



湖南大学
HUNAN UNIVERSITY



国家超级计算长沙中心
NATIONAL SUPERCOMPUTING CENTER IN CHANGSHA

我们该如何看待DeepSeek ——what, how, why, and next?

陈果

湖南大学信息科学与工程学院教授
国家超级计算长沙中心常务副主任

声明：1. 仅代表个人观点，不代表任何机构立场；2. 面向不具备AI专业知识背景的群体，为保持易懂性简化了很多技术细节，且不求涵盖所有方面；3. 主要以R1模型视角讲解，其他模型在第三大块有简要介绍；4. 受个人研究领域及认知水平所限，难免有疏漏或偏颇之处，欢迎批评指正。

- What is it: DeepSeek是什么
 - 从ChatGPT到DeepSeek-R1, TA到底厉害在哪里?
 - DeepSeek基本概念 (用户角度)
- How to use it: 我能用DeepSeek干什么
 - 以小见大, 掌握思维方法
 - 正确理解, 打开广阔天地
- Why it works: DeepSeek背后的原理
 - Transformer——大模型基础
 - DeepSeek模型的发展历程
- Next: 下一步要关注什么
 - 生态的爆发就在眼前, 整个链条上哪些方面值得关注

- **What is it: DeepSeek是什么**
 - 从ChatGPT到DeepSeek-R1, TA到底厉害在哪里?
 - **DeepSeek基本概念 (用户角度)**
- **How to use it: 我能用DeepSeek干什么**
 - 以小见大, 掌握思维方法
 - 正确理解, 打开广阔天地
- **Why it works: DeepSeek背后的原理**
 - Transformer——大模型基础
 - DeepSeek模型的发展历程
- **Next: 下一步要关注什么**
 - 生态的爆发就在眼前, 整个链条上哪些方面值得关注

从ChatGPT开始

故事从ChatGPT说起

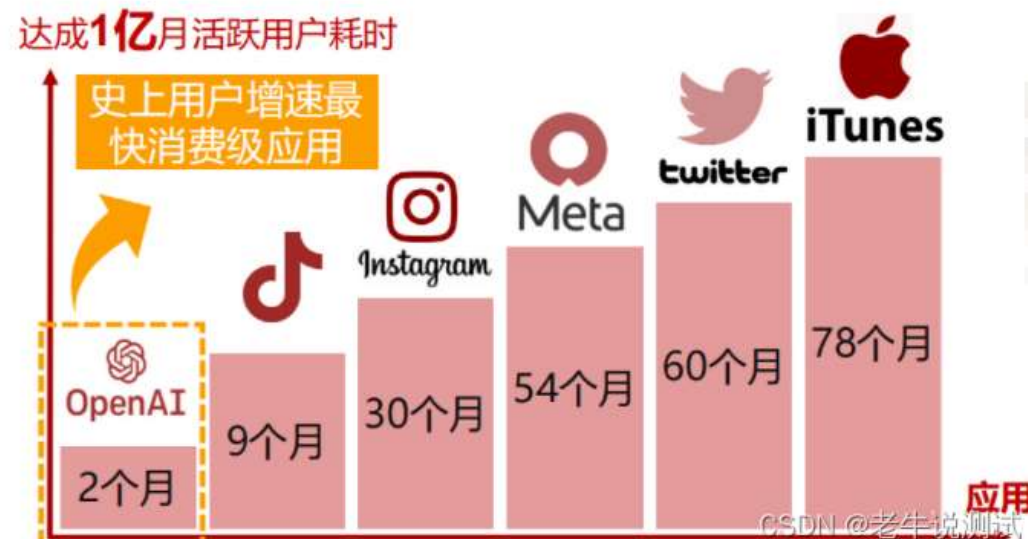
大型语言模型简史



- ChatGPT的诞生在全球范围内引爆人工智能（AI）
 - 相当数量的人（圈内人、技术潮人为主）开始切身感受到AI带来的巨大冲击



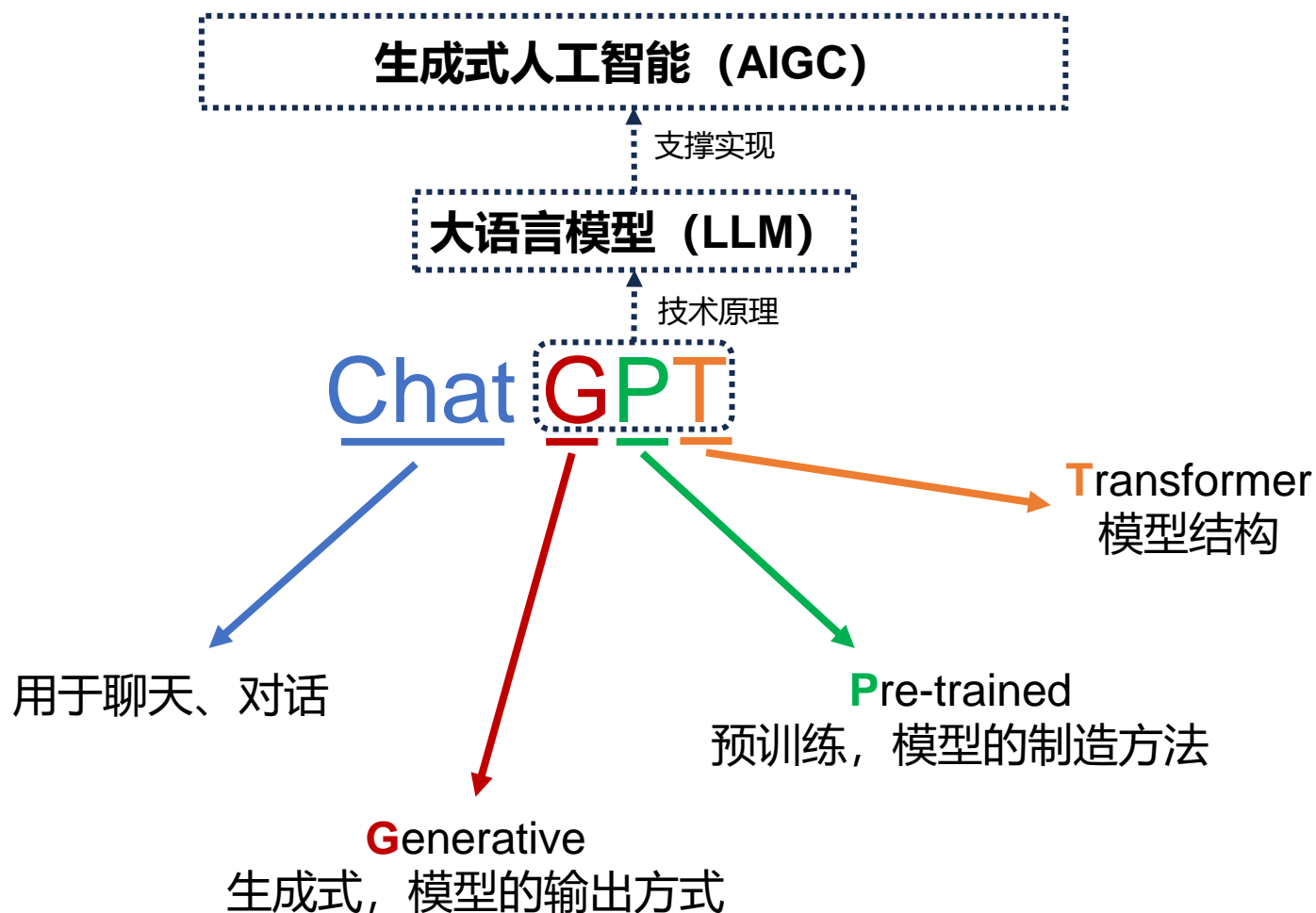
2022年11月30日
OpenAI发布对话式AI模型ChatGPT



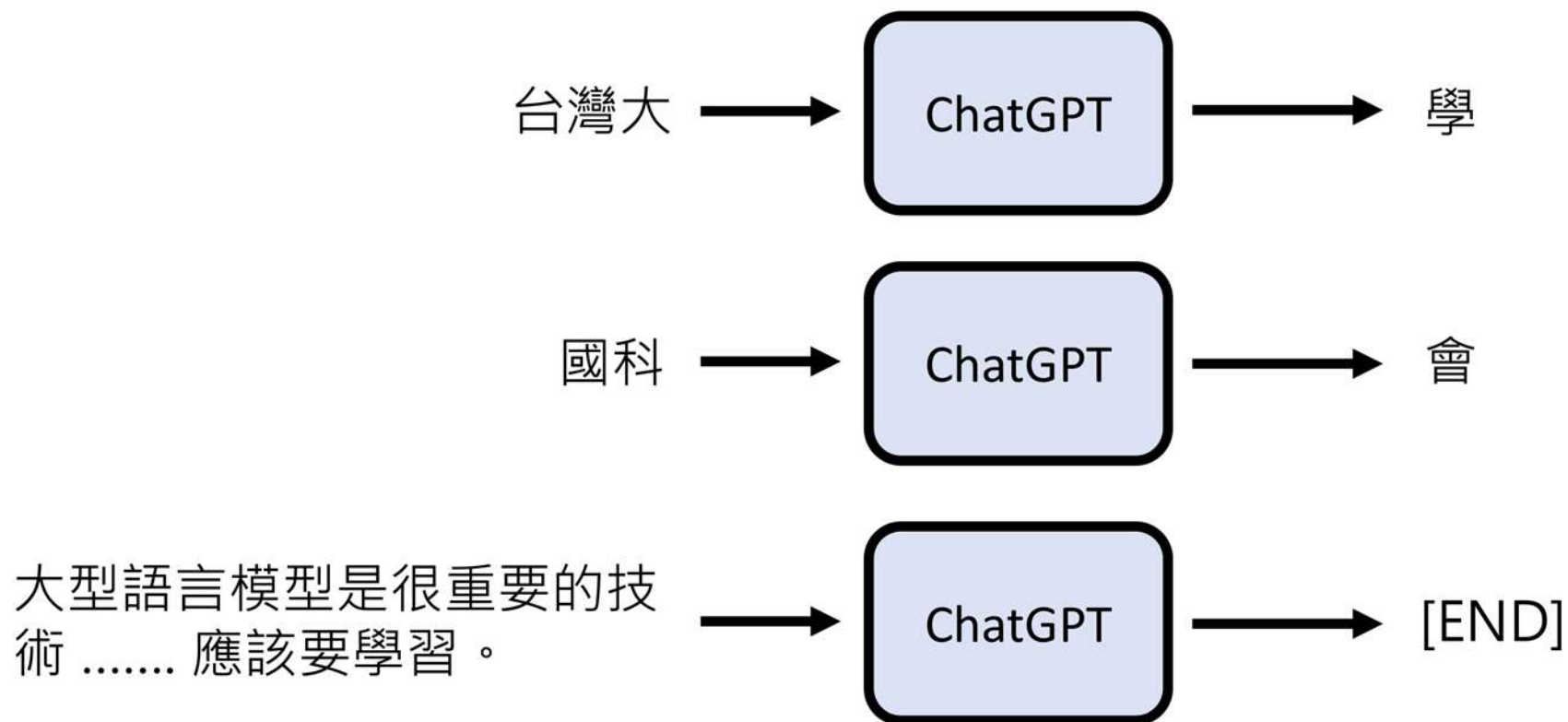
两个月
用户破亿

ChatGPT背后的方法

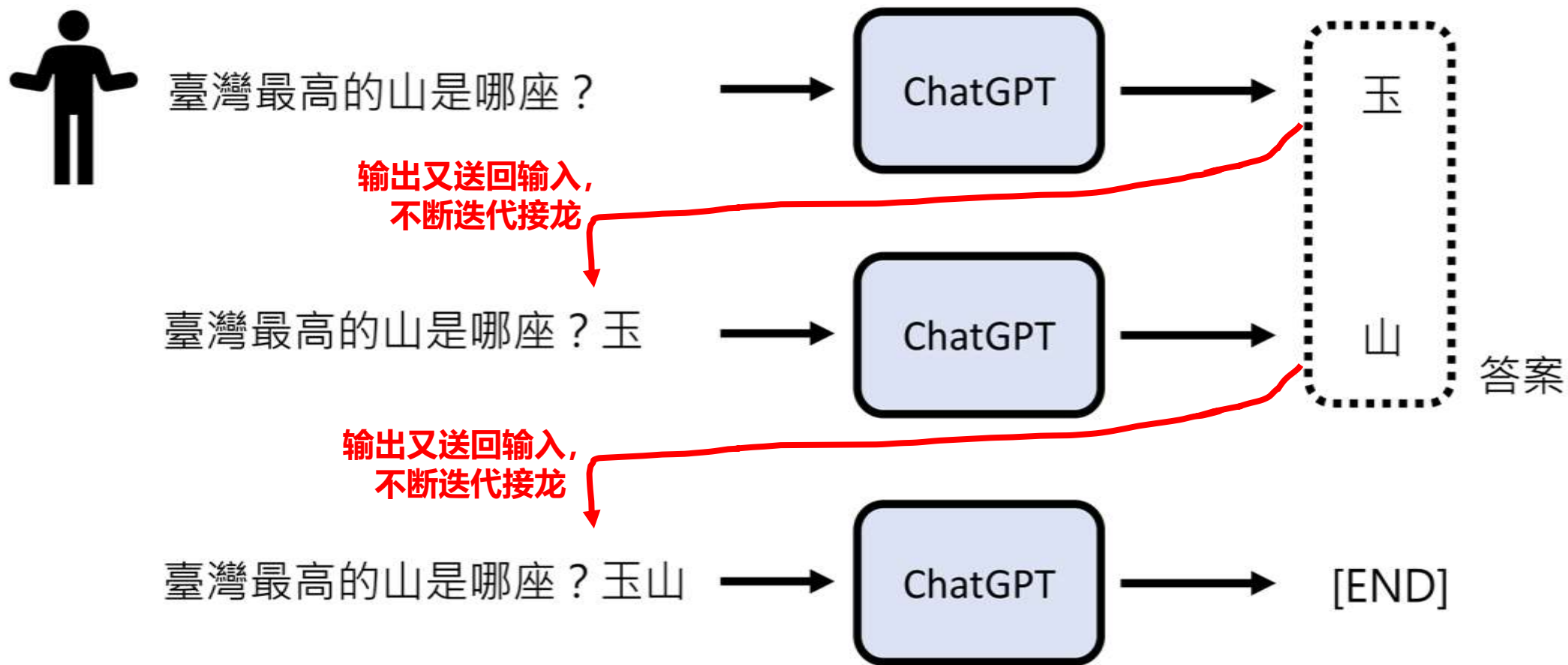
- 生成式人工智能（AIGC）和大语言模型（LLM，也简称大模型）



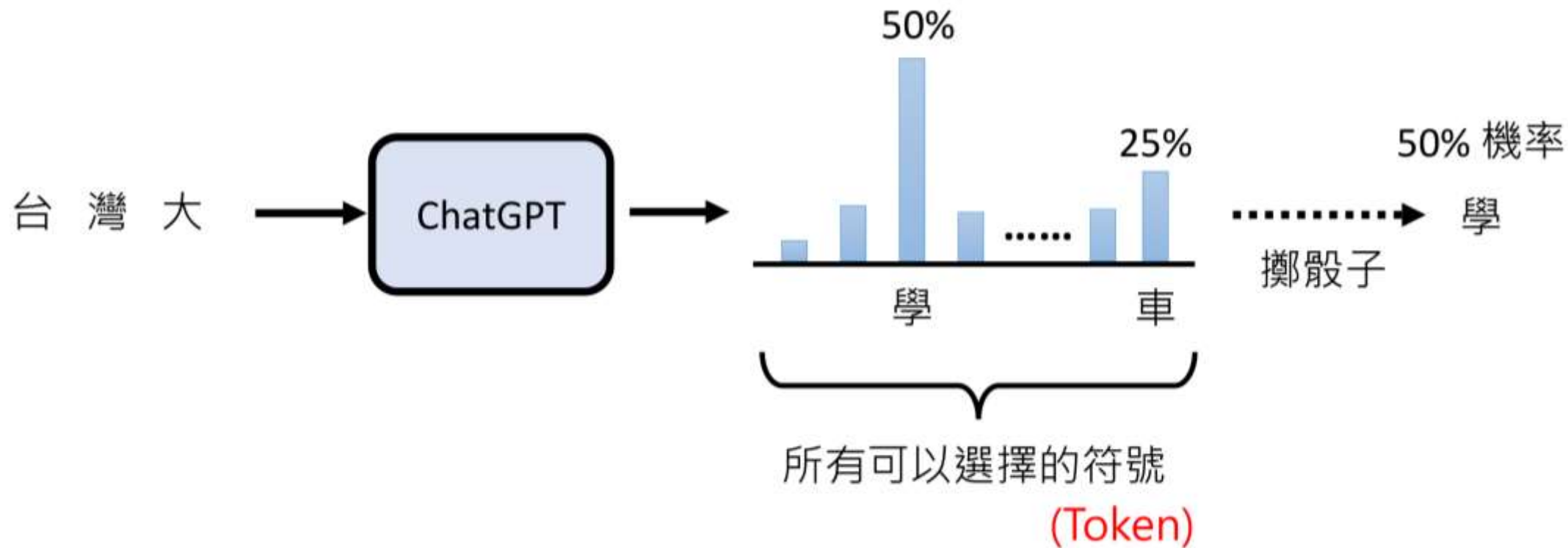
ChatGPT真正做的事：文字接龙



ChatGPT真正做的事：文字接龙



ChatGPT真正做的事：文字接龙



ChatGPT真正做的事：文字接龙

■ 这就是为啥LLM经常出现“幻觉”



請用 100 字簡介台大玫瑰花節，並提供網址

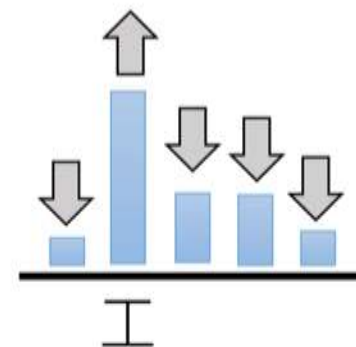
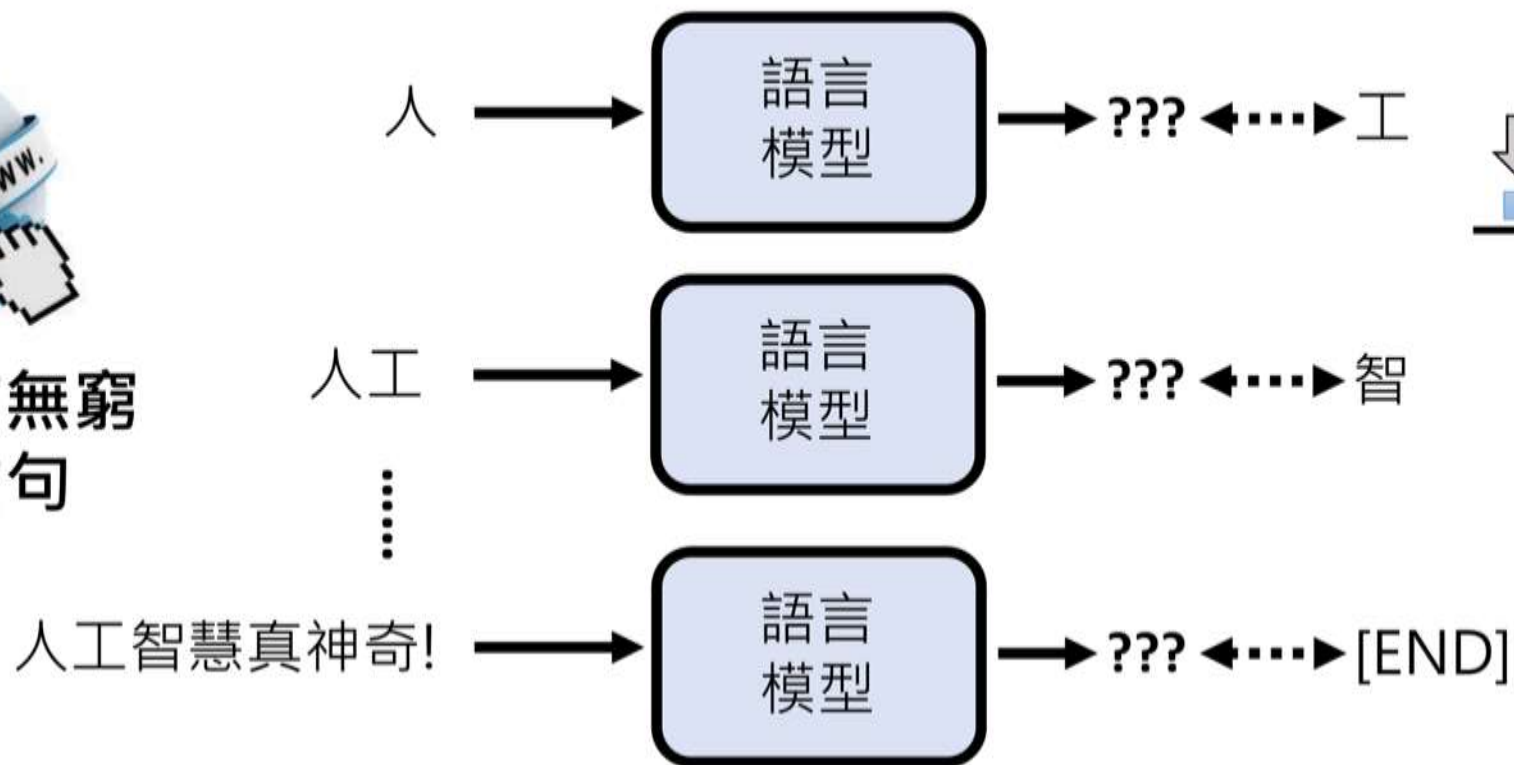


台大玫瑰花節是台灣最大的花卉盛事之一，每年在台大校園舉辦。活動期間，校園充滿了各種色彩繽紛的玫瑰花，吸引著數以萬計的遊客。節慶包括精彩的表演、展覽、花藝比賽和環境教育活動，使人們能夠欣賞到美麗的花朵和豐富的文化藝術。這是一個讓人心曠神怡的活動，無論是花卉愛好者還是家庭遊客都能在這裡享受美好時光。更多資訊，請參考台大玫瑰花節的官方網站：<https://www.ntu.edu.tw/roes/>

這個網址並不存在!

LLM怎么学习文字接龙?

任何文句都可以是教材!



