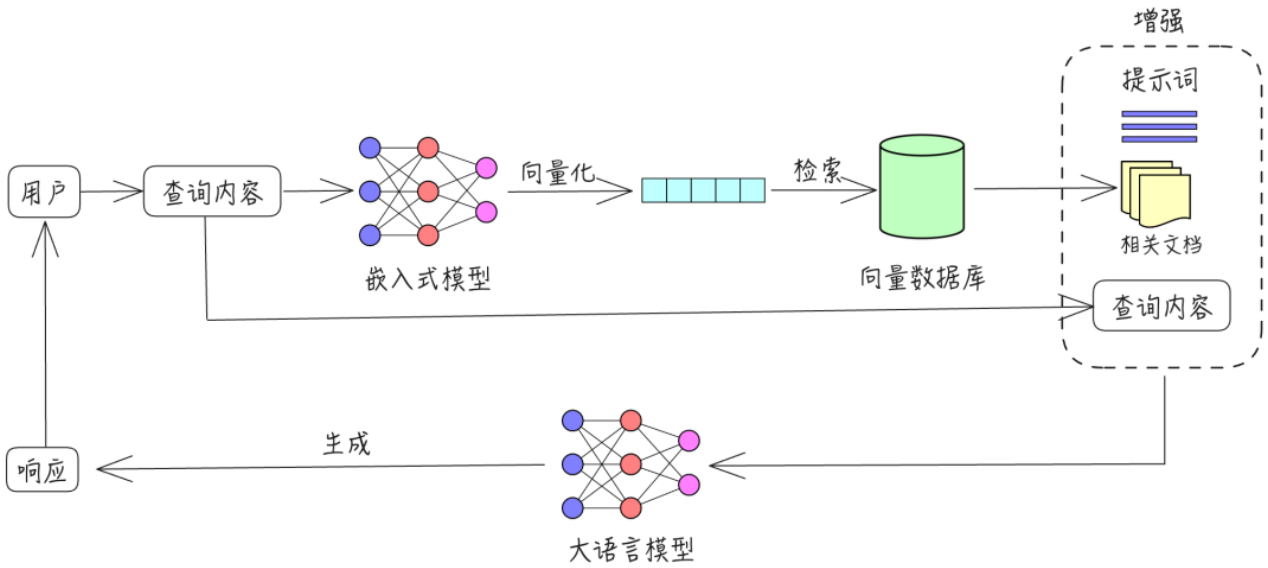
通用大模型

检索增强生成（Retrieval-Augmented Generation, RAG）

- 把人类能理解的信息转换为机器能理解的数字
- 嵌入模型把文字、图像等各种信息，转换成向量
- 向量间的距离表示相似度





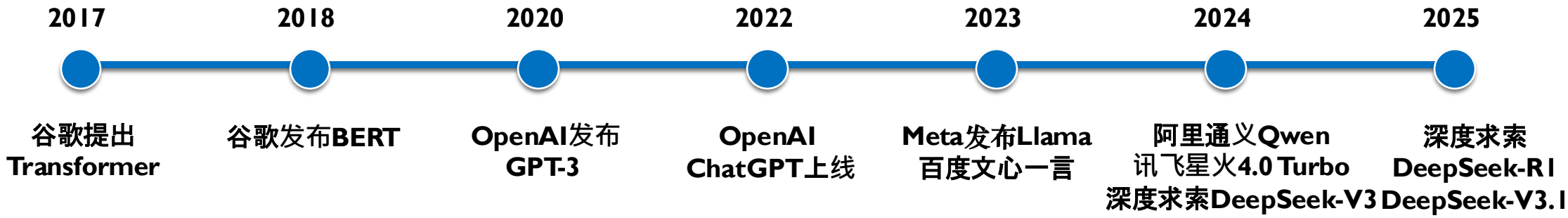


https://qdrant.tech/articles/what-are-embeddings/
https://www.gptsecurity.info/2024/05/26/RAG/

通用大模型

大模型发展时间线

- ChatGPT：
  - 2022年，GPT-3.5推出聊天界面，带来公众AI热潮
- DeepSeek-R1
  - 国产开源大模型，2025年1月发布，仅用一周时间，月活跃用户已达1亿