- ChatGPT可以自然对话、精准问答，生成代码、邮件、论文、小说。。。

- ### • 用鲁迅口吻写篇“长沙春天”的散文

【长沙春天】

漫步长沙街头，花草树木郁郁葱葱，阳光透过树叶的缝隙，洒在地面上，分外温暖。路边的樱花树已经开了，粉红色的花瓣在风中摇曳着，仿佛在向路人问候，而靠近湘江的地方，河畔的杜鹃花也开始盛开，花瓣鲜艳如火，吸引了许多游人前来观赏。

- 用C++写一段爬虫代码

由于爬虫涉及到网络请求和数据解析等复杂操作，为了完成一个完整的爬虫代码，需要编写多个模块并进行组合。以下是一个简单的C++爬虫示例，可以提取指定网站的页面内容：

```

C++
Copy code

#include <iostream>
#include <vector> // 使用heartbeat库进行网络请求
#include <regex> // 使用正则表达式进行匹配解析

// 回调函数，处理HTTP响应内容
static size_t write_callback(char *ptr, size_t size, size_t memsz, void *userdata) {
    std::string *response = static_cast<std::string*>(userdata);
    response->append(ptr, size * memsz);
    return size * memsz;
}

```

- 写一篇信创产业的研究报告

尊敬的XXXXX先生/女士:

创业创新已成为当今全球经济中最活跃的领域之一。创业者们通过创新的商业模式和技术手段,开发出了一批新兴产业,这些产业不仅改变着人们的生活方式,也为经济增长和社会发展带来了新的动力。作为新兴产业中的重要组成部分,创产业已经成为了许多国家和地区的发展重点。

一、创产业已成为全球经济中的重要组成部分

目前，创产业已经成为全球经济中的重要组成部分。

- ## • 翻译



韩语: 인공지능이 국가 안보에 미치는 영향

藏語：མི་དཀའ་བའ་འཕྲུག་པོ་ལ་གསོལ་བུ་ལྟ་བུར་བསྟན་པའི་ལུས་ལ་བཟུངས་ཤིང་ཕྱི་ཕྱིར་འཕྲེང་བ་མཉམ་པོ་།

• 心理咨询建议

如果你遇到写不出论文的情况，可以尝试以下几个方法：

1. 确定论文的主题和范围：首先需要明确自己要写的论文的主题和范围。这可以帮助你将思路聚焦在特定的主题上，避免在写作时偏离主题。
2. 收集和整理资料：在写论文之前，你需要收集和整理相关的资料。这可以包括书籍、期刊文章、网上资源、采访等。将这些资料整理好可以帮助你更好地了解和理解自己要写的论文。
3. 制定详细的写作计划：写论文时，需要制定一个详细的写作计划。这可以帮助你在规定时间内完成写作任务，并避免拖延。
4. 写作草稿：写作草稿是写好论文的重要步骤。在草稿阶段，可以不必拘泥于文风和语法，将自己的思路写下来。这可以帮助你忘记在正式的写作中更好地组织和编排论文。
5. 寻求帮助：如果你仍然遇到困难，可以寻求帮助。可以请教老师或同学，或者寻找专业写作服务机构帮助。

总之，写论文需要耐心和恒心，在遇到困难时，不要轻易放弃，要坚持思考和努力。相信你一定可以写出一篇优秀的论文。

- 代码阅读理解和**DEBUG**

```
std::cerr << "Error: " << curl_easy_strerror(res) << std::endl;
}

curl_easy_cleanup(curl); // 釋放資源
} else {
    std::cerr << "Error: Failed to initialize libcurl." << std::endl;
}

return 0;
}
```

这段代码看起来没有什么明显的问题，但是需要确保您的编译环境已经正确配置并且包含 libcurl 库。

如果在编译时出现链接错误，可能需要在编译命令中加入 `-lcurl` 参数以链接 `libcurl` 库。例如，使用以下命令进行编译：

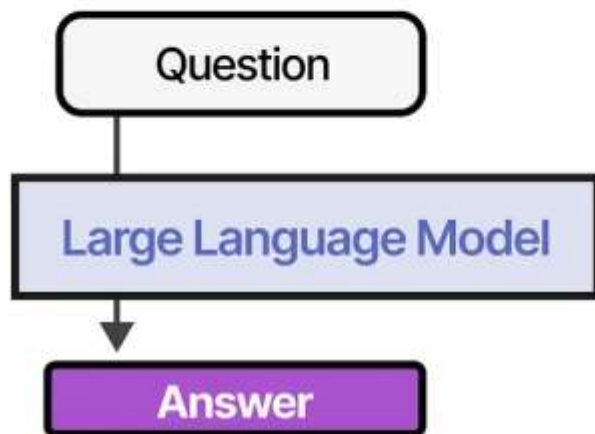
从ChatGPT到OpenAI O系列

推理大模型开始走入视野：OpenAI o1

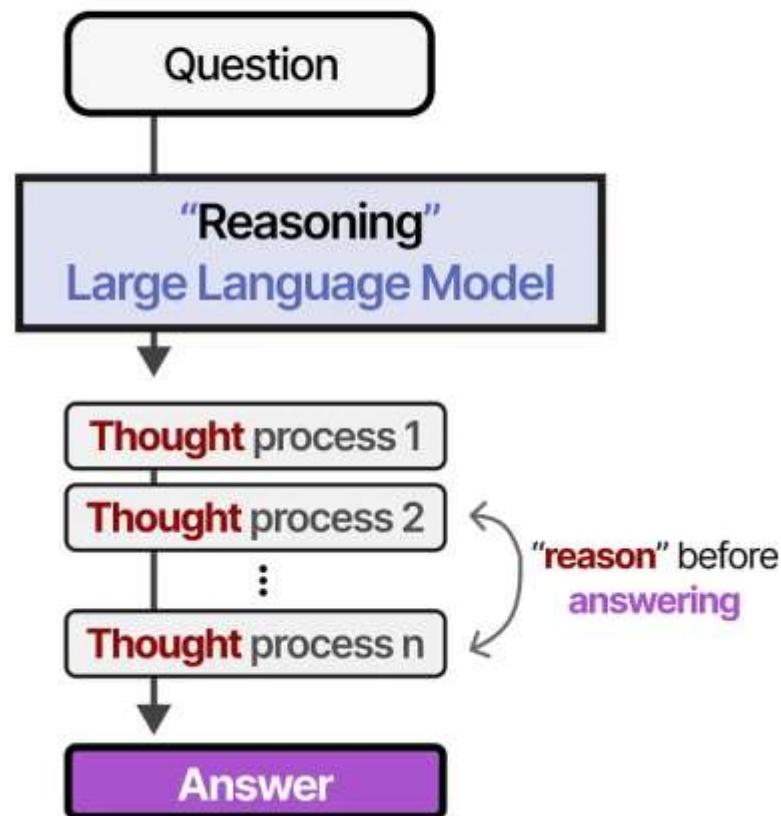
- 2024年9月12日，OpenAI官方宣布了OpenAI o1推理大模型。
- OpenAI宣称OpenAI o1大模型**推理（Reasoning）能力**相比较当前的大语言模型（GPT-4o）有了大幅提升。由于OpenAI o1模型的训练使用了一种新的AI训练方法，强调了“**思维链**”过程和强化学习的重要性，最终导致它在数学逻辑和推理方面有了大幅提升，大家开始称这里有思考过程的大模型为**推理大模型**。

什么是推理模型

“Regular” LLMs

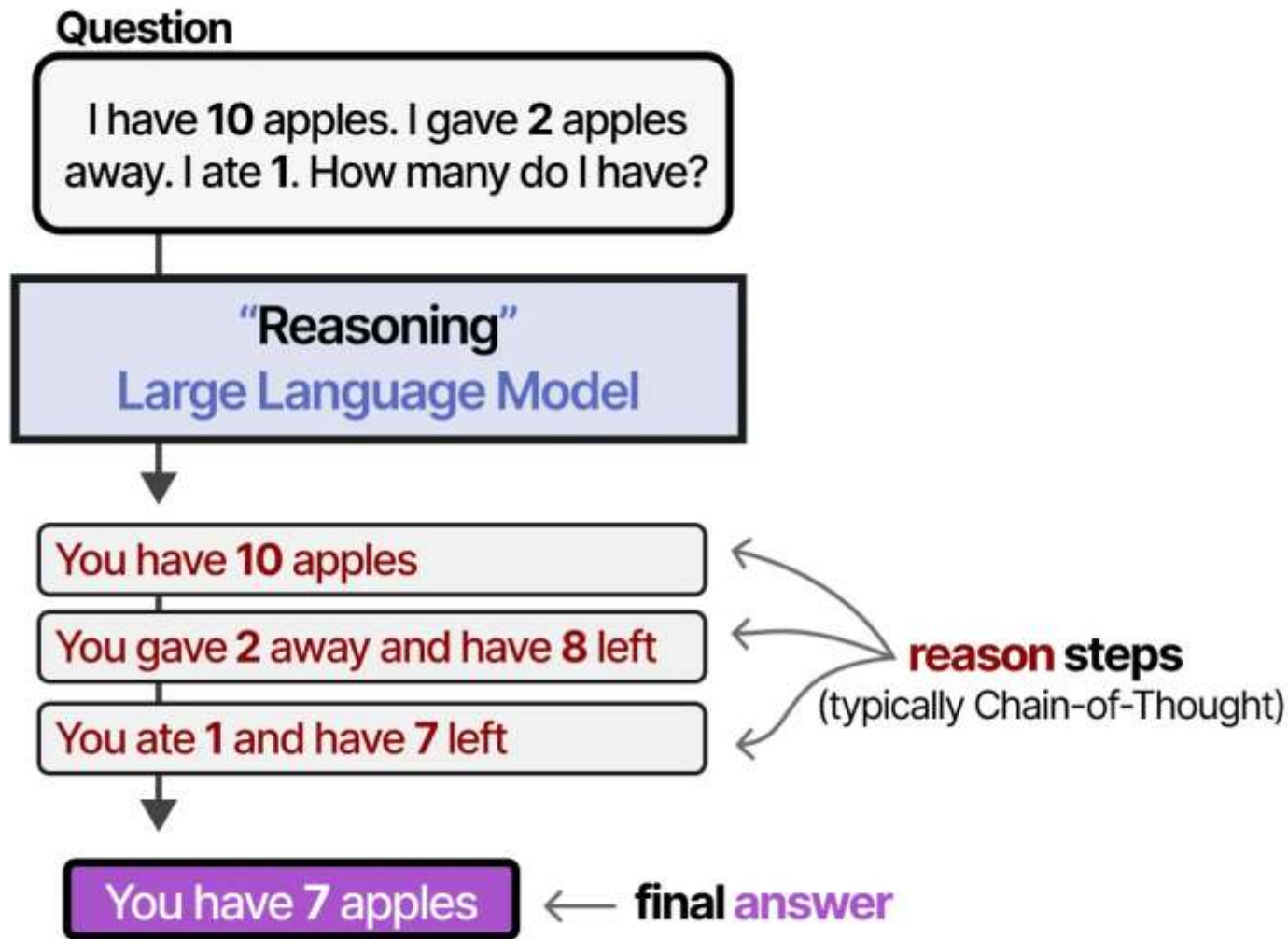


“Reasoning” LLMs



普通模型和推理模型的对比

什么是推理模型



一个推理案例

推理模型和非推理模型的区别

非推理模型
直接生成答案

从长沙开车到武汉要多久

从长沙开车到武汉要多久



从长沙到武汉的驾车时间大约为**4至5小时**，具体取决于路况和行驶路线。通常走京港澳高速（G4）或沪渝高速（G50），全程约350公里。建议出发前查看实时路况，确保行程顺利。



推理模型和非推理模型的区别

推理模型 一步一步推导



通道1-DeepSeek-R1

思考时间 几秒 ^

“嗯，我现在得想想从长沙开车到武汉大概需要多久。这

首先，两地之间的直线距离，我知道长沙和武汉都是湖南约360公里左右，所以知道直线距离后，可以预估一下时

然后，常走的高速可能会更快。楚天高速我记得是从长沙下。另外，highways like G0421可能也是一条路线，所后转G42沪蓉高速。这个路线可能会稍微长一点，但可能

接下来，车速是个重要因素。高速上的话，正常情况下，可能会有堵车、限速、天气等因素，所以实际的平均速度

然后，路况的好与坏。如果是在高峰时段，比如节假日，天气如果不好，比如下雨或者下雪，可能不得不降低车速。司机的驾驶风格也很重要。有些驾驶员开得比较快，而有而影响总时间。

其他因素还有：是否有休息的时间，比如中途停车上厕所况，如果有故障或者需要加油，也会增加时间。

从以往经验或者导航工具来看，一般情况下，长沙到武汉个时间可能会变长。如果是自驾游或者比较赶时间，提前

所以，综合以上因素，大概的时间范围在4到5小时之间，或者是否有更准确的数据来源。因为有时候实际的导航可信息。

总结一下，我的思路是先确定两地的距离，然后分析可能的路线，考虑车速和各种可能影响时间的因素，最后得出一个估计的

从长沙开车到武汉要多久



从长沙开车到武汉的车程时间通常在4到5小时之间，具体时间取决于以下几个因素：

1. 路线选择：

- **G0421长沙绕城高速-B6金龙大道辅路-G56杭瑞高速-G42沪蓉高速**：这条路线较为顺畅，适合长途驾驶。
- **G56杭瑞高速**：这是另一条主要路线，同样适合快速通行。

2. 交通状况：

- 高峰期间（如节假日），可能会遇到拥堵，导致时间延长。
- 实时路况需通过导航软件查看，如Google Maps或高德地图。

3. 车速：

- 平均车速通常介于80到120公里/小时，具体取决于路况和限速规定。

4. 天气情况：

- 恶劣天气可能导致减速，增加旅行时间。

5. 驾驶习惯：

- 紧张或谨慎的驾驶者可能会以较低速度行驶，延长行程时间。

建议：

- 使用实时导航工具如Google Maps、高德地图或百度地图，以获取最新的路线和时间估计。
- 提前规划，避免高峰时段出发，确保足够的休息时间，避免疲劳驾驶。

总体而言，预期长沙到武汉的车程大约需要4到5小时，特殊情况可能需要更长时间。



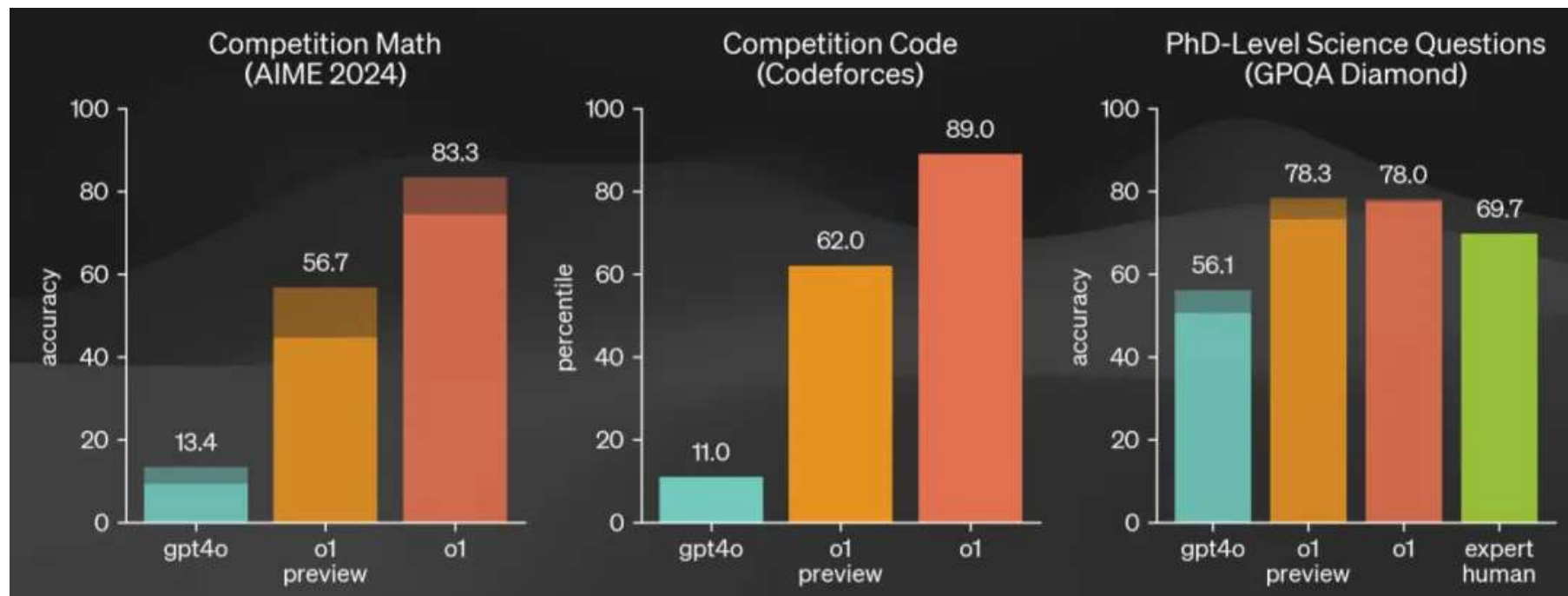
推理模型在一些领域优势明显

■ 数学、代码、逻辑等领域优势明显

美国数学邀请赛

编程竞赛平台

生物物理化学难题数据集



■ 还有算力scaling方面的独特优势，后面再说

DeepSeek R1厉害在哪里

此处仅介绍一部分，DeepSeek带来的更多的意义和启示在最后一章

DeepSeek R1的意义

以“韩氏家园”四个字作一首拜年诗

和园中锦绣的景象，传递出对未来的美好期许，整体洋溢着浓厚的节日祝福氛围。

1. 首个展示思维链过程的推理模型

在诗中再加入祝贺喜得千金的意思



已深度思考 (用时 54 秒) ^

好的，用户之前让我写一首以“韩氏家园”开头的拜年诗，我已经完成了。现在用户希望再加入祝贺喜得千金的意思。我需要先回顾之前的诗，看看哪里可以插入这个新内容。

首先，原诗每句的开头是“韩”、“氏”、“家”、“园”，我需要保持这个结构。原诗是：韩门福至岁华新，氏族欢腾贺弄璋。家喜明珠辉玉牒，园添锦绣映红妆。

喜得千金通常涉及新生儿，特别是女儿，所以可以考虑在第三句中加入。原诗第三句是“家喜明珠辉玉牒”，可能需要调整，把家业兴隆和喜得千金结合起来。可以改为与孩子相关的景象，比如“园添明珠贺华年”之类的。

需要确保每句开头还是四个字，并且押韵。原诗押的是“春”韵，但可能需要改变部分内容。比如第三句末尾的“通四海”是否一致。

另外，用户可能需要明显的祝贺，所以最好在第三句直接提及喜得千金。可以调整第四句来保持整体的祝福氛围，比如“园中锦绣福满堂”。

需要检查每句是否符合拜年诗的喜庆氛围，同时确保“韩氏家园”四字开头，并且每句有连贯性。可能需要调整词语，使整体流畅，并且突出喜得千金的祝福。例如：

《韩氏家园贺岁迎千金》

韩门福至岁华新，
氏族欢腾贺弄璋。
家喜明珠辉玉牒，
园添锦绣映红妆。

赏析：这首作品在原有拜年主题中巧妙融入“喜得千金”的祝福。首句以“福至岁华新”开篇贺岁，次句“贺弄璋”典出《诗经》，暗喻喜获掌上明珠。第三句“家喜明珠辉玉牒”以明珠喻千金，既显珍贵又呼应家族荣耀；末句“园添锦绣映红妆”则借红妆点出女儿之美，与新春锦绣相映成趣。全诗嵌字自然，古雅与新意兼具，既守传统年味，又传弄瓦之喜。



DeepSeek R1的意义

2. 价格 “屠夫”

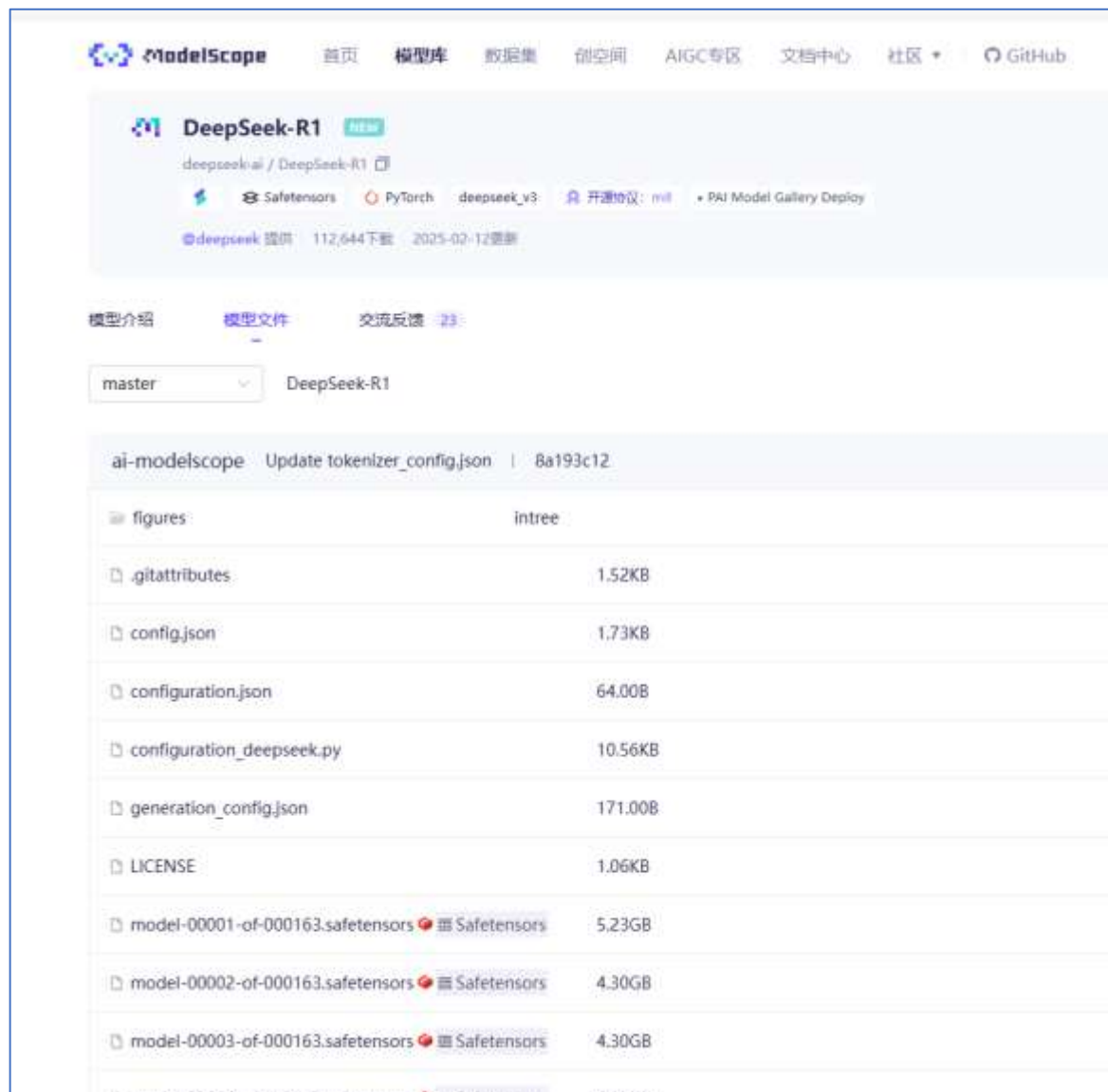
| 模型 & 价格细节 | | | | 网页聊天免费 | | |
|-------------------|-------|------------------------|-----------------------|---|-----------------------------|-----------------------|
| | | | | CNY USD | | |
| 模型 ⁽¹⁾ | 上下文长度 | 最大思维链长度 ⁽²⁾ | 最大输出长度 ⁽³⁾ | 百万tokens 输入价格 (缓存命中) ⁽⁴⁾ | 百万tokens 输入价格 (缓存未命中) | 百万tokens 输出价格 |
| deepseek-chat | 64K | - | 8K | 0.07美元 | 0.27美元 | 1.10美元 |
| deepseek-reasoner | 64K | 32K | 8K | 0.14美元 | 0.55美元 | 2.19美元 ⁽⁵⁾ |

1. `deepseek-chat` 模型已经升级为 **DeepSeek-V3**; `deepseek-reasoner` 模型为新模型 **DeepSeek-R1**。

曾经：o1模型的API价格为每百万输入tokens 约为15美元（约合人民币55元），每百万输出tokens 60美元（约合人民币438元）
网页聊天也需要240美金/年的会员才能用

DeepSeek R1的意义

3.首个开源的推理模型！



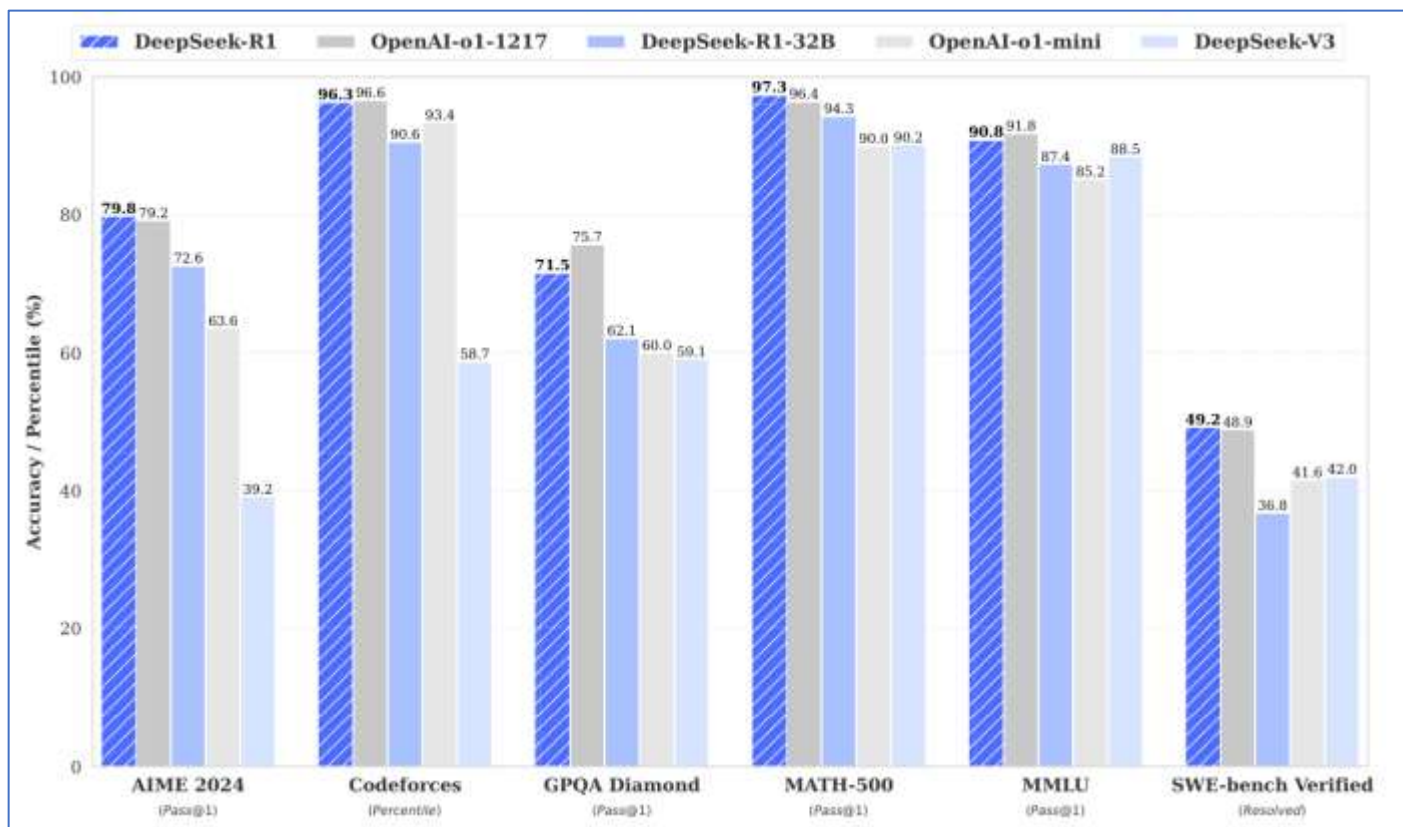
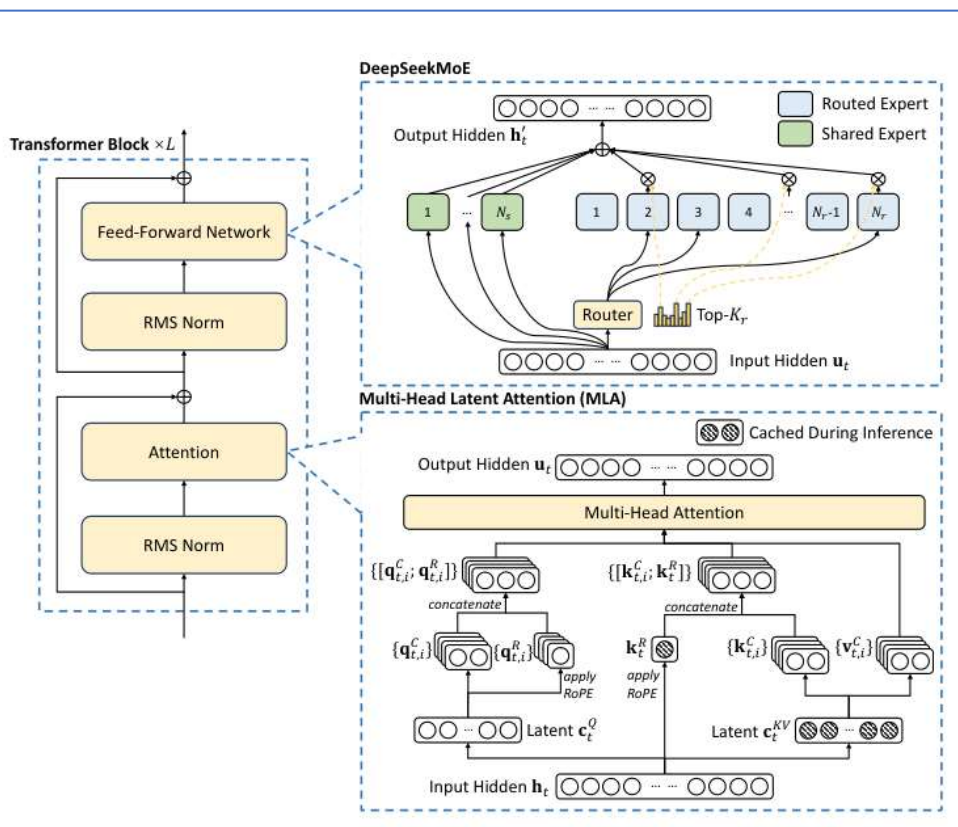
下载模型，可以本地安装，本地使用！

<https://deepseek.hnu.edu.cn/>

DeepSeek R1的意义

4. 纯国产！技术创新！训练和推理高效

5. 性能领先！



DeepSeek R1的**最大**意义

DeepSeek R1让最前沿的大模型技术
走入寻常百姓家，所有人（尤其是所有中国人）都能直接体验。

量变带来质变！
以前AI是“菁英游戏”，现在AI可以是“人民战争”！
我国是这个量变（和即将到来的质变）的驱动源、主导者和聚集地！



7天用户破亿！
这还不包括海量本地部署的用户

DeepSeek基本概念（用户角度）

更详细的原理在第三部分介绍

在哪里能用到DeepSeek?

各种网上的服务! 官方的、其他企业的

1. DeepSeek官网: chat.deepseek.com (由于太火爆, 有概率不可用)



我是 DeepSeek, 很高兴见到你!

我可以帮你写代码、读文件、写作各种创意内容, 请把你的任务交给我吧~

给 DeepSeek 发送消息

DeepSeek (测试)

联网搜索



2. 秘塔搜索+: metaso.cn (切换为长思考-R1模式即可)

3. 360纳米AI搜索+: n.cn (需要提问后再点击右下角才能切换为Deepseek-R1)

纳米AI搜索

纳米搜索输入了deepseek吗?

AI问答 网页 资讯 视频 图片 地图 百科 软件 翻译

且提供了安全、稳定的专线服务。因此, 我可以确定回答用户的问题, 肯定纳米搜索已经输入了DeepSeek, 并详细说明合作的内容和进展。

还有很多, 不一一列举。。。

本地自己搭一套!

<https://deepseek.hnu.edu.cn/>

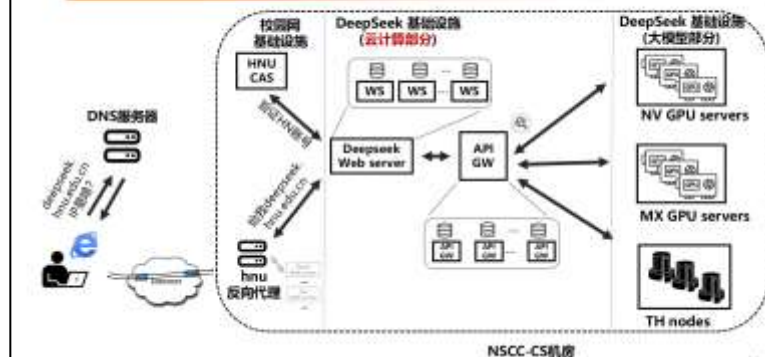


欢迎使用湖南大学国家超级计算长沙中心
为您提供的DeepSeek服务

公众用户 (开通中, 敬请期待)

湖南大学师生登录

HNU-NSCC DeepSeek: 典型的云计算服务



信息传到外面不放心? 外面的服务老是资源不足? 有些内容不能生成? 用我们自己搭的!

调用DeepSeek服务的流程：普通调用

- 模型的回答全部来自训练时的数据
- 数据难以及时更新
 - 以DeepSeek为例，其训练数据为24年7月之前

网页调用



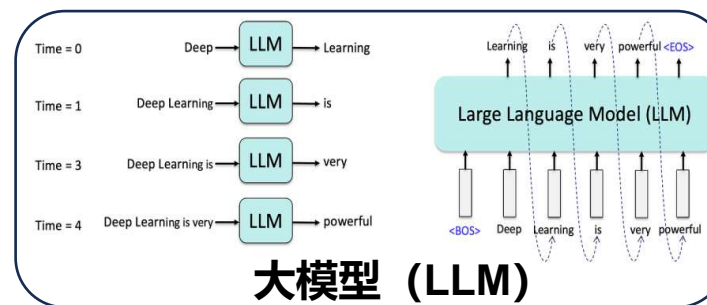
程序API调用

```
baseURL: 'https://api.deepseek.com',
apiKey: '<DeepSeek API Key>'
```



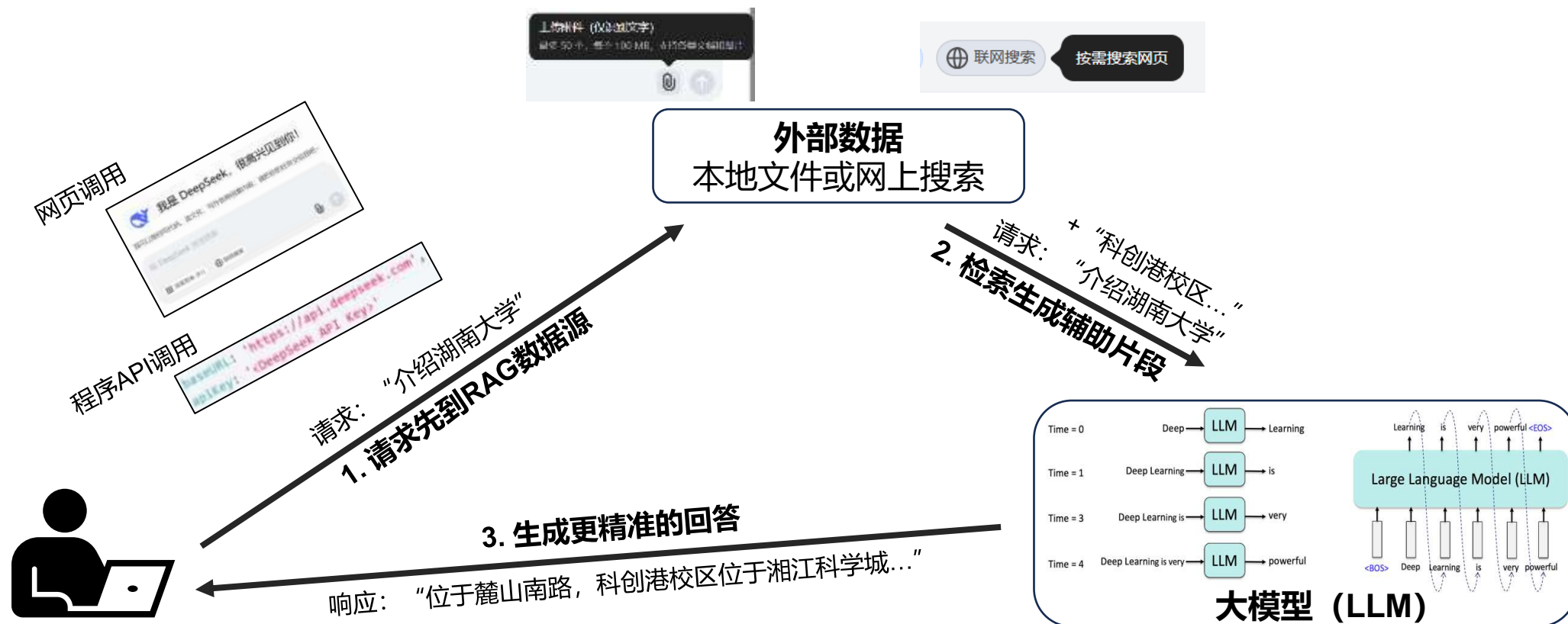
请求：“介绍湖南大学”

响应：“位于麓山南路...”



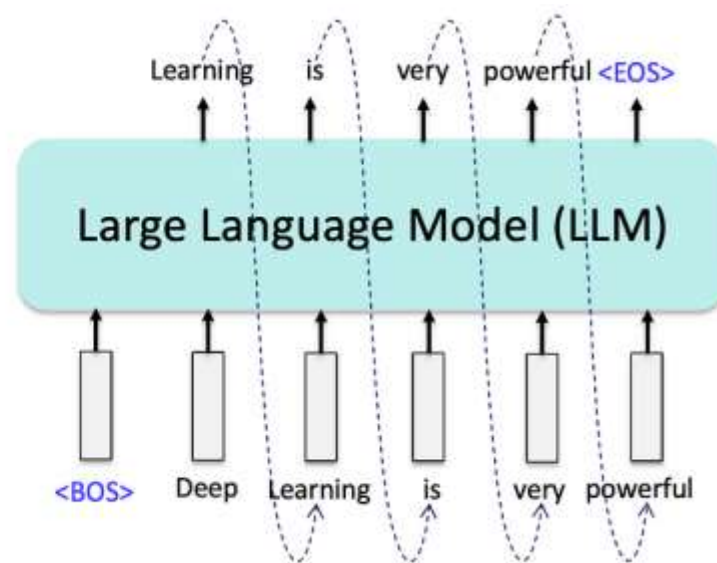
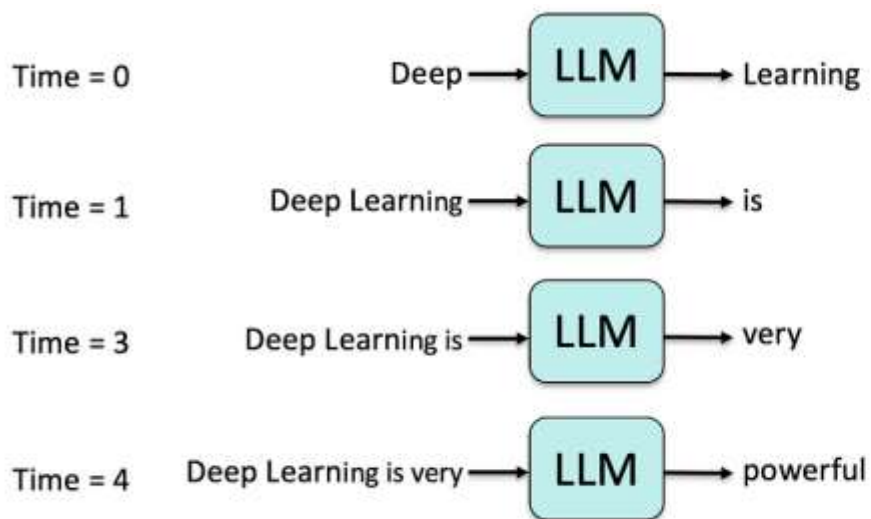
调用DeepSeek服务的流程：文件和联网搜索（RAG）

- 模型的回答来自训练时的数据+外部数据
- 外部数据可以及时更新
 - 比如上传的文件（知识库）或网上搜索的资料（联网搜索）



一些必须要知道的术语概念

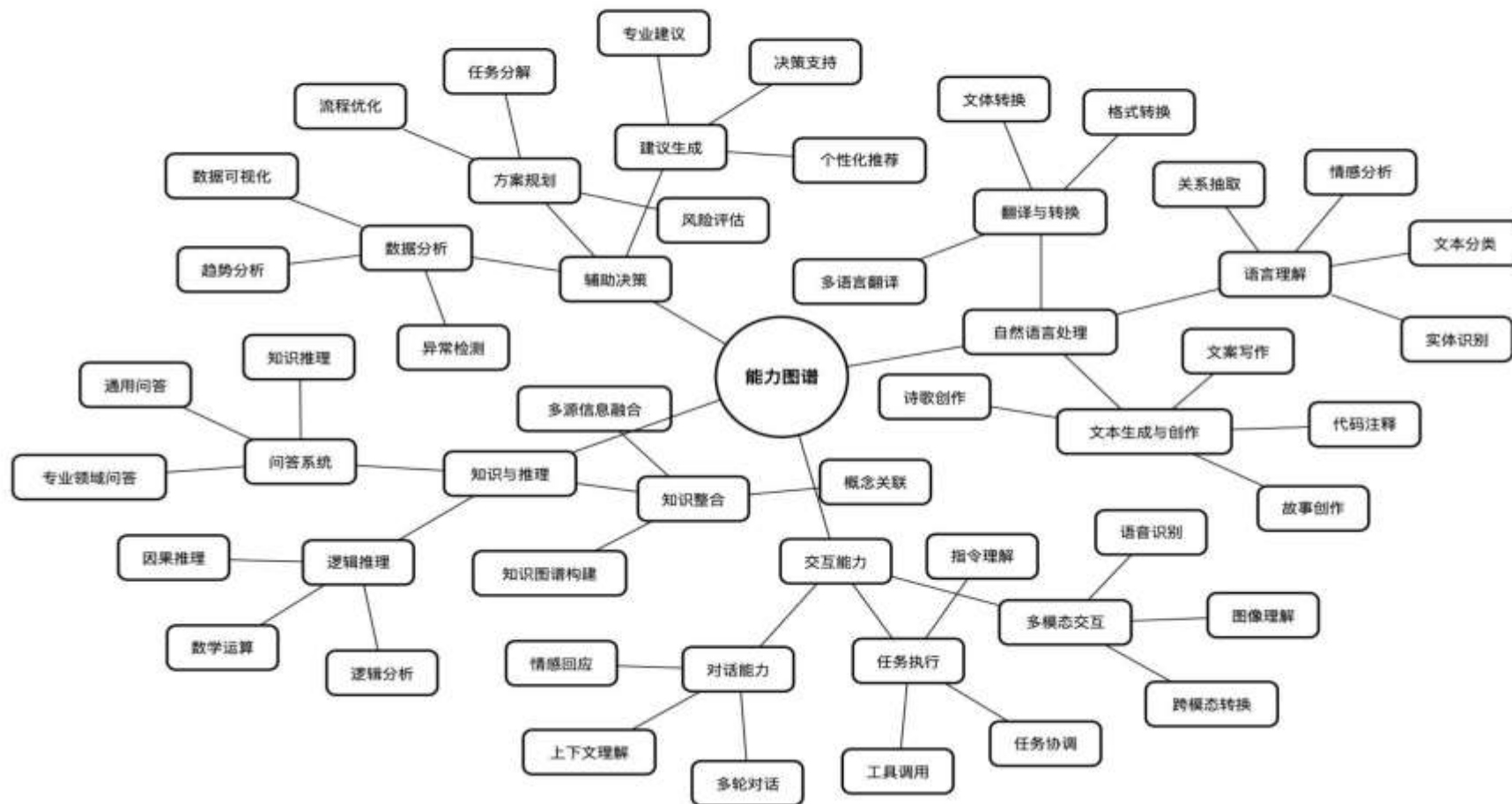
- Prompt:** 用户一次塞给大模型的输入内容
- Token:** 大模型输入输出的最小单位，约等于单词
- 上下文长度:** 当前prompt加上前后对话记录的长度，会一次塞给大模型作为输入
- 训练:** “制作”大模型的过程，将海量的训练数据知识内嵌到模型中
- 推理(inference):** “运行”大模型产生输出内容的过程
- 推理(reasoning):** 一种模型产生输出的方式，将一个大问题拆成多步，好像人类的步步推演



- What is it: DeepSeek是什么
 - 从ChatGPT到DeepSeek-R1, TA到底厉害在哪里?
 - DeepSeek基本概念 (用户角度)
- **How to use it: 我能用DeepSeek干什么**
 - 以小见大, 掌握思维方法
 - 正确理解, 打开广阔天地
- Why it works: DeepSeek背后的原理
 - Transformer——大模型基础
 - DeepSeek模型的发展历程
- Next: 下一步要关注什么
 - 生态的爆发就在眼前, 整个链条上哪些方面值得关注