通用的大语言模型不具备领域知识，
对特定领域相关问题的回答不够准确

训练/微调
领域大模型

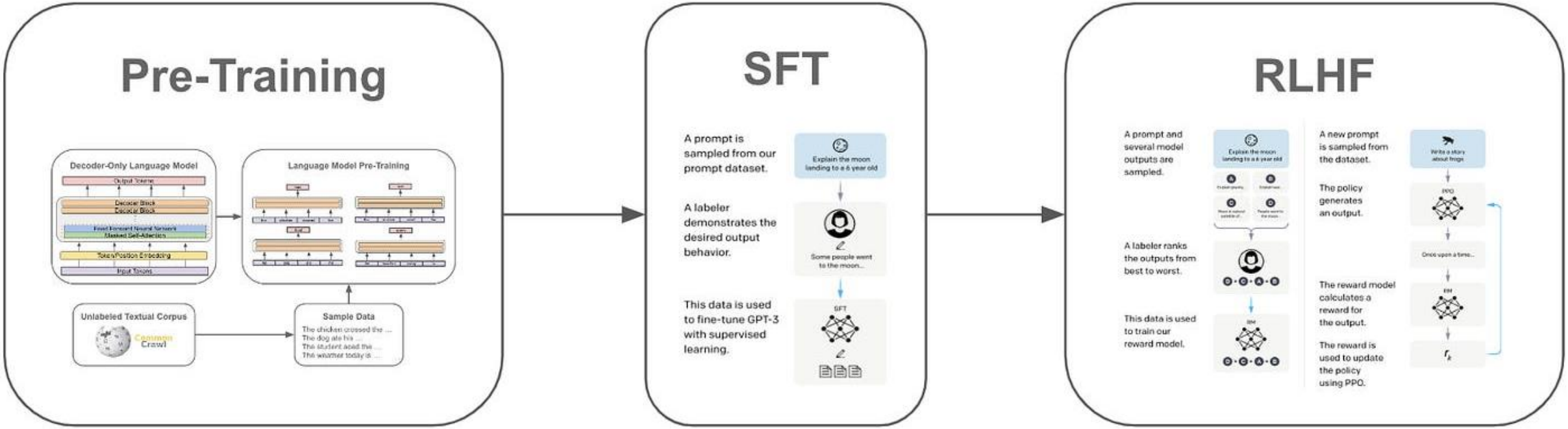
- 背景介绍
- Transformer与MoE框架
- 通用大模型
- 领域大模型
- HepAI模型进展
- 总结

领域大模型

预训练

监督微调

强化学习



领域模型优化策略

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022): 27730-27744.

领域大模型/平台

Llemma：

- 专为数学推理任务设计的大模型
- 7B和34B两个版本
- 以Code Llama为基础模型继续训练
- 训练数据：Proof-Pile-2（包含学术论文、数学网页、数学代码等的综合语料）

星语（StarWhisper）：

- 天文领域大模型
- 7B-72B等版本
- 训练数据：通过天文物理知识微调与强化学习

ChatLaw：

- 中文法律领域垂直大语言模型
- 13B、33B、Text2Vec等版本
- 训练数据：基于中国法律语料

更多领域模型：

- Deepseek-Math
- Qwen-Math
- MathGLM
- DeepSeek-Chem
- Qwen-Pharma
- BioGPT
- Qwen-Med
- DeepSeek-Science