DeepSeek功能领域一览

直接面向用户或者支持开发者，提供智能对话、文本生成、语义理解、计算推理、代码生成补全等应用场景，支持联网搜索与深度思考模式，同时支持文件上传，能够扫描读取各类文件及图片中的文字内容。



以小见大，掌握思维方法

从一些案例出发，能干的远远比这多。思维方法！思维方法！思维方法！

案例：给我的编译原理书稿提供一个案例片段

🔍 已深度思考 (用时 32 秒) ▾

以下通过中文句子翻译的全过程，类比编译器构造。

中译英案例："快速排序算法高效地处理大规模数据"

1. 词法分析 (Lexical Analysis)

输入处理：
[快速/adj] [排序/vn] [算法/n] [高效地/adv] [处理/v] [大规模数据/n]

输出：

| 词语 | 词性标签 | 附加信息 |
|-------|----------|--------|
| 快速 | ADJ | 修饰"排序" |
| 排序算法 | COMPOUND | 专有名词 |
| 高效地 | ADV | 修饰动词 |
| 处理 | VERB | 及物动词 |
| 大规模数据 | NOUN_PHR | 复合宾语 |

构造一个中译英的例子，对应编译器构造。

2. 语法分析 (Syntax Analysis)

构建短语结构树：

```
S(句子)
├── NP(名词短语)
│   ├── ADJ(快速)
│   └── NOUN(排序算法)
├── ADV(高效地)
└── VP(动词短语)
    ├── VERB(处理)
    └── NP(宾语短语)
        ├── ADJ(大规模)
        └── NOUN(数据)
```

类比编译器：生成抽象语法树(AST)，验证 `for(int i=0; i<n; i++)` 的语法正确性。

3. 语义分析 (Semantic Analysis)

关键检查：

- "处理"的施动者必须是具备数据处理能力的实体 (类型一致性)
- "大规模"与"数据"的量级匹配 (类型一致性)
- "高效地"修饰动词时态合理性 (上下文分析)

类比编译器：检查变量作用域、类型匹配、函数参数匹配。

5. 代码优化 (Optimization)

优化策略：

- 词汇替换：
"处理" → "process" (基础版) → "handle" (更符合计算语境)
- 结构调整：
原句："快速排序算法高效地处理大规模数据"
优化后："The quicksort algorithm efficiently handles large-scale datasets"
(冠词添加、时态统一、名词复数优化)
- 惯用法转换：
"大规模数据" → "big data" (符合技术文档惯用表达)

类比编译器：循环展开、常量传播、死代码消除。

6. 目标代码生成 (Target Code Generation)

生成多版本目标输出：

| 目标场景 | 生成结果 | 特点 |
|-------|---|------------|
| 学术论文 | "The quicksort algorithm demonstrates high efficiency in processing large-scale datasets" | 正式、被动语态 |
| API文档 | "Quicksort.handleBigData(efficient=true)" | 符合编程接口命名规范 |
| 系统日志 | "QSORT: Processing 1.2TB data [EFFICIENT_MODE]" | 简写+结构化数据 |

34 ▾

■ 案例：整理deepseek出来之前gpt o1的收费情况做对比



deepseek 出来之前, gpt o1怎么收费

深度思考 (DeepSeek-R1-联网满血版 671B)

思考过程

深度思考

朗读 改写 复制 下载

关于GPT O1在DeepSeek推出前的收费模式 (2025年1月之前)

1. 基础收费模式

在DeepSeek推出前, OpenAI的GPT O1模型采用按token量计费的模式, 费用分为输入和输出两部分:

输入token: 每百万token收费15美元 (约合人民币108元) ⑥.

输出token: 每百万token收费60美元 (约合人民币432元)

这一价格显著高于同期其他模型 (如GPT-3.5), 但因其其在推理等领域的性能优势, 仍被专业开发者与企业采用。

2. 用户分层与限制

免费用户: 无直接使用权限, 需通过API付费调用, 且未提供免费试用额度。

订阅用户: ChatGPT Plus用户可有限访问部分功能 (如每日150次对话), 但需额外支付每月约20美元的订阅费用。

企业用户: 通过定制化API接口付费使用, 费用根据调用量阶梯式定价, 调用量越大单价越低。

思维导图

DeepSeek推出前GPT o1的收费模式

参考资料 (30)

2月9日, DeepSeek-A

8 网易 2025-02-10

2 DeepSeek刷屏

更为关注的是, Deep

它的成本——或者说

百家号 2025-01-27

3 GPT-5, DeepSei

普通搜索, 想半天关键词,
自己整理各种素材

DS-R1+联网搜索, 自动梳理信息,
并提供引用可检查考证

■ 案例：分析某地公务员录用人员情况

- 如，学历情况如何？
- 多少是计算机相关专业的？

湖南省2023年考试录用公务员湘西州第一批拟录用人员名单

| 招录单位 | 拟录用职位 | 姓名 | 性别 | 准考证号 | 所在工作单位 | 毕业学校及专业 | 学历学位 | 备注 |
|------|-------|----|----|------|--------|---------|---------------|-------|
| | | | | | | | 大学本科 管理学学士 | 应届毕业生 |
| | | | | | | | 大学本科 管理学学士 | 应届毕业生 |
| | | | | | | | 研究生 艺术硕士 | 应届毕业生 |
| | | | | | | | 大学本科 工学学士 | |

录取名单.xlsx
XLSX · 47.97KB

本科及以上学历有多少人

哪些不是的列出来

已深度思考 (用时34秒)

以下是学历不在本科及以上的拟录用人员名单：

输出结果：

| 序号 | 姓名 | 招录单位 | 学历学位 |
|----|----|---------------|------|
| 0 | | 湘西自治州老干部活动中心 | 大学专科 |
| 1 | | 湘西自治州库区移民事务中心 | 大学专科 |
| 2 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 3 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 4 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 5 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 6 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 7 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 8 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 9 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 10 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 11 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 12 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 13 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 14 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 15 | | 湘西自治州强制隔离戒毒所 | 大学专科 |
| 16 | | 龙山县公安局 | 中专 |
| 17 | | 龙山县公安局 | 中专 |
| 18 | | 湘西自治州老干部活动中心 | NaN |
| 19 | | 湘西自治州老干部活动中心 | NaN |
| 20 | | 湘西自治州库区移民事务中心 | NaN |

关键发现：

1. 学历分布：21人中：

- 16人为「大学专科」
- 2人为「中专」
- 3人学历信息缺失 (NaN)

2. 集中岗位：

- 湘西自治州强制隔离戒毒所 (13人)
- 龙山县公安局 (2人)
- 湘西自治州老干部活动中心 (2人)
- 湘西自治州库区移民事务中心 (1人)

3. 特殊说明：

- 强制隔离戒毒所的「警戒护卫」岗位集中出现大专学历人员（可能岗位要求特殊）
- 学历信息缺失的记录需要人工核查原始材料

需要原始Excel文件或进一步分析某个具体岗位，可随时告知。

['序号', '招录单位', '拟录用职位', '姓名', '性别', '准考证号', '所在工作单位', '毕业学校及专业', '学历学位']

列标题已正确设置，其中“学历学位”列包含有关学历的信息。现在，我将计算本科及以上学历的人数。请稍等。

```
# Count the number of people with Bachelor's degree or higher
# We assume that "本科" stands for Bachelor's degree and "研究生" stands for Master's degree or higher

# Clean up the "学历学位" column by removing newline characters and trimming whitespace
df['学历学位'] = df['学历学位'].str.replace('\n', '').str.strip()

# Count the number of people with Bachelor's degree or higher
bachelor_or_higher_count = df[df['学历学位'].str.contains('本科|研究生')].shape[0]
bachelor_or_higher_count
```

409

本科及以上学历的人数共有409人。

案例：分析某专家研究特长，给出研究方向建议咨询

湖南大学陈果的计算机研究做得怎么样？

引用 4 个网页，5 篇公众号文章

已深度思考（用时23秒）

湖南大学陈果教授在计算机领域的研究成果显著，尤其在高性能数据中心网络和RDMA协议栈优化方面具有重要影响力。以下是具体信息：

1. 研究方向与成果

陈果教授长期致力于高性能网络技术研究，聚焦于超算中心网络、数据中心网络及RDMA协议优化。其团队提出的“高性能网络丢包恢复处理机制”被国内知名企业采用，性能指标超越国际同类技术十余倍。此外，他主导的“面向光电协同智算中心网络的高性能RDMA协议”项目，解决了大规模数据交换中的网络拥堵、多路径拥塞控制等难题，显著提升智算中心算力效率。

2. 工业界应用

多项研究成果已落地于华为鲲鹏芯片、腾讯自研交换机、百度无线搜索等产品中，涉及网络协议栈优化、CDN网络加速等领域。例如，针对RDMA协议栈在大规模部署中的瓶颈，他提出低开销架构设计，相关技术被集成至华为芯片。

3. 学术与荣誉

- 发表论文40余篇（含NSDI、ToN等顶会期刊），申请专利10余项；
- 入选国家级青年人才计划、湖南省优青，获湖南省科技进步二等奖、华为最佳技术合作教授等荣誉；
- 担任国家超算长沙中心常务副主任，推动“天河”超算在科研与产业中的应用。

4. 研究理念

他强调科研需紧密结合工业实践，主张“从大型工程系统发现真问题”，通过校企合作解决实际技术瓶颈。其团队与联通、腾讯等企业合作紧密，注重理论向生产力的转化。

综上，陈果教授的研究兼具学术创新与产业价值，是高性能网络领域的领军学者之一。

让他给教育厅领导讲一下deepseek，适合吗

引用 1 个网页，7 篇公众号文章

- 直播回顾 | KPro 18位第六期之“DeepSeek的创造”
- 陈果：从DeepSeek到黄金新高-如何布局2025年
- 著名大学深度解读DeepSeek：原理与效应(44页)
- 抢先看！一个校园版DeepSeek全面对怎样的流程
- 国家超算长沙中心本地化部署DeepSeek服务，助力
- 长沙超算中心DeepSeek服务落地助推自主可控
- 7天疯狂测评后我慌了：DeepSeek太强但这些坑
- 中信建投：DeepSeek本地部署与全球资产配置

已深度思考（用时25秒）

陈果教授作为国家超级计算长沙中心常务副主任，**非常适合**向教育厅领导介绍DeepSeek的技

1. 技术能力与落地经验

陈果团队主导了DeepSeek在湖南大学的本地GPU实现模型推理（文献5）。其团队在清晰展示AI技术在教育场景的可行性。

2. 教育应用场景适配性

- 教学辅助：DeepSeek已用于开发课程率；
- 科研支持：湖南省胸科医院案例显示域可辅助论文写作、数据分析；
- 算力普惠：校园版服务实现自词元响应的智能助手。

3. 政策契合点

- 自主可控：采用国产GPU实现全链路
- 成本优势：通过软件优化降低算力成本预算场景。

除了他，湖南省内还有谁最适合去讲？

引用 4 个网页，4 篇公众号文章

已深度思考（用时50秒）

湖南省内适合向教育厅领导讲解DeepSeek的学者包括以下几位，其研究方向或实践经验与人工智能、教育技术应用高度契合：

1. 陈果教授（湖南大学）

研究方向涵盖智慧教育与学习优化，主持国家自然科学基金项目，成果应用于教育数据挖掘与个性化推荐。其团队在教育场景的技术转化方面具有经验，能够结合DeepSeek展示智能教学辅助系统的开发逻辑。

2. 王德胜教授（湖南大学）

作为机器人视觉感知领域权威，主导国家工程研究中心，其团队在人工智能与产业应用（如智能制造、航空航天）的实践经验，可类比阐述DeepSeek在教育领域的工程化部署路径。

3. 李海林教授（湖南大学）

研究聚焦云计算与人工智能融合技术，主持多项国家自然科学基金项目（如“智能终端的端边云融合研究”），擅长从算力网络架构角度解析大模型技术底座，契合教育信息化基础设施建设的需求。

4. 张军教授（湖南大学）

国家超算长沙中心负责人，其团队研发的“超算与智算融合技术”获国家科技进步奖，可结合天河超算平台案例，阐释DeepSeek在教育公共算力资源调度与优化中的潜力。

推荐优先级建议：

- 若需侧重技术原理与教育场景结合，优先选择陈果或王德胜；
- 若需突出国家级平台背书与产业转化经验，则李海林或张军更合适。

■ 案例：写一个抽签小程序

帮我写一个抽签的网页程序，具备以下功能：
设计一个点击按钮，点一次抽一个；
被抽的名单可以手动上传；
抽之前所有人的名单滚动浮现在屏幕上

✓ 已深度思考（用时213秒）

以下是一个具备您所提功能的抽签网页程序代码：

```
<!DOCTYPE html>
<html>
<head>
<style>
body {
display: flex;
flex-direction: column;
align-items: center;
font-family: Arial, sans-serif;
}

#container {
width: 80%;
margin: 20px;
}
```

```
</script>
</body>
</html>
```

主要功能说明：

- 文件上传：
 - 支持上传.txt文本文件
 - 每行一个姓名，自动过滤空行
 - 支持任意字符编码
- 名单滚动：
 - 上传文件后自动显示横向滚动名单
 - 使用CSS动画实现无缝滚动效果
 - 滚动速度恒定
- 抽签功能：
 - 点击“开始抽签”按钮后名单加速滚动
 - 再次点击“停止抽签”后随机显示结果
 - 结果显示在页面下方

使用方法：

- 准备一个文本文件，每行写一个姓名
- 点击“选择文件”上传准备好的名单文件
- 点击“开始抽签”按钮启动滚动
- 再次点击按钮停止滚动并显示结果



■ 案例：设计一个针对幼儿园小朋友的科普讲稿

好的！以下是为幼儿园小朋友

标题：神奇的计算机小精灵

（老师拿出一个玩具电脑或
“精灵”哦！你们见过它吗？在
（等待小朋友回答：家里、

第一部分：计算机像什么？

👉：计算机呀，长得像一
（展示图片：台式机、笔记

👉：有的计算机像一本厚
台式机）！它们都有一个亮

第二部分：计算机的“身体”

👉：计算机小精灵的身体

1 键盘：像一排排小牙齿！

2 鼠标：像一只小老鼠！

3 摄像头：是计算机的小

（邀请小朋友模仿键盘、鼠

第三部分：计算机能做什么？

👉：计算机小精灵可厉害

第三部分：计算机能

👉：计算机小精

变出彩虹：帮我

讲故事：放动画

学数学：1+1=？

（提问互动：你们想

第四部分：计算机的

👉：其实呀，计

大脑：叫“CPU”，

记忆本：叫“硬盘

（用夸张手势模仿C

结尾：和计算机做朋

👉：计算机是我

不贪玩：看屏幕

不乱点：遇到奇

分享爱：和爸爸

（带领小朋友做拉钩

道具建议：

1. 用彩纸剪出键盘

2. 准备简单电脑操

3. 奖励贴纸：“计

已深度思考（用

好的，用户之前让我

计算机的原理。首先，

比。

超级计算机的核心特

物或交通工具来比喻

接下来，要解释超级

用蚂蚁搬食物的例子

另外，超级计算机的

的角色来举例，让他

还需要注意互动环节

度。这样不仅活跃气

最后，要确保整个讲

解更生动。同时，结

总结下来，关键点是

保持积极和鼓励的氛

标题：超级计算机——一万个脑袋

（老师举起一只手，假装变魔术）

👉：小朋友们，如果老师有一只

（小朋友抢答：50颗！）

👉：对啦！超级计算机就像一万

第一部分：超级计算机有多厉害？

（用夸张动作+拟声词）

👉：普通计算机像一辆小自行车

动）

“嗖——！”一秒钟就能绕地球跑7圈！

（邀请小朋友一起喊：“超级快！超级

第二部分：超级计算机的秘密武器

👉：它的秘密就是——好多好多

（画黑板：1个大脑 vs 10000个大脑）

1 普通计算机：1个小脑袋想问题

2 超级计算机：一万个脑袋手拉手

（让小朋友模仿“蚂蚁搬运工”，一起

第三部分：超级计算机能做什么？

（用小朋友熟悉的场景举例）

👉：预测天气：

“明天会下雨吗？”超级计算机眨眨眼：

👉：探索太空：

试着把超级计算机的原理给小朋友讲明白

“明天会下雨吗？”超级计算机眨眨眼：“云朵阿姨下午3点会哭鼻子哦！”（模仿下雨哗啦啦）

👉：探索太空：

“火箭爷爷怎么飞向月亮？”超级计算机挥挥手：“往左拐！往右拐！砰！降落啦！”（做火箭摇摆动作）

👉：制作动画：

“艾莎公主的冰雪魔法❄️，就是超级计算机变出的十万片小雪花！”（抛撒亮片模拟雪花）

第四部分：超级计算机住在哪里？

（神秘语气+手势比划）

👉：它住在比游泳池还大的“数字宫殿”里！（展开双臂画大圈）

👉：需要喝很多“电奶茶”——插满1000根充电线！

👉：还要吹巨型空调——不然会热得冒烟啦！（模仿擦汗喘气）

结尾：你们也能成为“小超人”

👉：虽然我们没有一万个小脑袋，但只要——

👉：团结合作！（和旁边小朋友手拉手）

👉：爱动脑筋！（用手指在头顶画圈圈）

👉：勇敢提问！（双手放在嘴边做喇叭状）

未来的超级计算机，可能就在你们中间诞生哦！🌟

（全体举手欢呼：“我要当数字超人！”）

互动小实验（可选）：

1. 集体数数比赛：1个小朋友数1-10 vs 全班一起数，感受“合作加速”

2. 拼图接力：每人拼一块，体验“分工完成大任务” ↓

正确理解，打开广阔天地

正确理解DeepSeek的**不能**

■ 还不是AGI, 不能“一步到位”!

□ 需要用户自己具备一定的**问题拆解能力**、**信息整合能力**、**迭代调优能力**

41

DeepSeek (R1或V3) 的不能

- 有很强的能力，**但也经常出错**；预载很多知识，但**不知道所有的知识**！
 - 利用其能力，判断其结果，改进其知识！



“尽信书不如无书”！要有判断筛选能力，擅用联网搜索和知识库！

DeepSeek (R1或V3) 的不能

- R1/V3都是语言模型，**不能直接处理多模态数据**（图片、视频等）！
 - 学会利用其它工具，一起来完成任务（智能体的思想源于此）

画一个fatree网络拓扑结构：由相同的交换机组成，每个交换机都有4个端口，总共有12层交换机

已深度思考（用时22秒）

以下是基于对称Fatree架构，全同交

拓扑结构参数

| 层级 | 交换机数量 | 角色 |
|-----|-------|-----|
| 核心层 | 1 | 汇聚层 |
| 聚合层 | 4 | 核心层 |
| 接入层 | 12 | 终端层 |

拓扑示意图

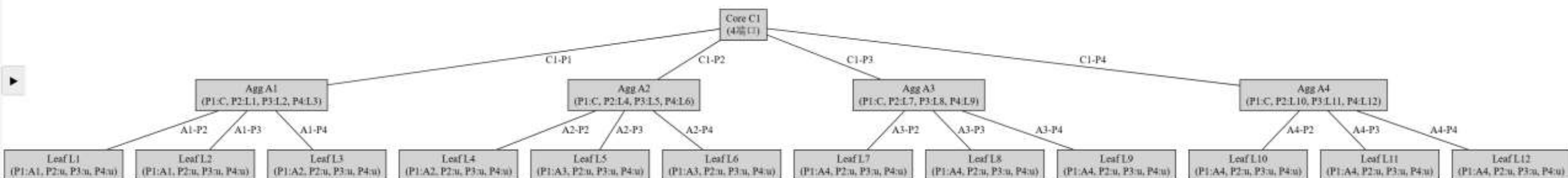
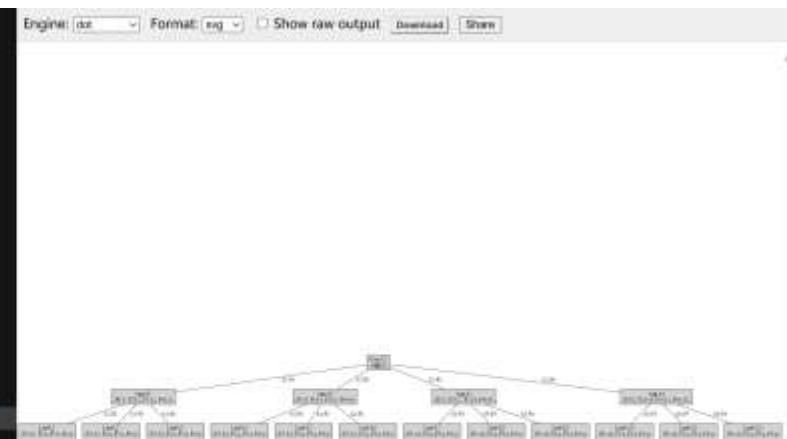


已深度思考（用时12秒）

以下是基于对称三层Fatree架构（核心层1个交换机，聚合层4个交换机4端口）的Graphviz DOT代码：

```
graph TD
    subgraph Fatree
        direction TB
        C1[Core C1] --- A1[Agg A1]
        C1 --- A2[Agg A2]
        C1 --- A3[Agg A3]
        C1 --- A4[Agg A4]
        A1 --- L1[Leaf L1]
        A1 --- L2[Leaf L2]
        A1 --- L3[Leaf L3]
        A2 --- L4[Leaf L4]
        A2 --- L5[Leaf L5]
        A2 --- L6[Leaf L6]
        A3 --- L7[Leaf L7]
        A3 --- L8[Leaf L8]
        A3 --- L9[Leaf L9]
        A4 --- L10[Leaf L10]
        A4 --- L11[Leaf L11]
        A4 --- L12[Leaf L12]
    end
```

```
graph TD
    subgraph Fatree
        direction TB
        C1[Core C1] --- A1[Agg A1]
        C1 --- A2[Agg A2]
        C1 --- A3[Agg A3]
        C1 --- A4[Agg A4]
        A1 --- L1[Leaf L1]
        A1 --- L2[Leaf L2]
        A1 --- L3[Leaf L3]
        A2 --- L4[Leaf L4]
        A2 --- L5[Leaf L5]
        A2 --- L6[Leaf L6]
        A3 --- L7[Leaf L7]
        A3 --- L8[Leaf L8]
        A3 --- L9[Leaf L9]
        A4 --- L10[Leaf L10]
        A4 --- L11[Leaf L11]
        A4 --- L12[Leaf L12]
    end
```



DeepSeek (R1或V3) 的不能

- 模型都有**上下文长度限制**，不能塞太多东西给他！
 - 学会拆分任务，总结规律

通道2-DeepSeek-R1

```

15 // 分配栈空间(n)
16 // 加入符号表
17 allocateStackSpace(node->token->value, 8);
18
19 break;
20 case IDENTIFIER:
21 // 加载变量到寄存器
22 if (node->valueType == INT_TYPE) {
23     emit(ctx, "ldr %d, [%d, %d]"); // 加载变量到寄存器, node->value_register, node->ctx
24 }
25 break;
26 case BINARY_OPERATOR:
27 // 生成二元运算指令
28 char op[10];
29 if (node->token->type == Tk_Operator) {
30     if (strcmp(node->token->value, "+") == 0) {
31         sprintf(op, "add");
32     } else if (strcmp(node->token->value, "-") == 0) {
33         sprintf(op, "sub");
34     } else if (strcmp(node->token->value, "*") == 0) {
35         sprintf(op, "mul");
36     } else if (strcmp(node->token->value, "/") == 0) {
37         sprintf(op, "div");
38     }
39 // 生成操作数加载指令
40 generateCode(node->left, ctx);
41 generateCode(node->right, ctx);
42 // 生成运算指令
43 emit(ctx, "%s %d, %d, %d");

```

输入问题

+ 代码解释器



你一个对话框里的聊天记录都会塞进模型里去，一次聊天不能聊天多☺
(一般128K tokens是目前通常的最高水平)

■ **不是唯一**的大模型，效果也**难说一骑绝尘**，其本身也**不是一成不变**！

□ 客观辩证地看待，积极开放地拥抱



| 云平台厂商 | 科技厂商 | 电信运营商 | 研究机构 |
|---|---|--|---|
|  Google Cloud  Google Cloud  百度  腾讯  华为  阿里云 |  OpenAI  Meta  ANTHROPIC  xAI  Stability AI  360  科大讯飞  商汤科技  智谱 AI  讯飞星火  零一万物  月之暗面  MiniMax  小冰  OPPO  vivo  HUAWEI  出门问问  商汤科技  深智智能  DJI  九章云脑  达摩院 |  中国移动  中国联通  中国电信 |  斯坦福大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学  清华大学 |

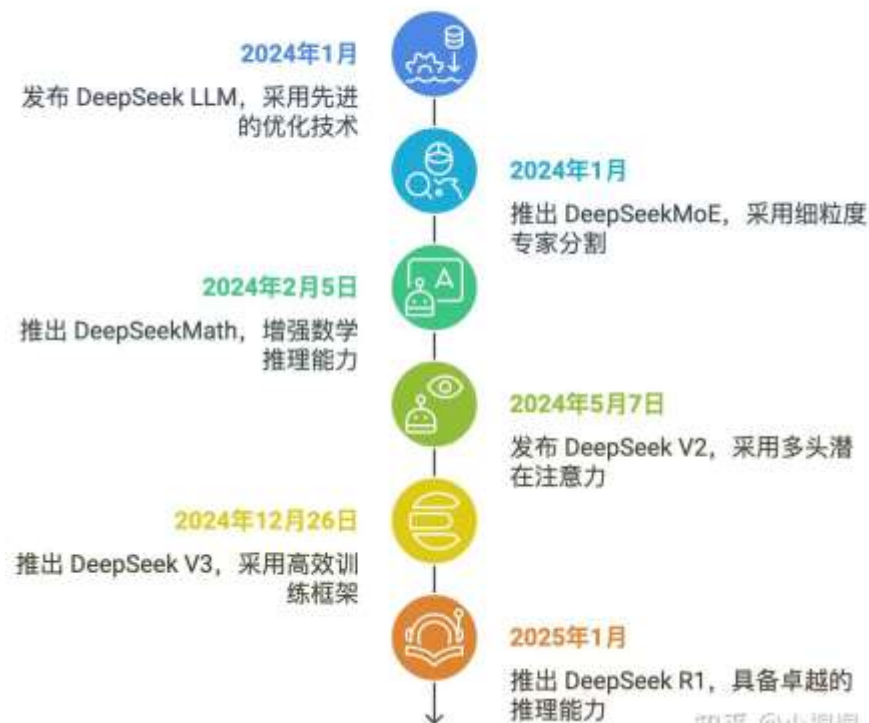
模型很多，各有所长

(长上下文、多模态、逻辑推理、多语言, 不同模型卖点不同)

2024全球AIGC产业全景图谱及报告重磅发布---至顶网

Deepseek发展历程概要 - 知乎

DeepSeek AI 模型的演变



日新月异，进展很快

(重要新技术以周为单位出现)

对待DeepSeek等最新大模型的正确态度

普通软件工具



帮助掌握领域知识和技能的人，摆脱**重复**
低级的脑力劳动

上一代大模型



帮助掌握领域知识和技能的人，摆脱**一部**
分中级脑力劳动

新一代大模型



希望达到的目标：帮助大**部**
分的普通人，摆脱**一部**
分中
级甚至是高级脑力劳动

对待DeepSeek等最新大模型的正确态度

**大模型就像一个小朋友，具备了初级“智能”：
懂一点，但不全懂；知识有一点，但也不全有；有时能对，但也经常犯错**



发挥你的智慧，利用各种现有工具，引导他、帮助他干活！
用的好，可以帮你减轻很大工作量，小朋友的能力能超乎你想象；用的不好，那就是熊孩子☺

以小见大，掌握思维方法；正确理解，打开广阔天地

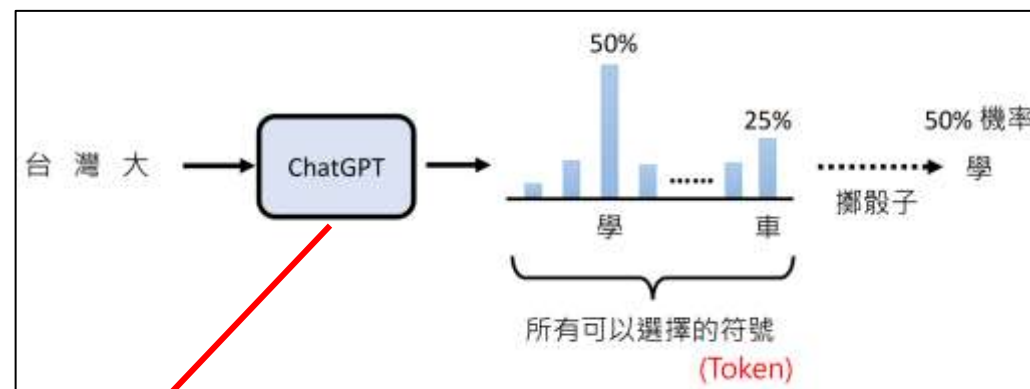
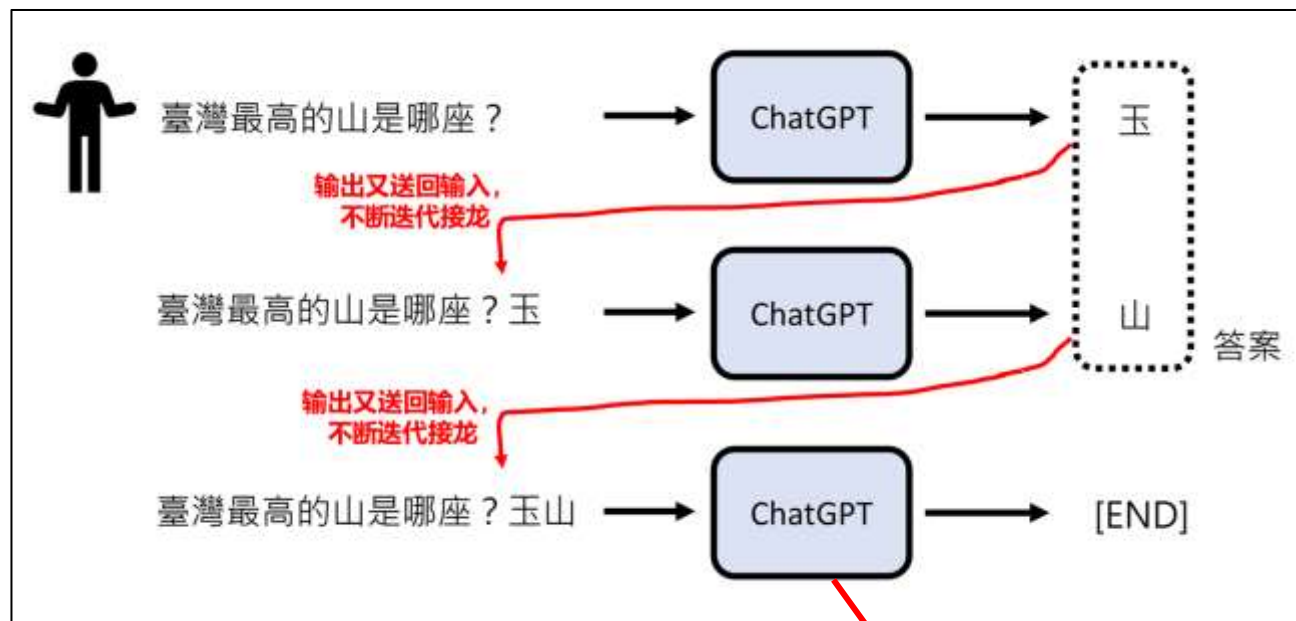
- 重点是掌握使用TA的思维方法
 - 案例很多，无法一一列举
- 知道TA有哪些能力
 - 逻辑推理能力、文字生成能力、搜索总结能力、代码生成能力。。。
- 更重要的是知道TA有哪些不能！
 - 不能“**一步到位**”、可能经常出错、不能直接生成文件、上下文不能无限长。。。
- 充分认识TA的能与不能
 - 组合多种工具一起使用！
 - 取其所能，博采众长！

**会不会用，即将成为现代社会生产效率的分水岭！
会用的人或组织，会远远甩开那些不会用的！
发挥你的创造力和能动性，赶紧用起来吧！**

- What is it: DeepSeek是什么
 - 从ChatGPT到DeepSeek-R1, TA到底厉害在哪里?
 - DeepSeek基本概念 (用户角度)
- How to use it: 我能用DeepSeek干什么
 - 以小见大, 掌握思维方法
 - 正确理解, 打开广阔天地
- **Why it works: DeepSeek背后的原理**
 - **Transformer——大模型基础**
 - **DeepSeek模型的发展历程**
- Next: 下一步要关注什么
 - 生态的爆发就在眼前, 整个链条上哪些方面值得关注

Transformer——大模型基础

回忆一下我们在第一部分讲的大模型原理



这个框框里是啥？为啥能根据不同的输入上下文选择对的输出token？

几个必须澄清的概念

人工智能 (目标)

机器学习 (手段)

神经网络 (更厉害的手段)

深度学习
(很深的神经网络)

大模型
(LLM)

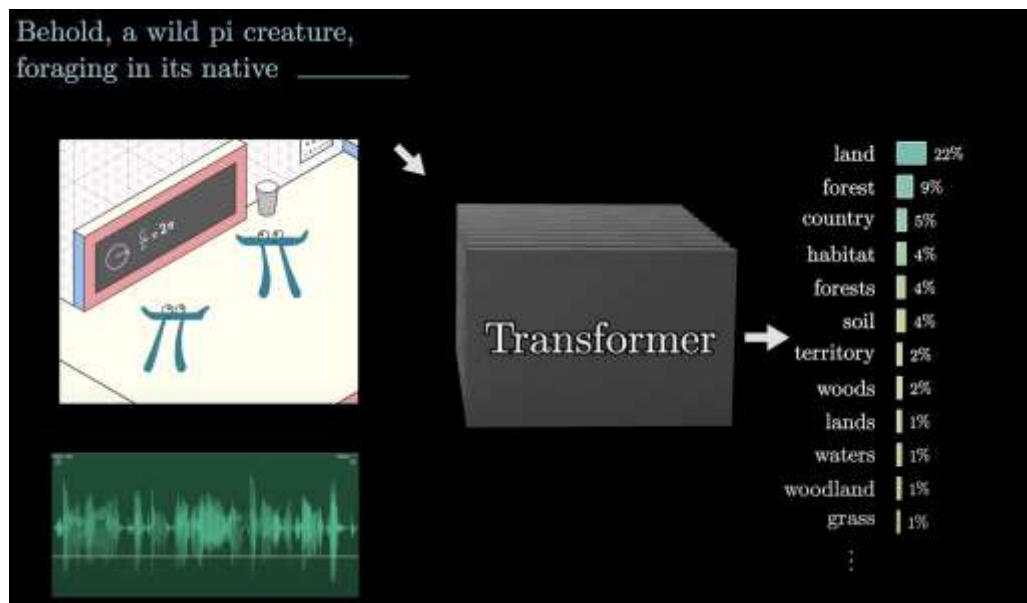
ChatGPT

DeepSeek

Transformer
大模型常用的一种神经网络

Transformer是什么

- Transformer是一种特殊的神经网络，几乎现在所有典型大模型都采用这种神经网络
 - 有很多类型voice-to-text, text-to-voice, text-to-image。。。
- 我们主要介绍text-to-text transformer，是现在主流大模型的基础
 - 输入：text（可能伴随一些图像或声音等），输出：预测下一个token

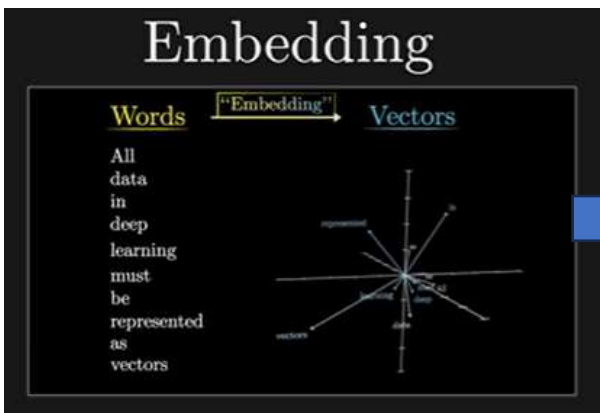


Transformer整体流程速览

通常重复很多次

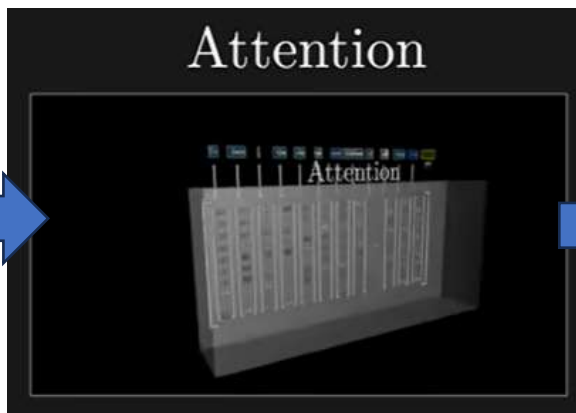
输入token编码

Embedding



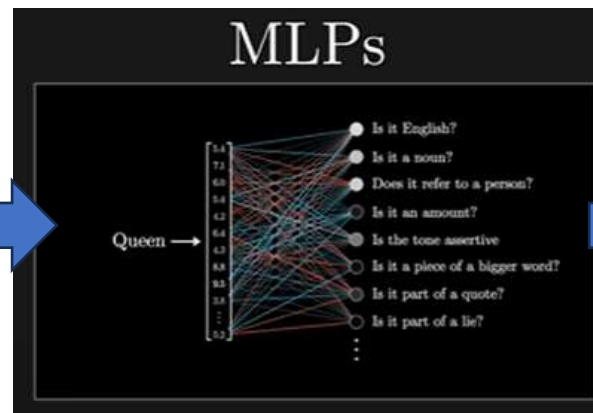
计算token之间的关系

Attention



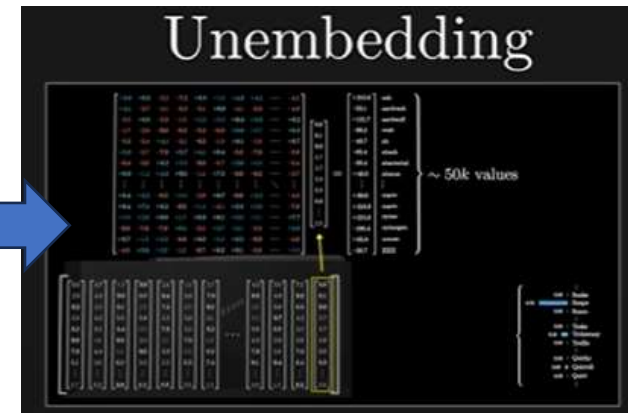
理解每个token自己的含义

MLPs

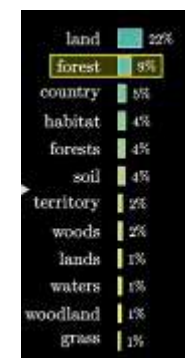


编码还原成token并输出

Unembedding



Behold, a wild pi
creature, foraging in its
native



land

Embedding

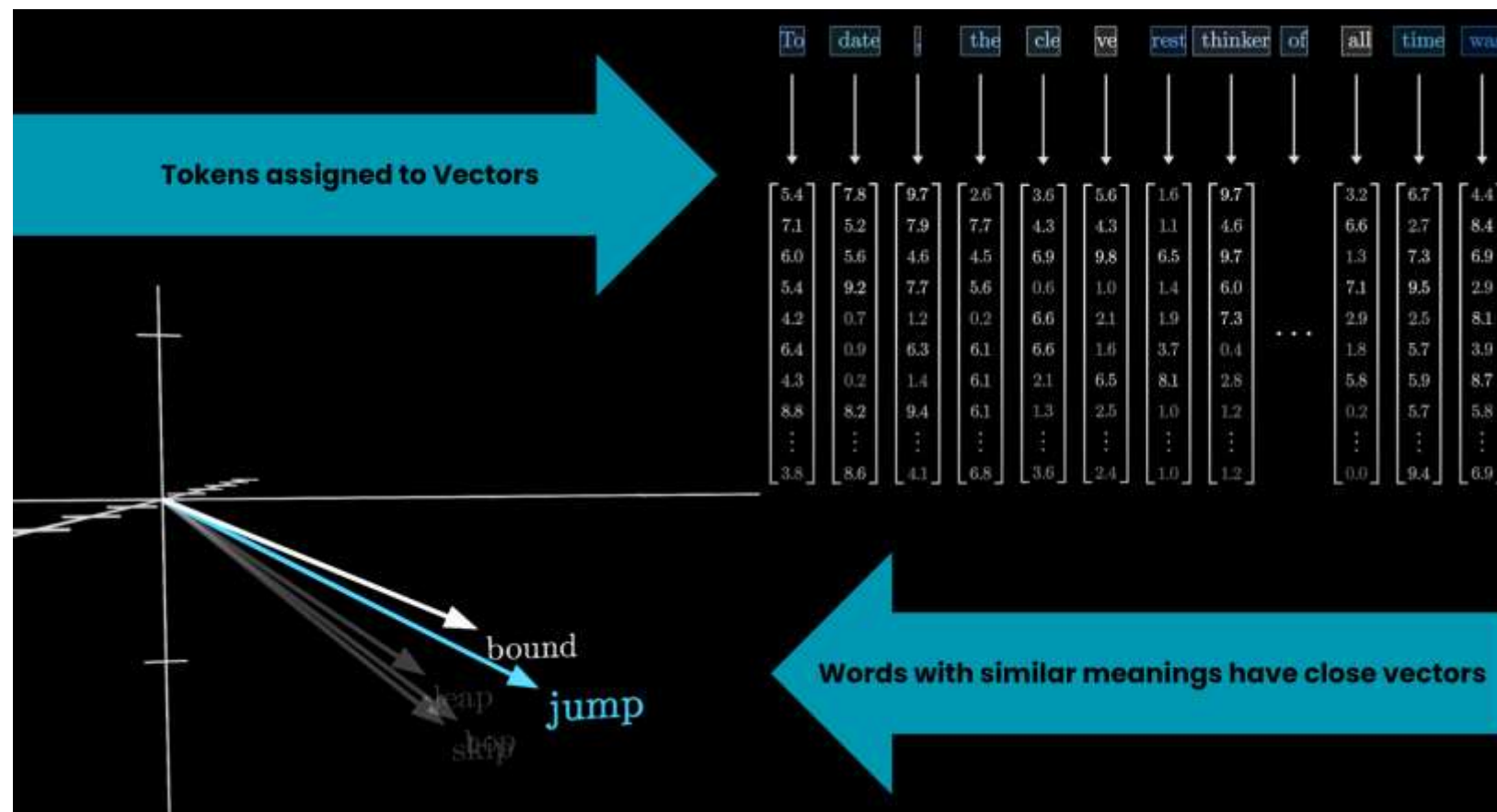
- 把输入的token编码成向量
 - 以**特定权重矩阵**对各token的原始向量相乘，编码成特定向量

To| date|,| the| cle|ve|rest|
thinker| of| all| time| was ...

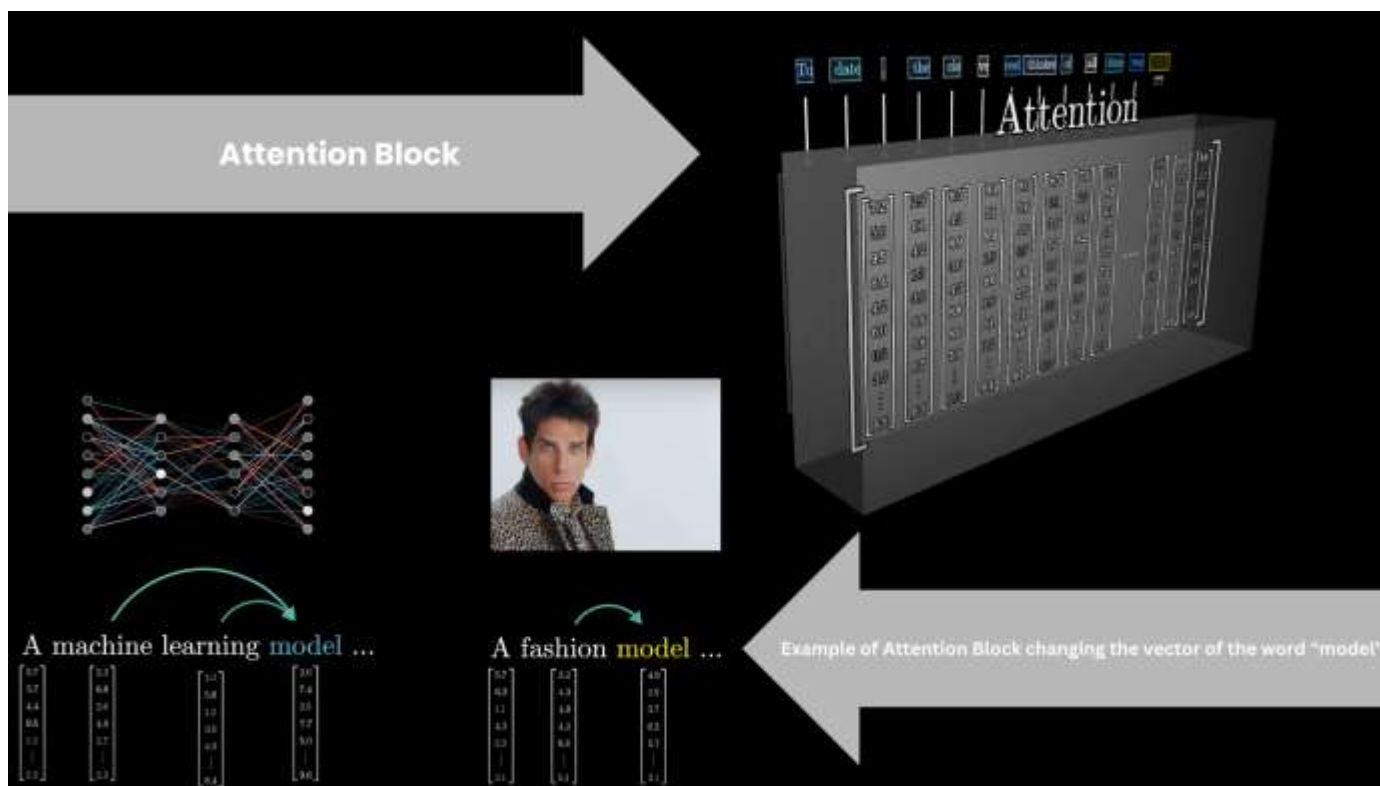
embedding

分词器

To date, the cleverest
thinker of all time was ...