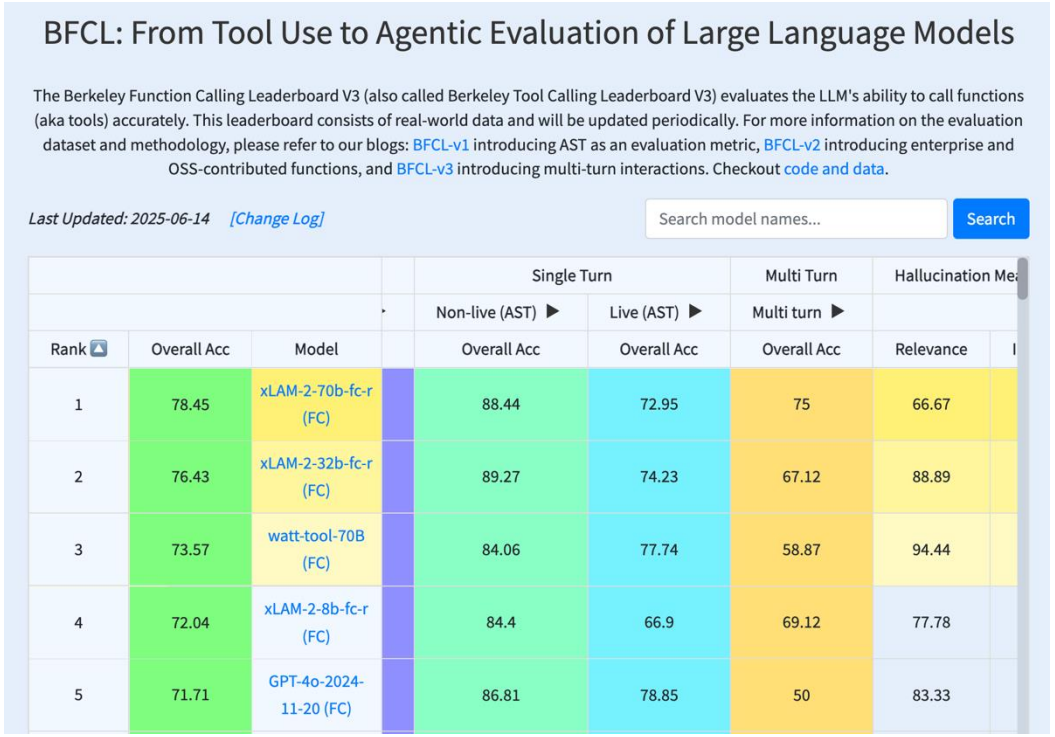
- 大语言模型的发展向多模态、智能体方向
- 单一通过微调获得的领域大模型逐渐减少
- 通过加入RAG、多智能体等形成领域科研平台

领域大模型/平台

特定任务优化的模型

针对function calling功能优化的模型

- xLAM-2-70b-fc-r
- watt-tool-70B
- ToolACE-2-Llama-3.1-8B



BFCL: From Tool Use to Agentic Evaluation of Large Language Models

The Berkeley Function Calling Leaderboard V3 (also called Berkeley Tool Calling Leaderboard V3) evaluates the LLM's ability to call functions (aka tools) accurately. This leaderboard consists of real-world data and will be updated periodically. For more information on the evaluation dataset and methodology, please refer to our blogs: BFCL-v1 introducing AST as an evaluation metric, BFCL-v2 introducing enterprise and OSS-contributed functions, and BFCL-v3 introducing multi-turn interactions. Checkout code and data.