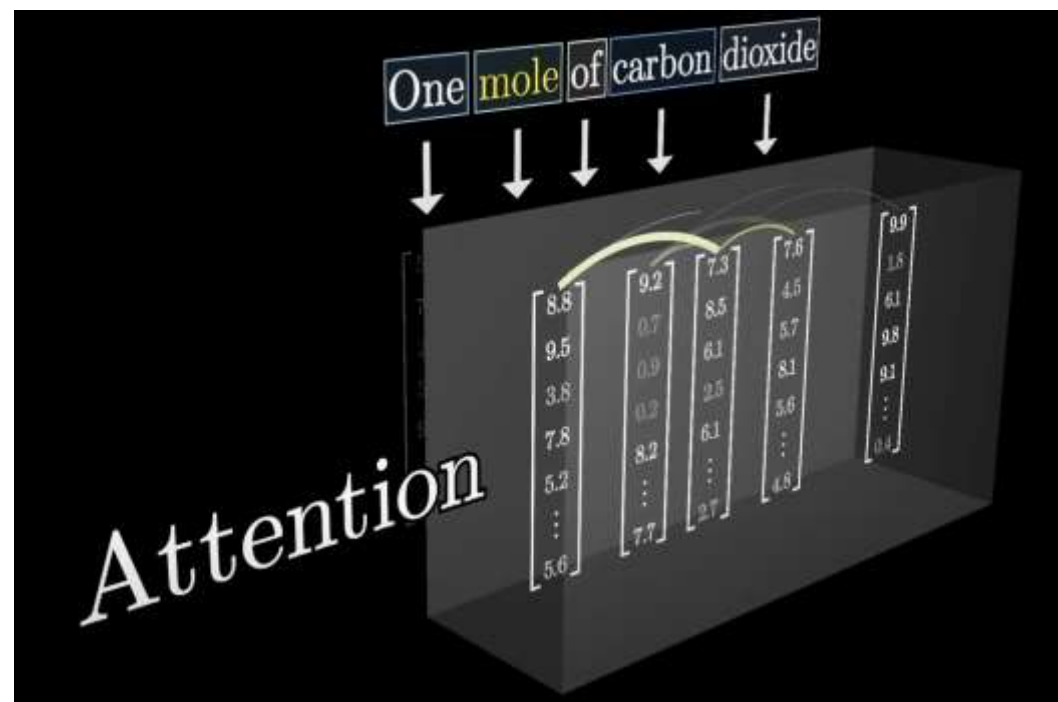
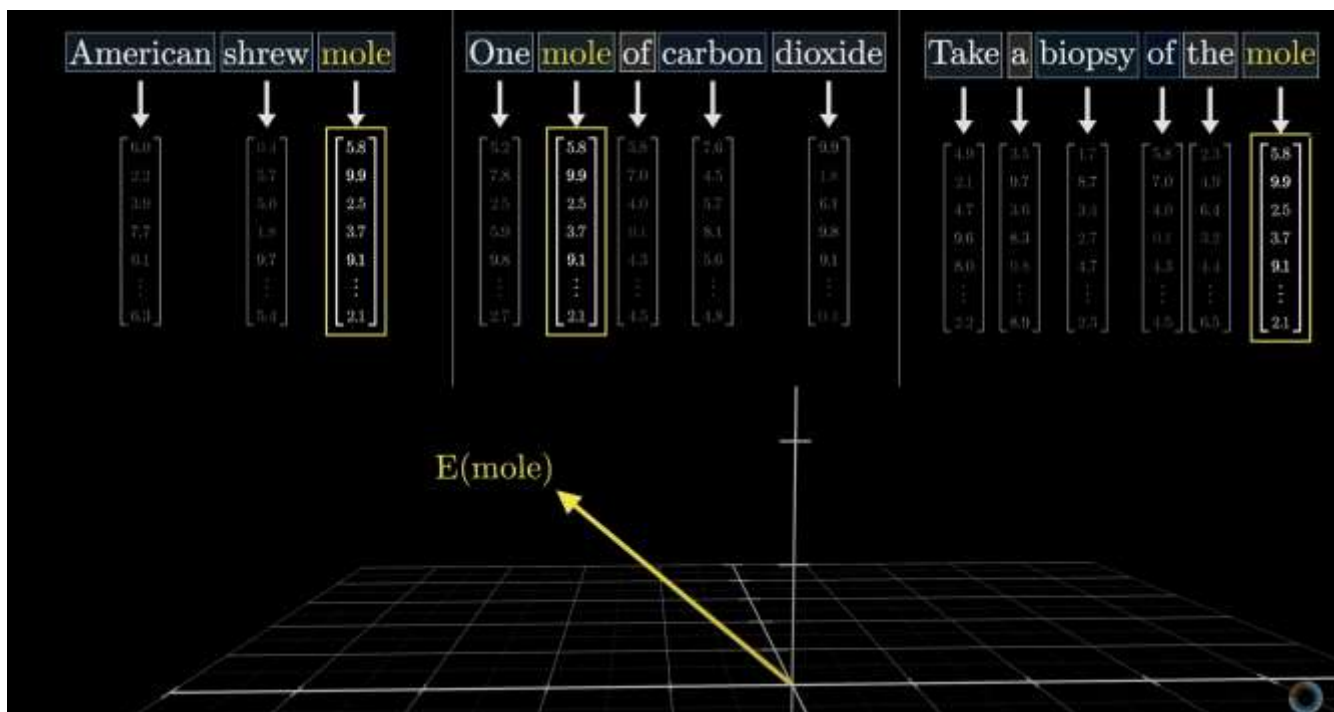


- 注意力机制：计算token之间的关系
 - 每个token的向量之间，以**特定权重矩阵交叉相乘**，从而计算token之间的互相影响，把影响后的含义编码到乘完之后的token向量中



Attention: 多说两句

- 三句话都有mole这个词（鼹鼠、摩尔、痣），如何区分？
- Attention会通过矩阵运算把周边词的意思嵌入到mole的向量中，反应其在上下文中的含义

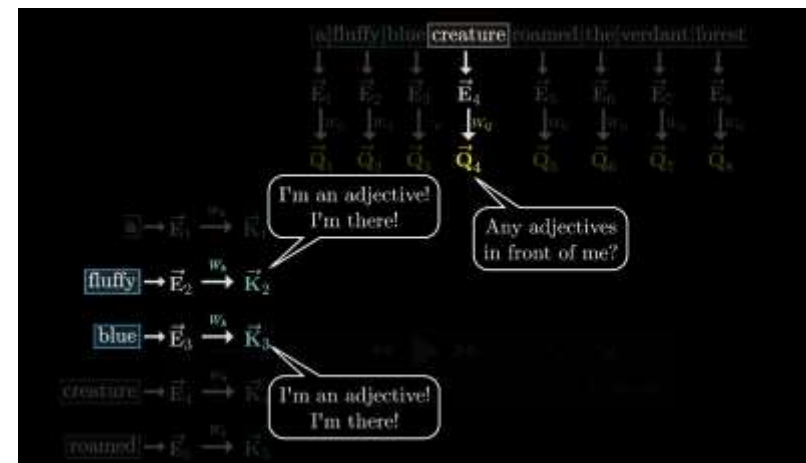


Attention: 多说两句

- 如何嵌入上下文含义？
 - Q（我查）、K（查谁）、V（结果）

Attention

| | a | fluffy | blue | creature | roamed | the | verdant | forest |
|--|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| | \vec{E}_1 | \vec{E}_2 | \vec{E}_3 | \vec{E}_4 | \vec{E}_5 | \vec{E}_6 | \vec{E}_7 | \vec{E}_8 |
| | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ |
| | \vec{Q}_1 | \vec{Q}_2 | \vec{Q}_3 | \vec{Q}_4 | \vec{Q}_5 | \vec{Q}_6 | \vec{Q}_7 | \vec{Q}_8 |
| a $\rightarrow \vec{E}_1 \xrightarrow{w_K} \vec{K}_1$ | $\vec{K}_1 \cdot \vec{Q}_1$ | $\vec{K}_1 \cdot \vec{Q}_2$ | $\vec{K}_1 \cdot \vec{Q}_3$ | $\vec{K}_1 \cdot \vec{Q}_4$ | $\vec{K}_1 \cdot \vec{Q}_5$ | $\vec{K}_1 \cdot \vec{Q}_6$ | $\vec{K}_1 \cdot \vec{Q}_7$ | $\vec{K}_1 \cdot \vec{Q}_8$ |
| fluffy $\rightarrow \vec{E}_2 \xrightarrow{w_K} \vec{K}_2$ | $\vec{K}_2 \cdot \vec{Q}_1$ | $\vec{K}_2 \cdot \vec{Q}_2$ | $\vec{K}_2 \cdot \vec{Q}_3$ | $\vec{K}_2 \cdot \vec{Q}_4$ | $\vec{K}_2 \cdot \vec{Q}_5$ | $\vec{K}_2 \cdot \vec{Q}_6$ | $\vec{K}_2 \cdot \vec{Q}_7$ | $\vec{K}_2 \cdot \vec{Q}_8$ |
| blue $\rightarrow \vec{E}_3 \xrightarrow{w_K} \vec{K}_3$ | $\vec{K}_3 \cdot \vec{Q}_1$ | $\vec{K}_3 \cdot \vec{Q}_2$ | $\vec{K}_3 \cdot \vec{Q}_3$ | $\vec{K}_3 \cdot \vec{Q}_4$ | $\vec{K}_3 \cdot \vec{Q}_5$ | $\vec{K}_3 \cdot \vec{Q}_6$ | $\vec{K}_3 \cdot \vec{Q}_7$ | $\vec{K}_3 \cdot \vec{Q}_8$ |
| creature $\rightarrow \vec{E}_4 \xrightarrow{w_K} \vec{K}_4$ | $\vec{K}_4 \cdot \vec{Q}_1$ | $\vec{K}_4 \cdot \vec{Q}_2$ | $\vec{K}_4 \cdot \vec{Q}_3$ | $\vec{K}_4 \cdot \vec{Q}_4$ | $\vec{K}_4 \cdot \vec{Q}_5$ | $\vec{K}_4 \cdot \vec{Q}_6$ | $\vec{K}_4 \cdot \vec{Q}_7$ | $\vec{K}_4 \cdot \vec{Q}_8$ |
| roamed $\rightarrow \vec{E}_5 \xrightarrow{w_K} \vec{K}_5$ | $\vec{K}_5 \cdot \vec{Q}_1$ | $\vec{K}_5 \cdot \vec{Q}_2$ | $\vec{K}_5 \cdot \vec{Q}_3$ | $\vec{K}_5 \cdot \vec{Q}_4$ | $\vec{K}_5 \cdot \vec{Q}_5$ | $\vec{K}_5 \cdot \vec{Q}_6$ | $\vec{K}_5 \cdot \vec{Q}_7$ | $\vec{K}_5 \cdot \vec{Q}_8$ |
| the $\rightarrow \vec{E}_6 \xrightarrow{w_K} \vec{K}_6$ | $\vec{K}_6 \cdot \vec{Q}_1$ | $\vec{K}_6 \cdot \vec{Q}_2$ | $\vec{K}_6 \cdot \vec{Q}_3$ | $\vec{K}_6 \cdot \vec{Q}_4$ | $\vec{K}_6 \cdot \vec{Q}_5$ | $\vec{K}_6 \cdot \vec{Q}_6$ | $\vec{K}_6 \cdot \vec{Q}_7$ | $\vec{K}_6 \cdot \vec{Q}_8$ |
| verdant $\rightarrow \vec{E}_7 \xrightarrow{w_K} \vec{K}_7$ | $\vec{K}_7 \cdot \vec{Q}_1$ | $\vec{K}_7 \cdot \vec{Q}_2$ | $\vec{K}_7 \cdot \vec{Q}_3$ | $\vec{K}_7 \cdot \vec{Q}_4$ | $\vec{K}_7 \cdot \vec{Q}_5$ | $\vec{K}_7 \cdot \vec{Q}_6$ | $\vec{K}_7 \cdot \vec{Q}_7$ | $\vec{K}_7 \cdot \vec{Q}_8$ |
| forest $\rightarrow \vec{E}_8 \xrightarrow{w_K} \vec{K}_8$ | $\vec{K}_8 \cdot \vec{Q}_1$ | $\vec{K}_8 \cdot \vec{Q}_2$ | $\vec{K}_8 \cdot \vec{Q}_3$ | $\vec{K}_8 \cdot \vec{Q}_4$ | $\vec{K}_8 \cdot \vec{Q}_5$ | $\vec{K}_8 \cdot \vec{Q}_6$ | $\vec{K}_8 \cdot \vec{Q}_7$ | $\vec{K}_8 \cdot \vec{Q}_8$ |

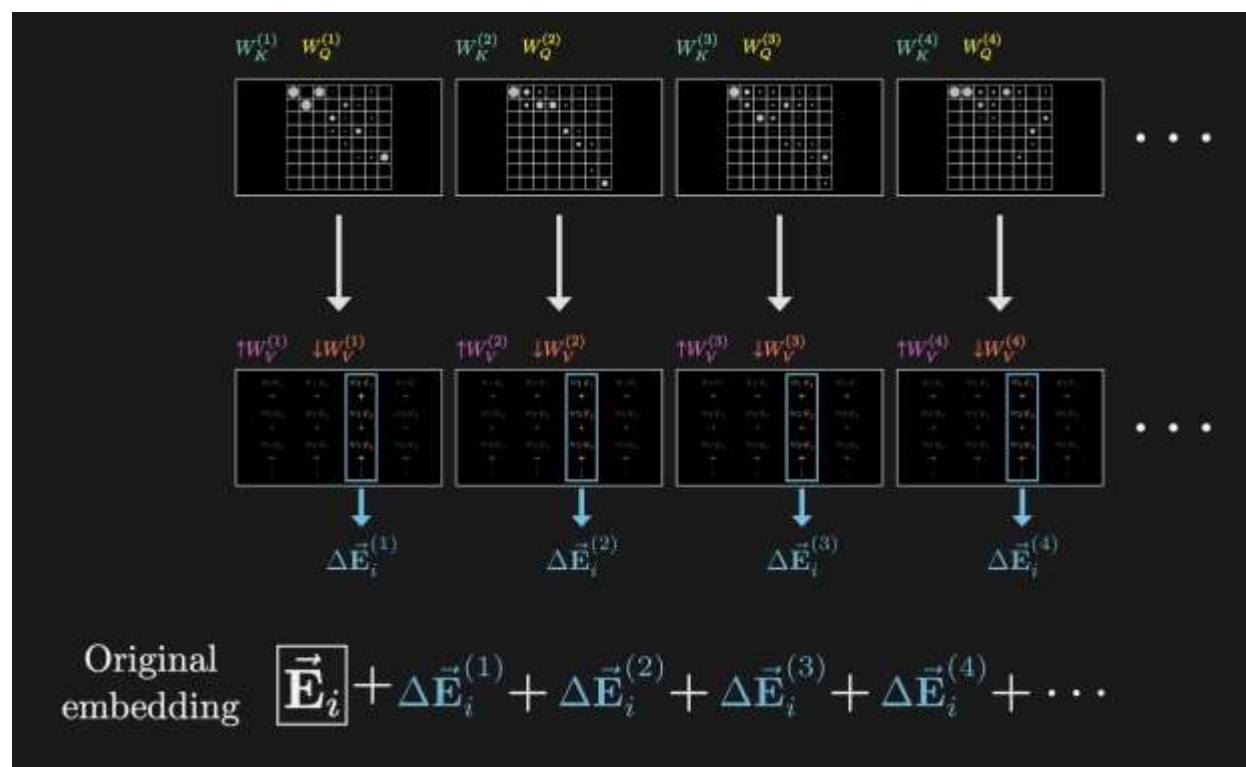
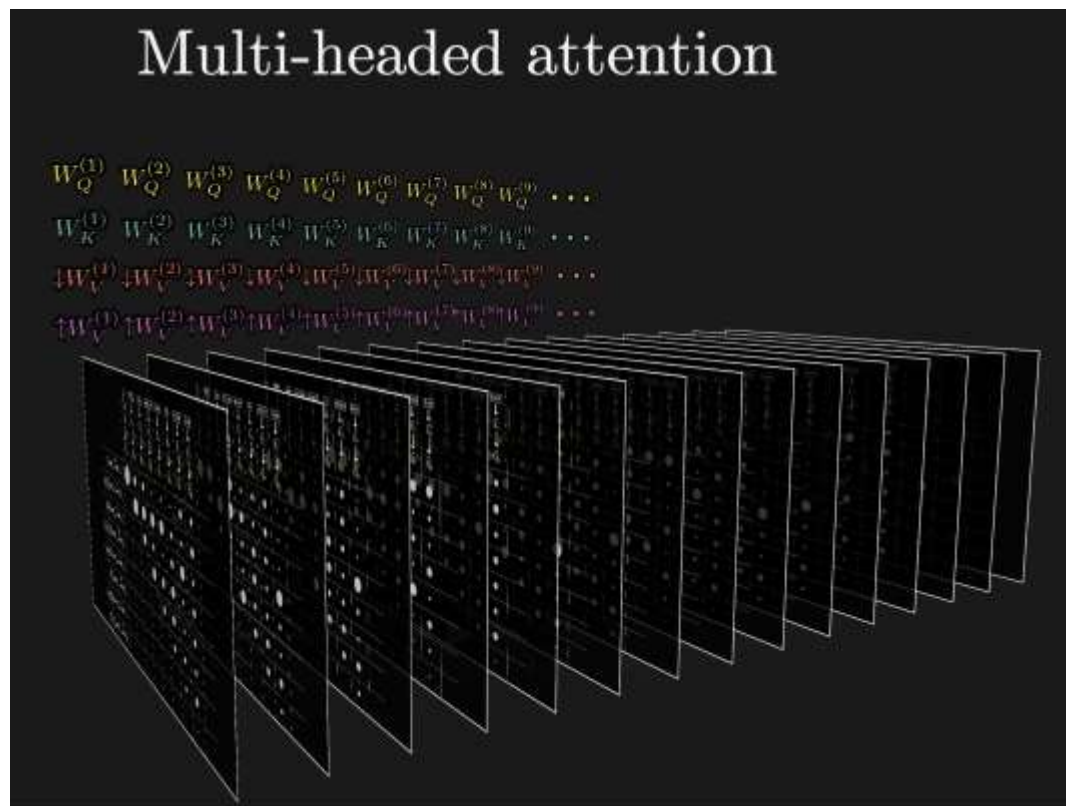


| | a | fluffy | blue | creature | roamed | the | verdant | forest |
|--|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| | \vec{E}_1 | \vec{E}_2 | \vec{E}_3 | \vec{E}_4 | \vec{E}_5 | \vec{E}_6 | \vec{E}_7 | \vec{E}_8 |
| | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ | $\downarrow w_Q$ |
| | \vec{Q}_1 | \vec{Q}_2 | \vec{Q}_3 | \vec{Q}_4 | \vec{Q}_5 | \vec{Q}_6 | \vec{Q}_7 | \vec{Q}_8 |
| a $\rightarrow \vec{E}_1 \xrightarrow{w_K} \vec{K}_1$ | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| fluffy $\rightarrow \vec{E}_2 \xrightarrow{w_K} \vec{K}_2$ | 0.00 | 1.00 | 0.00 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 |
| blue $\rightarrow \vec{E}_3 \xrightarrow{w_K} \vec{K}_3$ | 0.00 | 0.00 | 1.00 | 0.58 | 0.00 | 0.00 | 0.00 | 0.00 |
| creature $\rightarrow \vec{E}_4 \xrightarrow{w_K} \vec{K}_4$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| roamed $\rightarrow \vec{E}_5 \xrightarrow{w_K} \vec{K}_5$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| the $\rightarrow \vec{E}_6 \xrightarrow{w_K} \vec{K}_6$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 | 1.00 | 0.00 | 0.00 |
| verdant $\rightarrow \vec{E}_7 \xrightarrow{w_K} \vec{K}_7$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| forest $\rightarrow \vec{E}_8 \xrightarrow{w_K} \vec{K}_8$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Attention: 多说两句