Last Updated: 2025-06-14 *[Change Log]*

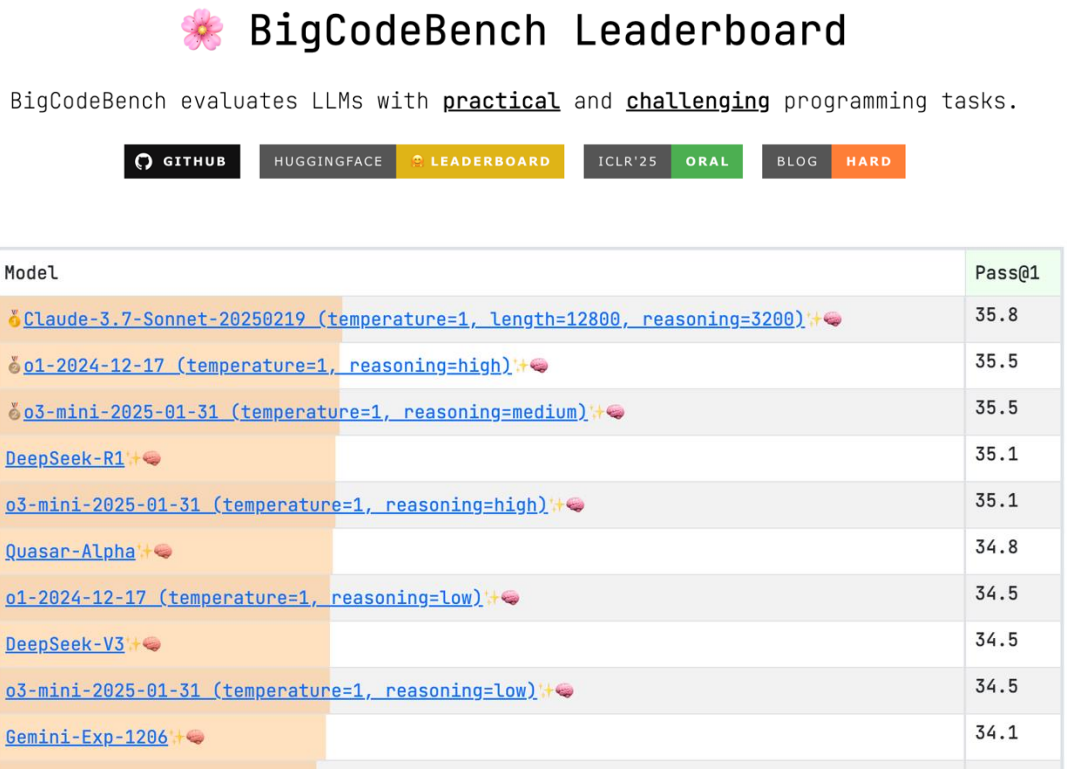
Search model names... Search

			Single Turn		Multi Turn	Hallucination Me
			Non-live (AST) ▶	Live (AST) ▶	Multi turn ▶	
Rank	Overall Acc	Model	Overall Acc	Overall Acc	Overall Acc	Relevance
1	78.45	xLAM-2-70b-fc-r (FC)	88.44	72.95	75	66.67
2	76.43	xLAM-2-32b-fc-r (FC)	89.27	74.23	67.12	88.89
3	73.57	watt-tool-70B (FC)	84.06	77.74	58.87	94.44
4	72.04	xLAM-2-8b-fc-r (FC)	84.4	66.9	69.12	77.78
5	71.71	GPT-4o-2024-11-20 (FC)	86.81	78.85	50	83.33

https://gorilla.cs.berkeley.edu/leaderboard.html

针对编程功能优化的模型

- Claude code
- Qwen3-coder



BigCodeBench Leaderboard

BigCodeBench evaluates LLMs with practical and challenging programming tasks.

GITHUB HUGGINGFACE LEADERBOARD ICLR'25 ORAL BLOG HARD

#	Model	Pass@1
1	Claude-3.7-Sonnet-20250219 (temperature=1, length=12800, reasoning=3200)	35.8
2	o1-2024-12-17 (temperature=1, reasoning=high)	35.5
2	o3-mini-2025-01-31 (temperature=1, reasoning=medium)	35.5
4	DeepSeek-R1	35.1
4	o3-mini-2025-01-31 (temperature=1, reasoning=high)	35.1
6	Quasar-Alpha	34.8
7	o1-2024-12-17 (temperature=1, reasoning=low)	34.5
7	DeepSeek-V3	34.5
7	o3-mini-2025-01-31 (temperature=1, reasoning=low)	34.5
10	Gemini-Exp-1206	34.1

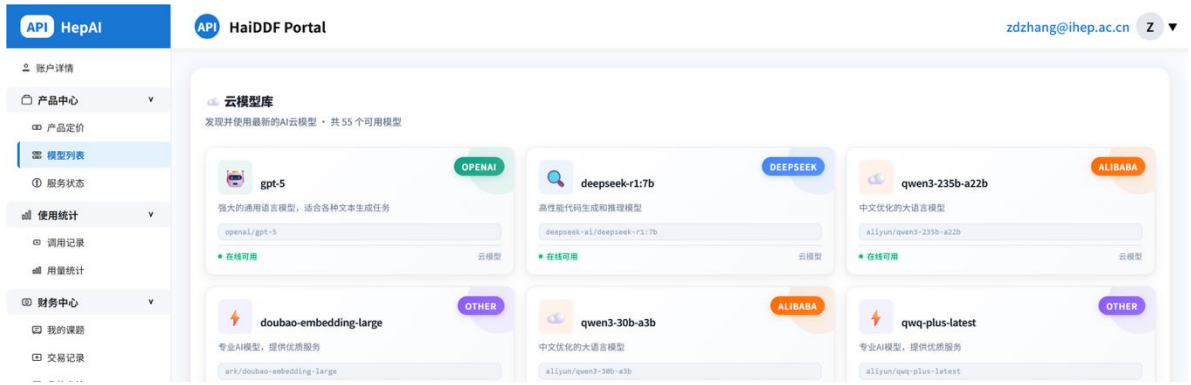
https://bigcode-bench.github.io/

- 背景介绍
- Transformer与MoE框架
- 通用大模型
- 领域大模型
- HepAI模型进展
- 总结

HepAI平台模型服务

https://aiapi.ihep.ac.cn

- 支持55+在线模型/工具
  - deepseek-v3, r1
  - GPT系列（4o, 5, o1, o3）等
  - 千问系列
  - Claude系列
  - 图像模型
  - 本地deepseek
  - 本地boss code worker等



- 通过API发起网络请求调用

```
from hepai import HepAI

client = HepAI(
    base_url="https://aiapi.ihep.ac.cn/apiv2",
    api_key=os.getenv("HEPAI_API_KEY")
)

response = client.chat.completions.create(
    model="hepai/deepseek-r1:671b",
    messages=[
        {"role": "user", "content": "你好"}
    ],
    stream=True
)

for chunk in response:
    if chunk.choices[0].delta.content:
        print(chunk.choices[0].delta.content, end="", flush=True)
```

大模型的本地部署

大模型在GPU、NPU、DCU上的本地部署

Generative Language Models

- DeepSeek-R1-671B
- DeepSeek-R1-Distill-Qwen-32B
- Qwen3-Coder

Agentic Language Models

- xLAM-2-70b-fc-r
- xLAM-2-32b-fc-r

Embedding & Reranking Models

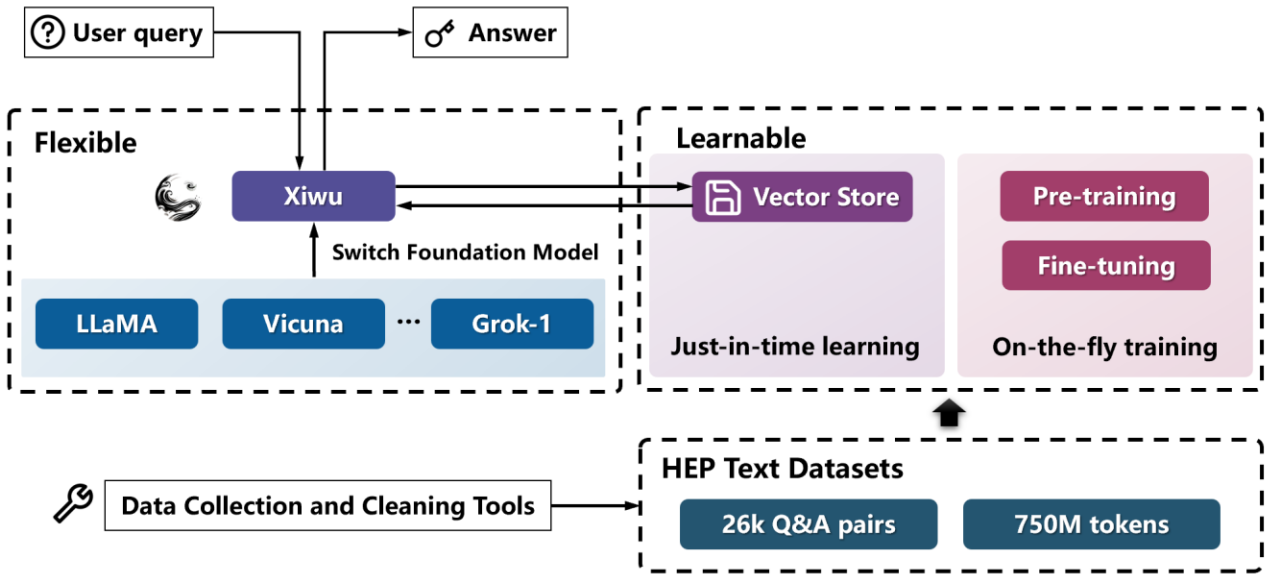
- BGE-M3
- BGE-Reranker-V2-M3

Specialization

- Text generation
- Code generation
- Dialogue systems
- Reasoning tasks
- Multi-turn dialogue
- Function calling
- Text representation,
- Information retrieval
- Relevance ranking