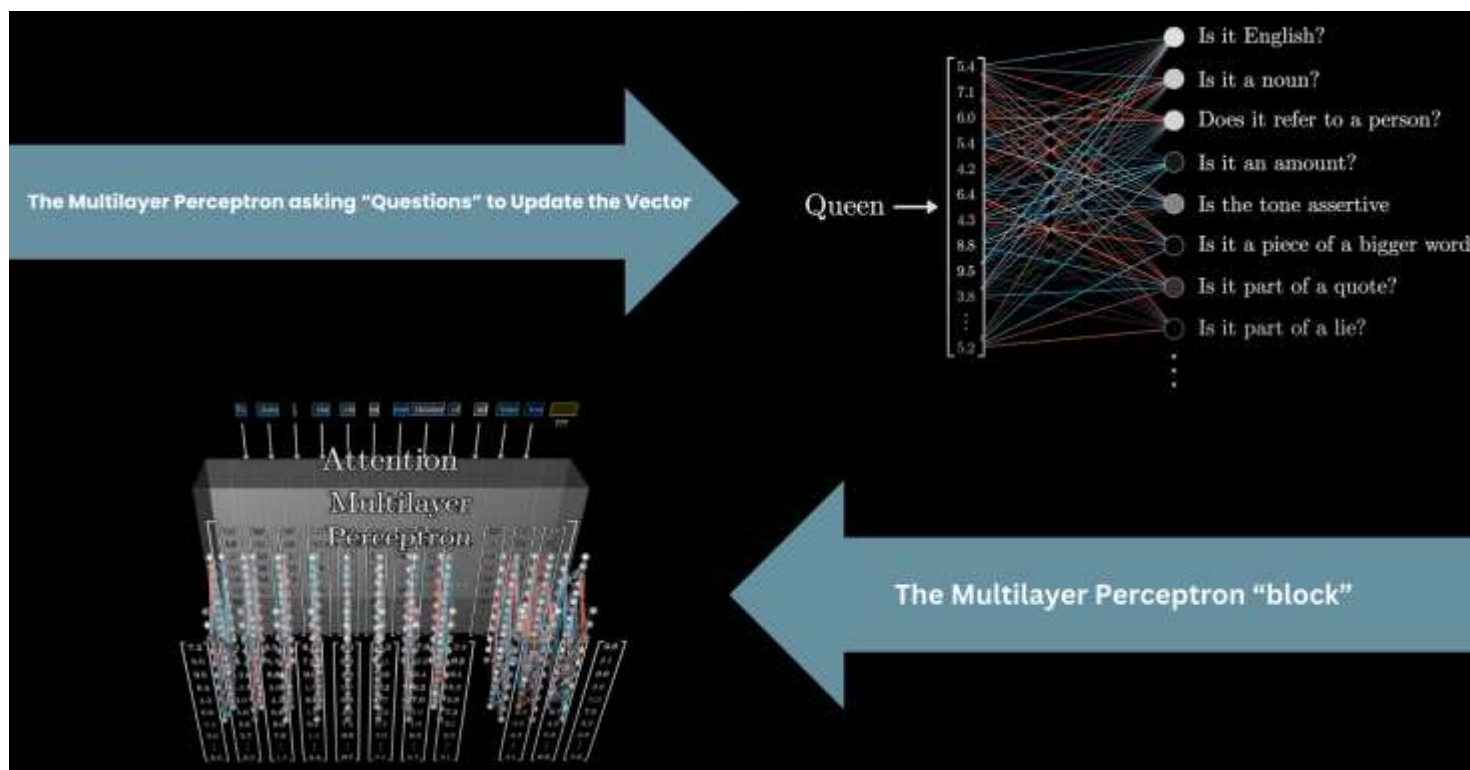
■ 多头注意力 (Multi-head Attention, MHA)

- 多个注意力矩阵，各自侧重不同方面，一起把上下文含义嵌入token向量



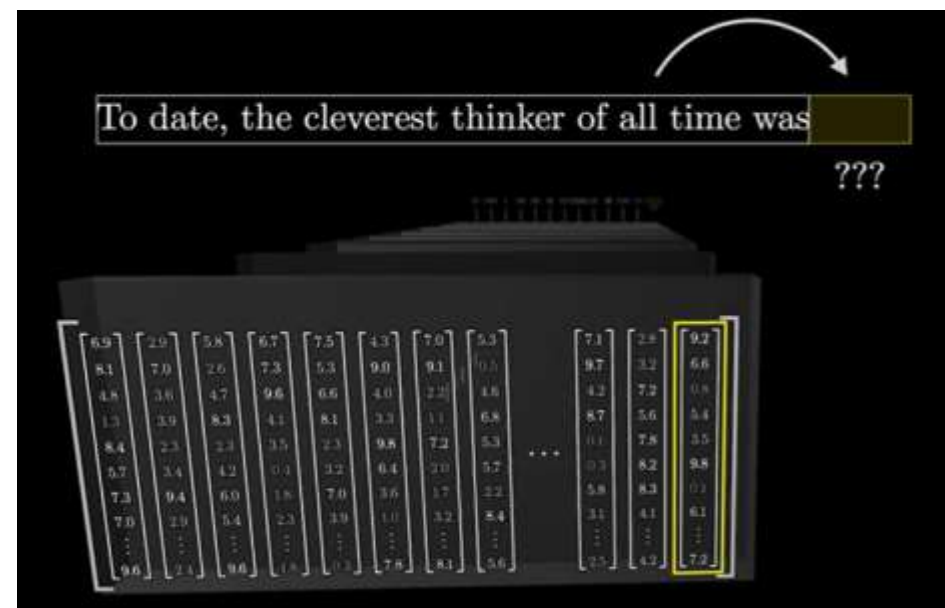
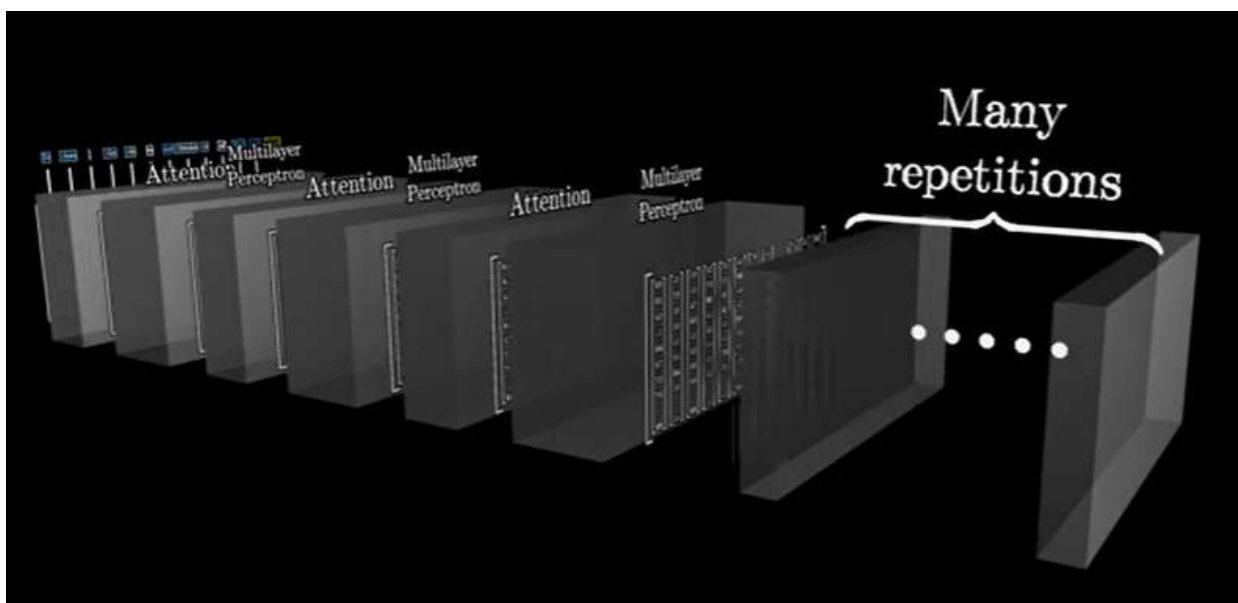
MLP(Multilayer Perceptron)

- 多层感知机：理解每个token自己的含义
 - 每个token的向量，独立的乘以自己的**特定权重矩阵**，好比在进一步理解这个token自身的含义，理解后的含义反映到乘完之后的token向量中



重复很多很多次Attention和MLP

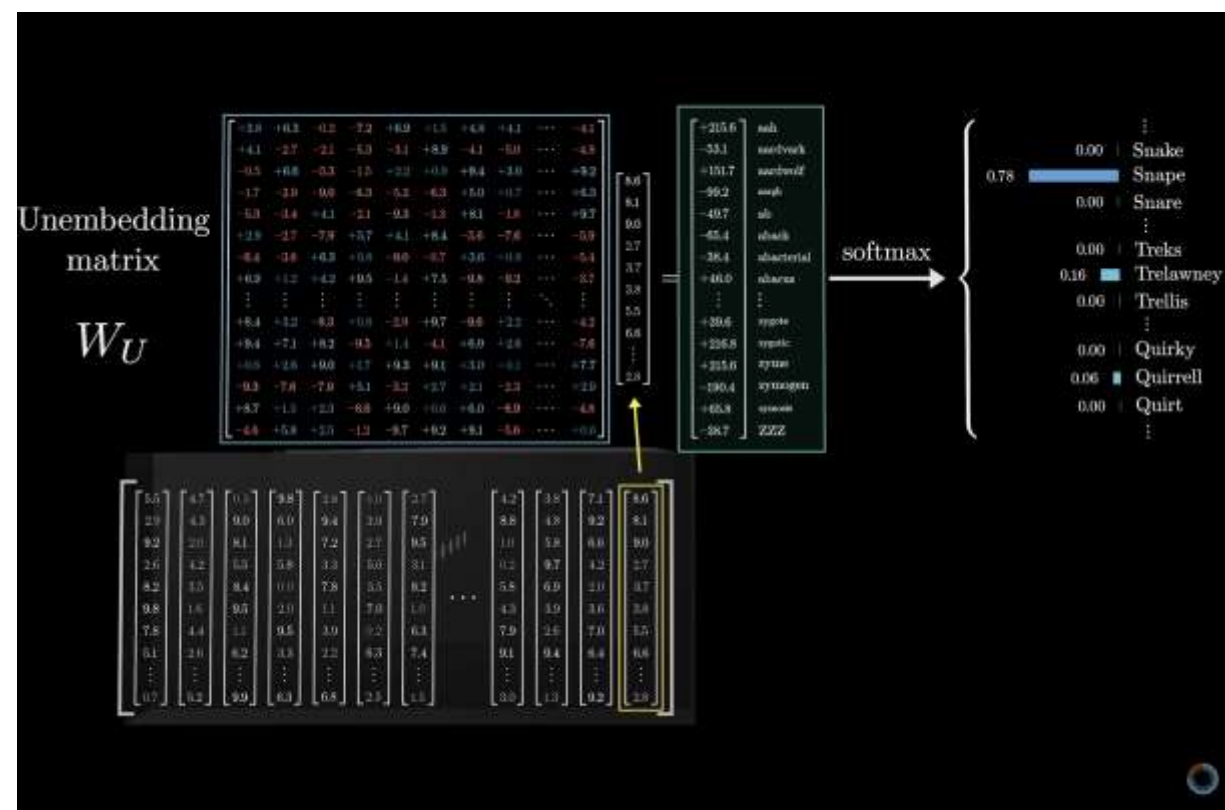
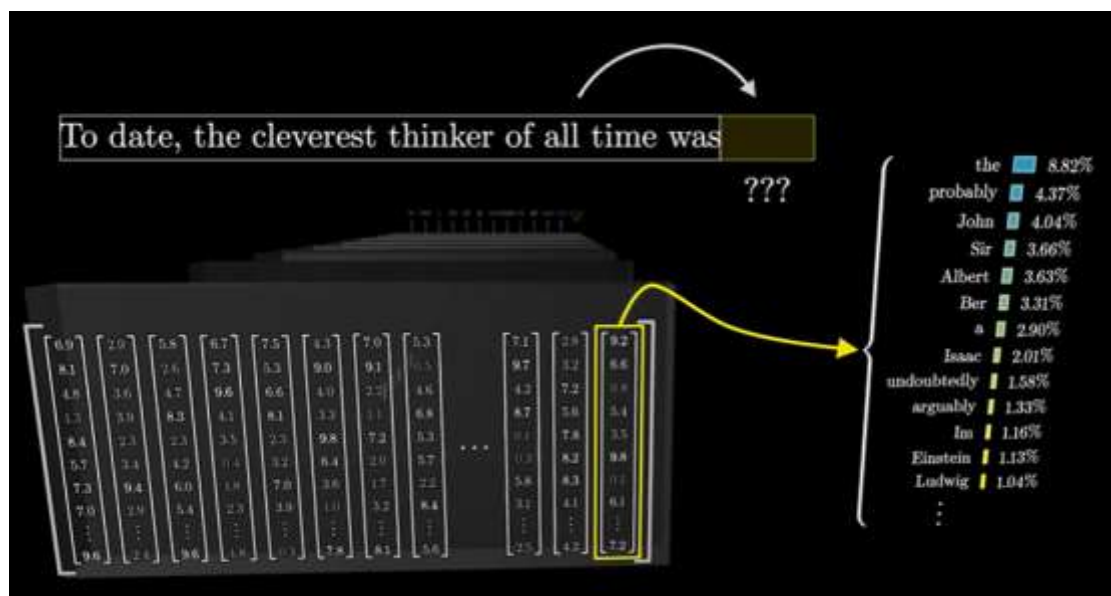
- 重复很多很多次Attention和MLP
 - 通常至少得几十次，每次都有**不同的权重矩阵**
 - 每个token之间和token自己的含义都被充分地加载到乘完的最后一个向量中



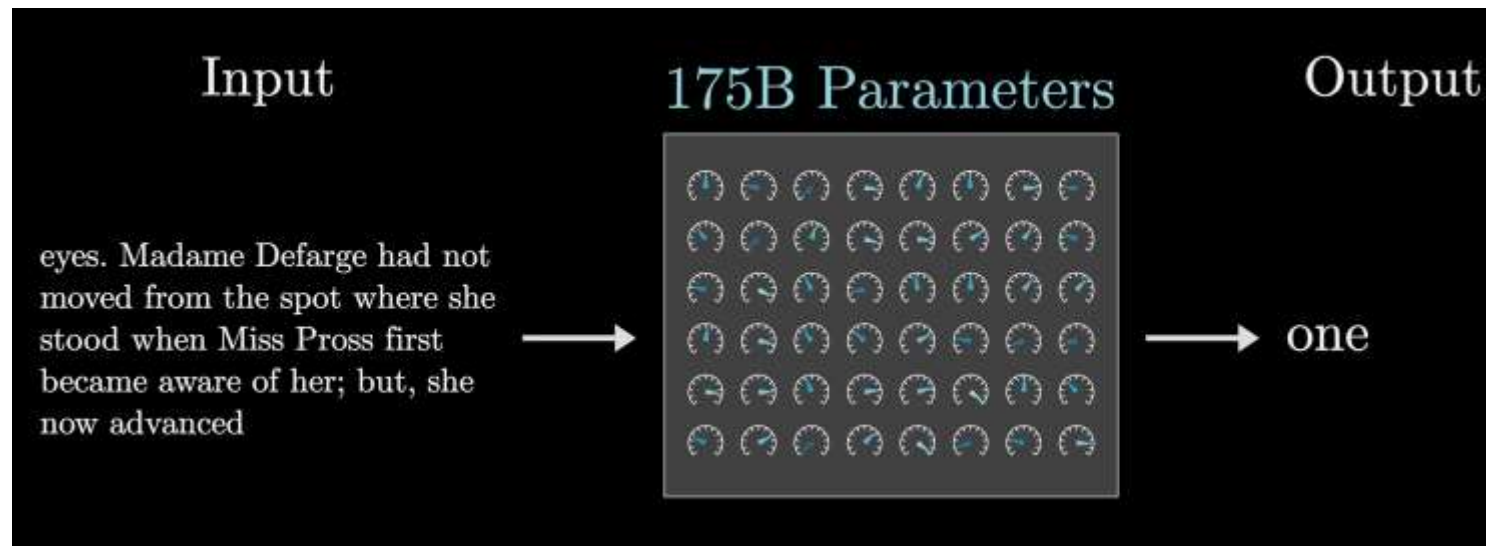
Unembedding

■ 向量编码还原成token并输出

- 乘以特定**权重矩阵**，形成输出词汇概率，通过特定函数采样输出
- 为何只用最后一层的最后一个token向量做输出？因为计算效率高

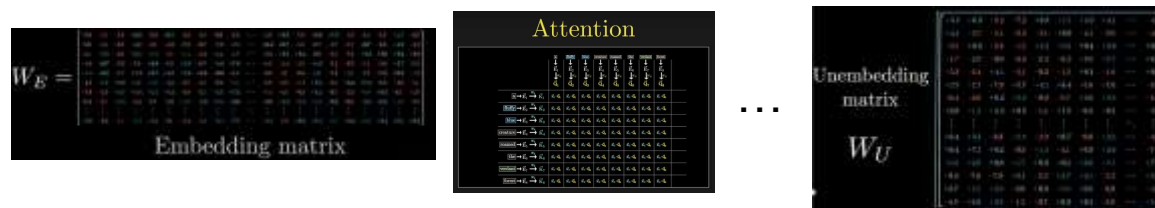


训练：LLM通过数据学习文字接龙的过程



用大量的基础数据
预训练(Pre-trained)

上千亿个参数（不断调整）



Transformer各个环节的权重矩阵里面的值。。。

湖南大→学

✓ 调好了，
就用它！

湖南大→车

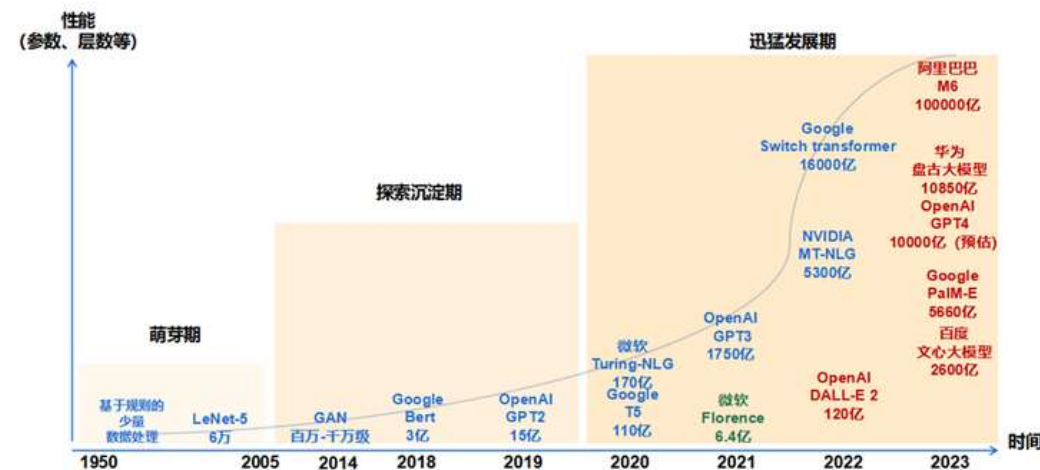
✗ 没调好，
再调！

不断生成内容(Generative)