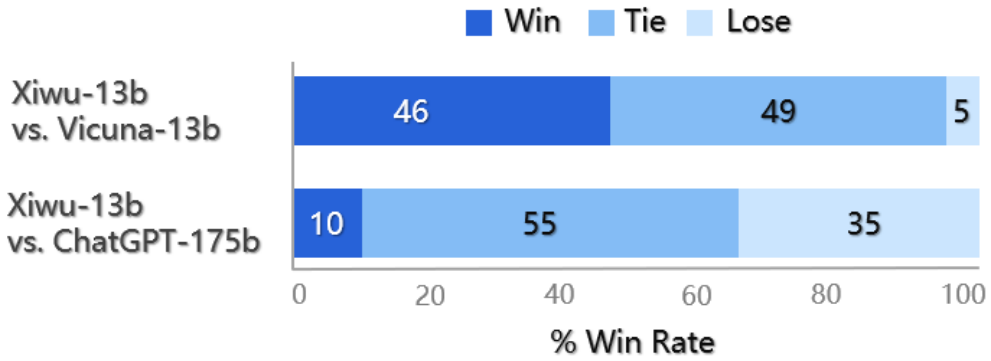
粒子物理领域大模型

Xiwu(溪悟):A Basis Flexible and Learnable LLM for High Energy Physics[1]



- 基于LlaMA1～3
- 在高能物理数据集上进行二次预训练、微调和强化学习
- 在GPU-A100和DCU-K100 AI训练
- 通过RAG提升准确性
- 在HEP Q&A和代码生成上效果优于通用大模型

Test Results



[1] arXiv:2404.08001

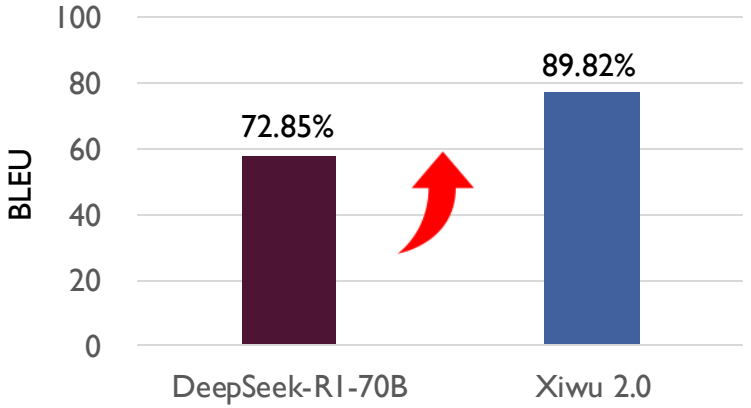
https://github.com/zhangzhengde0225/Xiwu

粒子物理领域大模型

Xiwu大模型的最新进展：

Xiwu 2.0

- Based on **DeepSeek-R1-Distill-Llama-70B**.
- Trainable params: 4B.
- Fine-tuned on 39k Q&A pairs, including mixed samples from both HEP and common corpus datasets.
- Trained on DCU-K100-AI.



Comparison on HEP domain-specific dataset.

为Dr.Sai智能体训练的大模型

Dr.Sai-Host

- Built on DeepSeek-R1-Distill-Llama-8B, Qwen3-4B and xLAM-2-32b-fc-r
- Finetuned using runtime data of Dr.Sai
- Optimized for LLM tool calling, specifically in expert agent selection.

Dr.Sai-Planner

- Currently under development, fine-tuned with runtime data of Dr.Sai
- Specialized in planning particle physics data analysis workflows.

总结

- 通过微调提升模型在领域数据集上的表现
- 针对具体的物理分析智能体，训练特定功能的模型
- 在HepAI平台实现本地部署与服务

- 未来工作
  - 领域数据的制造，尤其是带思维链（CoT）的数据和强化学习数据
  - 持续开发具有特定功能、支撑智能体运行的大模型

谢谢！