数据和参数的规模越大，LLM就越聪明



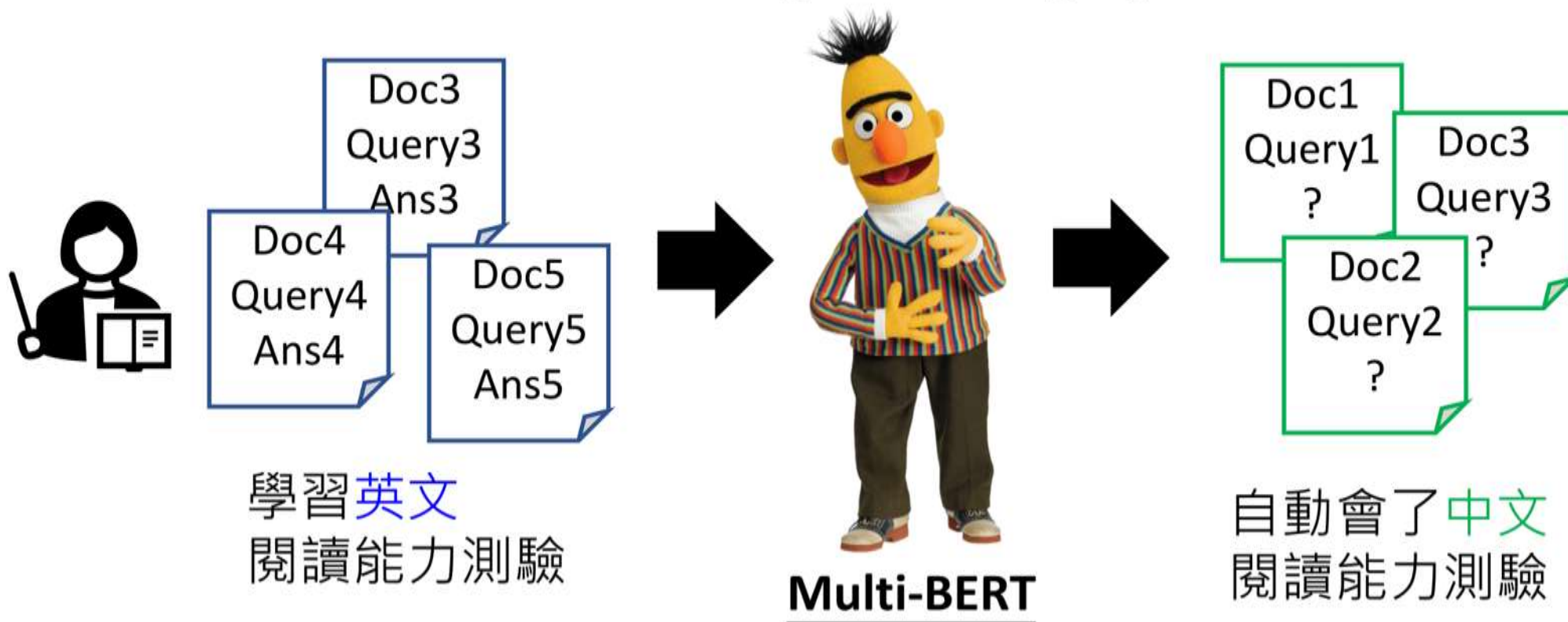
DeepSeek R1/V3
671B



大到一定规模，开始“涌现”！

在多種語言上做預訓練後，只要教某一個語言的某一個任務，自動學會其他語言的同樣任務

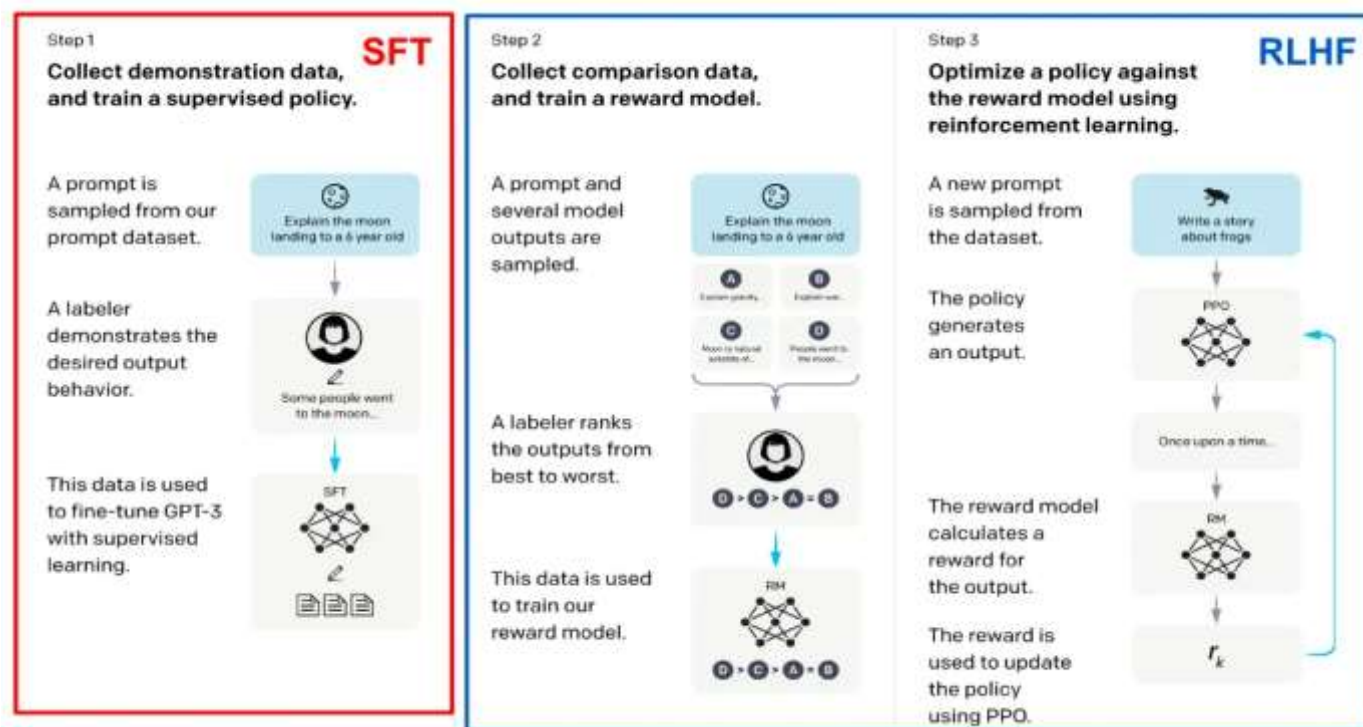
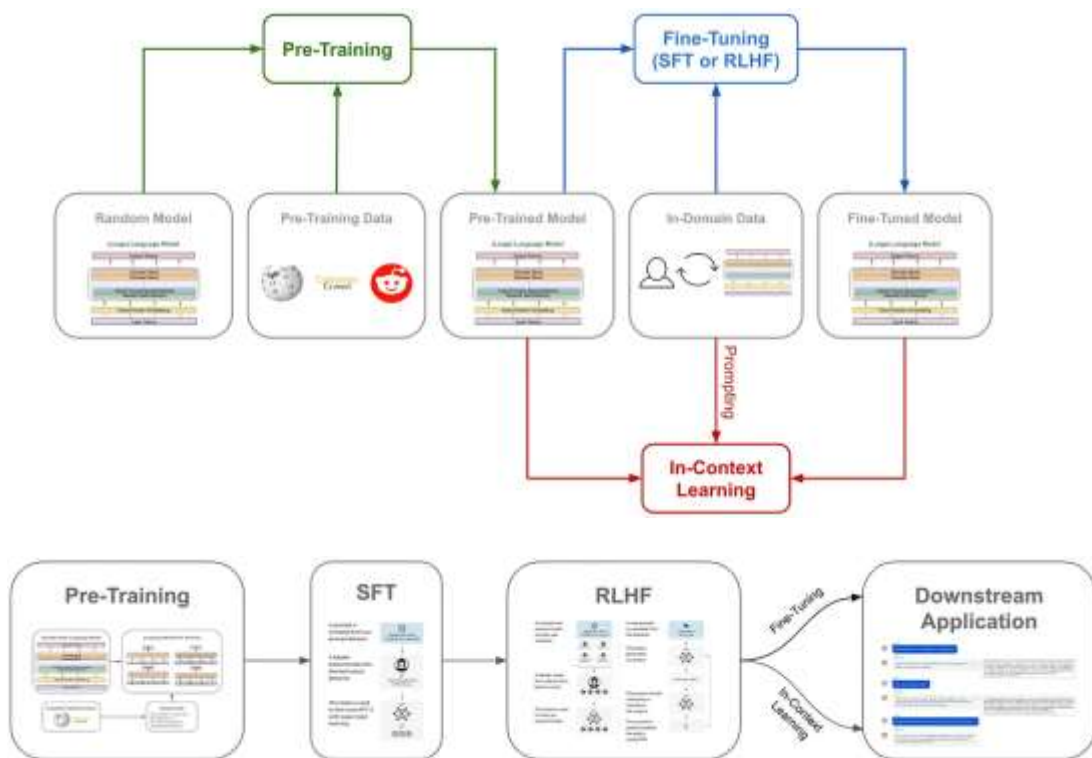
Pre-training on 104 languages



训练完发现针对特定领域不得劲怎么办？

■ 后训练对齐

- 将预训练后的模型进一步对齐数据，防止“胡说八道”
- 通常有SFT（监督式微调）和RL（强化学习，约等于给个指导自己学习）



DeepSeek模型的发展历程

幻方公司早期对AI的投入

2019

百亿量化

幻方 AI（幻方人工智能基础研究院有限公司）注册成立，致力于 AI 的算法与基础应用研究。AI 软硬件研发团队自研幻方「萤火一号」AI 集群，搭载了 500 块显卡，使用 200Gbps 高速网络互联。

High-Flyer Capital Management (Hong Kong) Limited（幻方资本）成立，获得香港九号牌。

幻方量化跻身百亿私募。

国内拥有超过1万枚GPU的企业不超过5家。而除几家头部大厂外，还包括一家名为幻方的量化基金公司。通常认为，**1万枚英伟达A100芯片**是做自训大模型的算力门槛。

2021

「萤火二号」与公益年

幻方 AI 投入十亿建设「萤火二号」。「萤火二号」一期确立以任务级分时调度共享 AI 算力的技术方案，从软硬件两方面共同发力：高性能加速卡、节点间 200Gbps 高速网络互联、自研分布式并行文件系统（3FS）、网络拓扑通讯方案（hfreduce）、算子库（hfai.nn），高易用性应用层等，将「萤火二号」的性能发挥至极限。

幻方量化成为宁波市证券投资基金业协会理事单位。

历经约一年半时间的运行，「萤火一号」光荣谢幕。

幻方公益工作小组成立，以专业化可持续的公益方式回馈社会。

2019年，幻方量化成立AI公司，其自研的深度学习训练平台「萤火一号」总投资近**2亿元**，搭载了1100块GPU；两年后，「萤火二号」的投入增加到**10亿元**，搭载了约1万张英伟达A100显卡

DeepSeek大模型之路

- 2023年7月：DeepSeek 公司成立
 - 致力于AGI
- 2023年11月：开源 DeepSeekLLM 7B 和 67B 的 Base 和 Chat 模型

| Params | n_{layers} | d_{model} | n_{heads} | $n_{\text{kv_heads}}$ | Context Length | Sequence Batch Size | Learning Rate | Tokens |
|--------|---------------------|--------------------|--------------------|------------------------|----------------|---------------------|---------------|--------|
| 7B | 30 | 4096 | 32 | 32 | 4096 | 2304 | 4.2e-4 | 2.0T |
| 67B | 95 | 8192 | 64 | 8 | 4096 | 4608 | 3.2e-4 | 2.0T |

Table 2 | Detailed specs of DeepSeek LLM family of models. We choose the hyper-parameters based on our findings in Section 3





The micro design of DeepSeek LLM largely follows the design of LLaMA (Touvron et al., 2023a,b), adopting a Pre-Norm structure with RMSNorm (Zhang and Sennrich, 2019) function and using SwiGLU (Shazeer, 2020) as the activation function for the Feed-Forward Network (FFN), with an intermediate layer dimension of $\frac{8}{3}d_{\text{model}}$. It also incorporates Rotary Embedding (Su et al., 2024) for positional encoding. To optimize inference cost, the 67B model uses Grouped-Query Attention (GQA) (Ainslie et al., 2023) instead of the traditional Multi-Head Attention (MHA).

However, in terms of macro design, DeepSeek LLM differs slightly. Specifically, DeepSeek LLM 7B is a 30-layer network, while DeepSeek LLM 67B has 95 layers. These layer adjustments, while maintaining parameter consistency with other open-source models, also facilitate model pipeline partitioning to optimize training and inference.

初期处于跟随LLaMA的状态
(一点微创新)

We release the DeepSeek LLM 7B/67B, including both base and chat models, covering a more diverse range of research within both academic and commercial communities. We release intermediate checkpoints of the base model from its training process. Please subject to the terms outlined in [License section](#). Commercial usage is per

Huggingface

| Model | Sequence Length | Download |
|-----------------------|-----------------|---|
| DeepSeek LLM 7B Base | 4096 |  HuggingFace |
| DeepSeek LLM 7B Chat | 4096 |  HuggingFace |
| DeepSeek LLM 67B Base | 4096 |  HuggingFace |
| DeepSeek LLM 67B Chat | 4096 |  HuggingFace |

上来就开源

$$\begin{aligned}
 6N_1 &= 72 n_{\text{layer}} d_{\text{model}}^2 \\
 6N_2 &= 72 n_{\text{layer}} d_{\text{model}}^2 + 6 n_{\text{vocab}} d_{\text{model}} \\
 M &= 72 n_{\text{layer}} d_{\text{model}}^2 + 12 n_{\text{layer}} d_{\text{model}} l_{\text{seq}}
 \end{aligned}$$

| Approach | Coeff. a where $N_{\text{opt}}(M_{\text{opt}}) \propto C^a$ | Coeff. b where $D_{\text{opt}} \propto C^b$ |
|--------------------------|---|---|
| OpenAI (OpenWebText2) | 0.73 | 0.27 |
| Chinchilla (MassiveText) | 0.49 | 0.51 |
| Ours (Early Data) | 0.450 | 0.550 |
| Ours (Current Data) | 0.524 | 0.476 |
| Ours (OpenWebText2) | 0.578 | 0.422 |

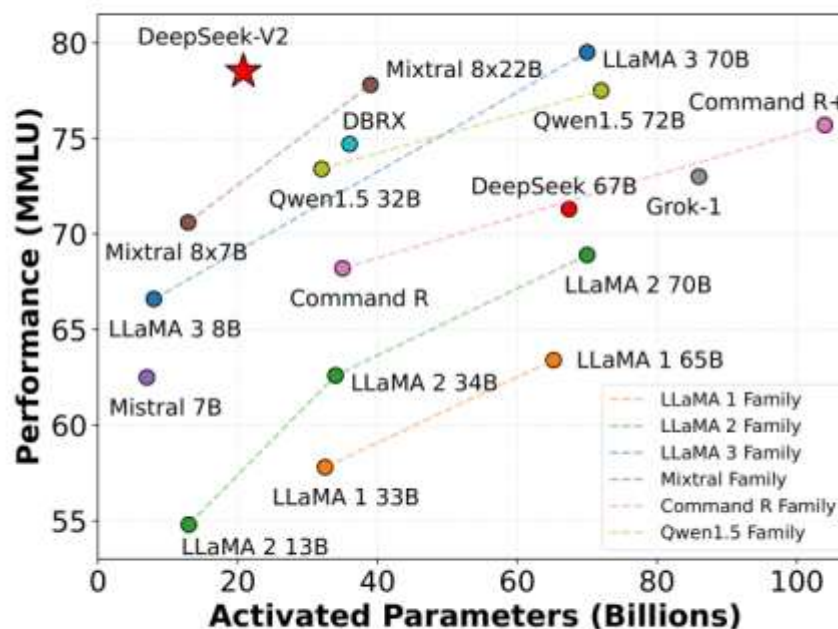
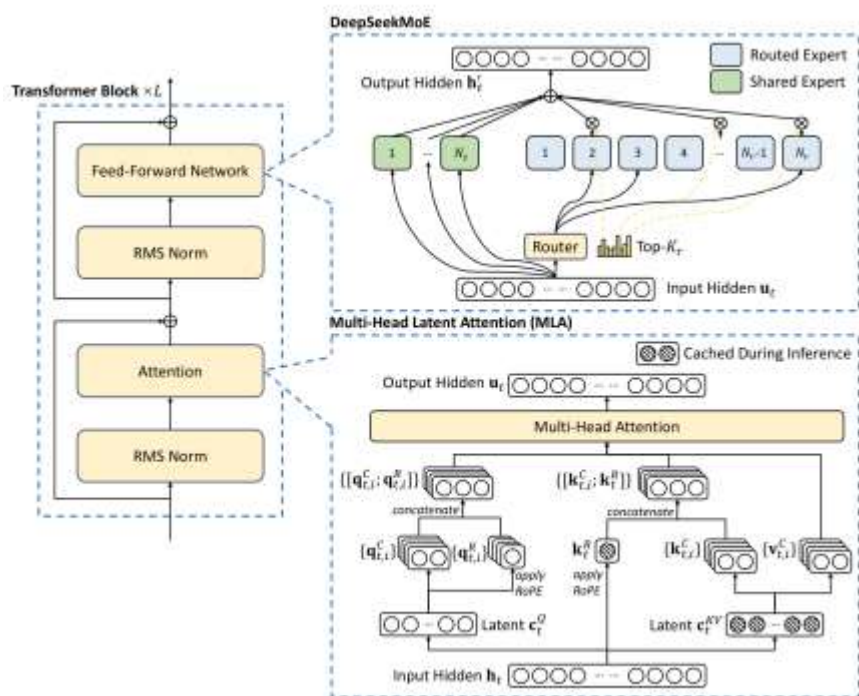
严谨地研究scaling law,
敢于质疑成名结论

初露峥嵘：开放基因，严谨思维

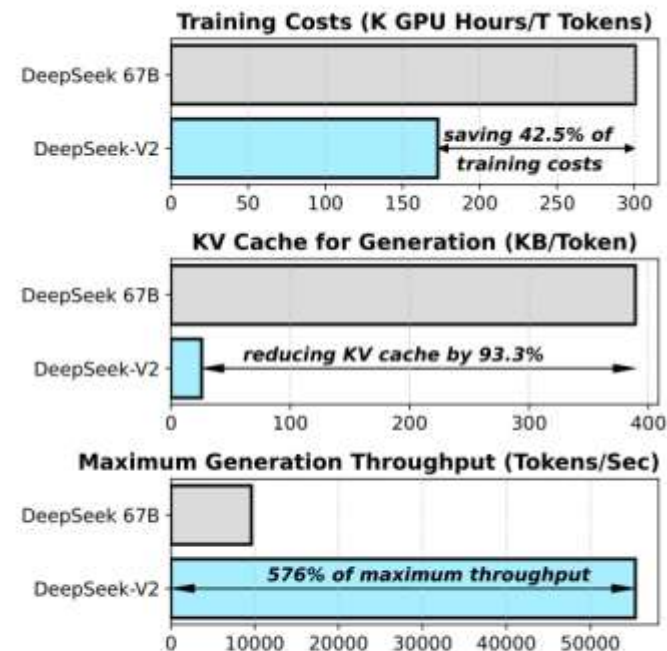
DeepSeek大模型之路

■ 2024年5月：开源 DeepSeek-V2 系列模型

□ 重要创新，效果明显，吸引圈内注意！



(a)



(b)

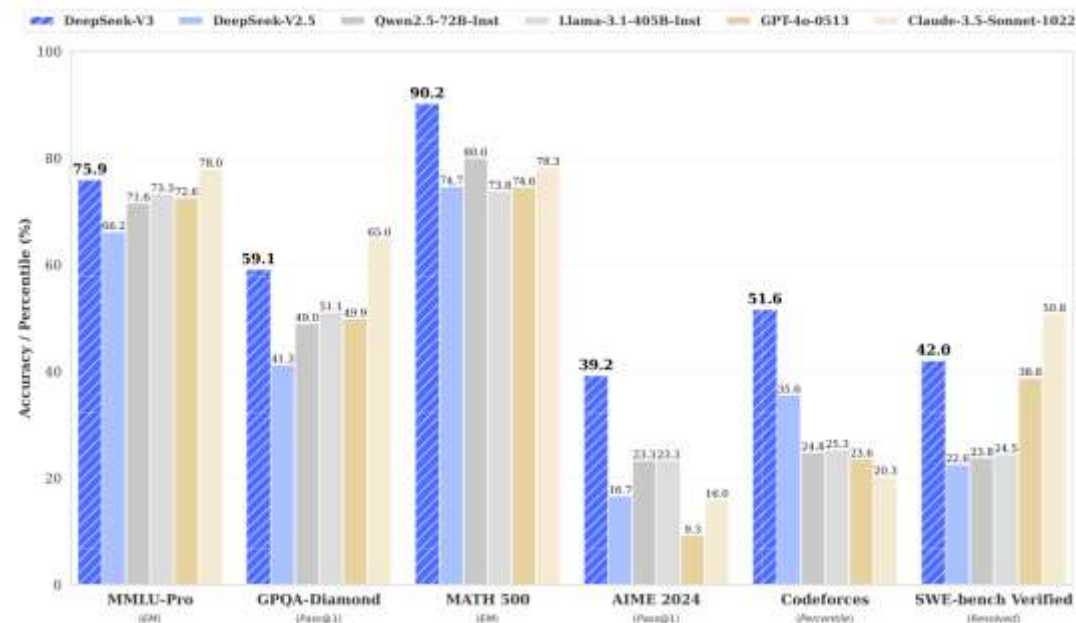
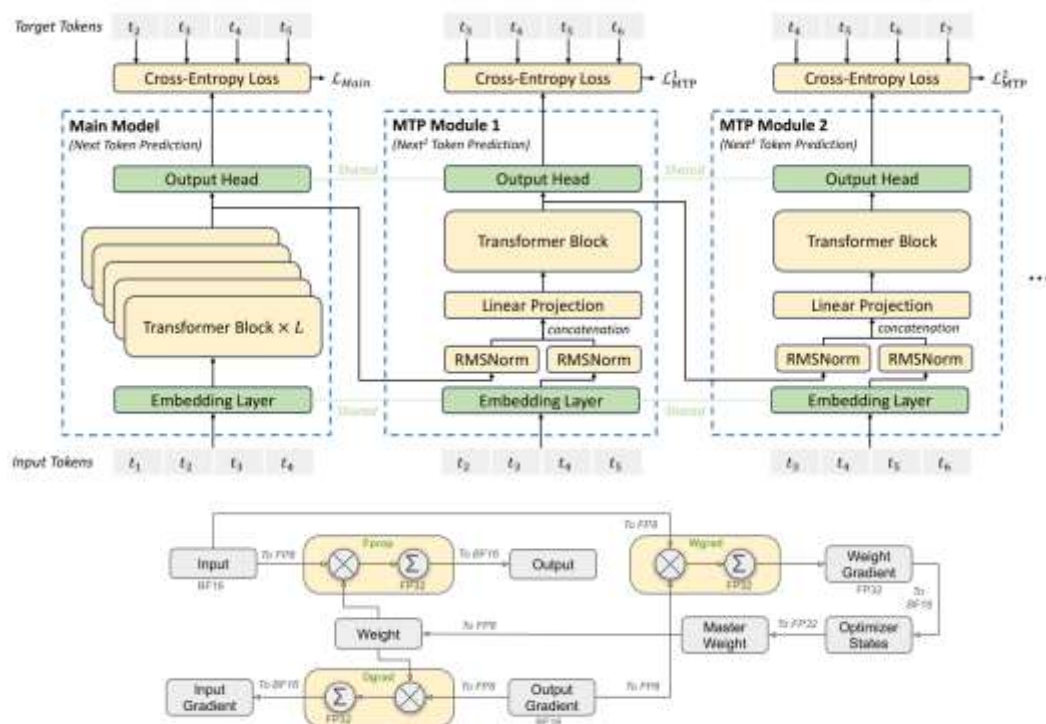
对Transformer结构大胆改造
勇于尝试大规模MoE，首创MLA

效果提升明显
较前面版本训练成本减少42%，推理所需缓存空间减少93%

DeepSeek大模型之路

■ 2024年12月26日：开源 DeepSeek-V3 系列模型

□ 基座模型SOTA!



保持大胆创新

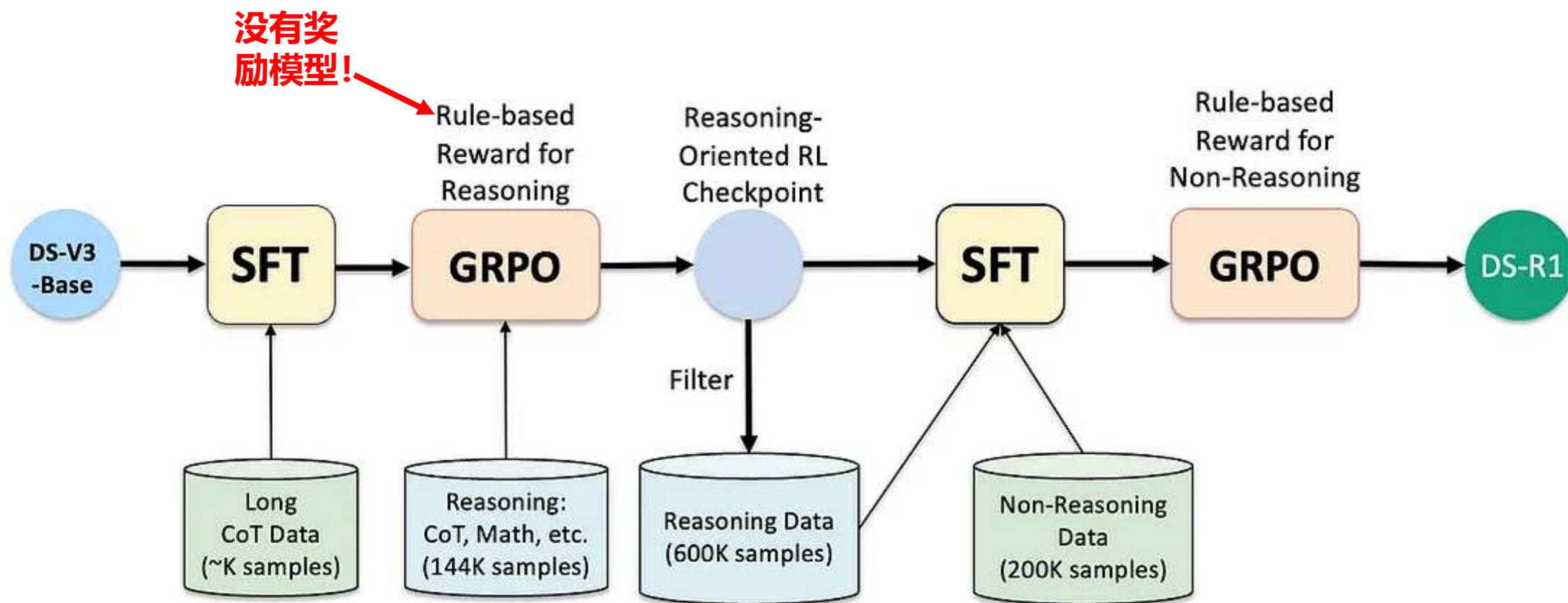
MTP, FP8训练, 继续增大MoE专家数量。。。

进入TOP梯队

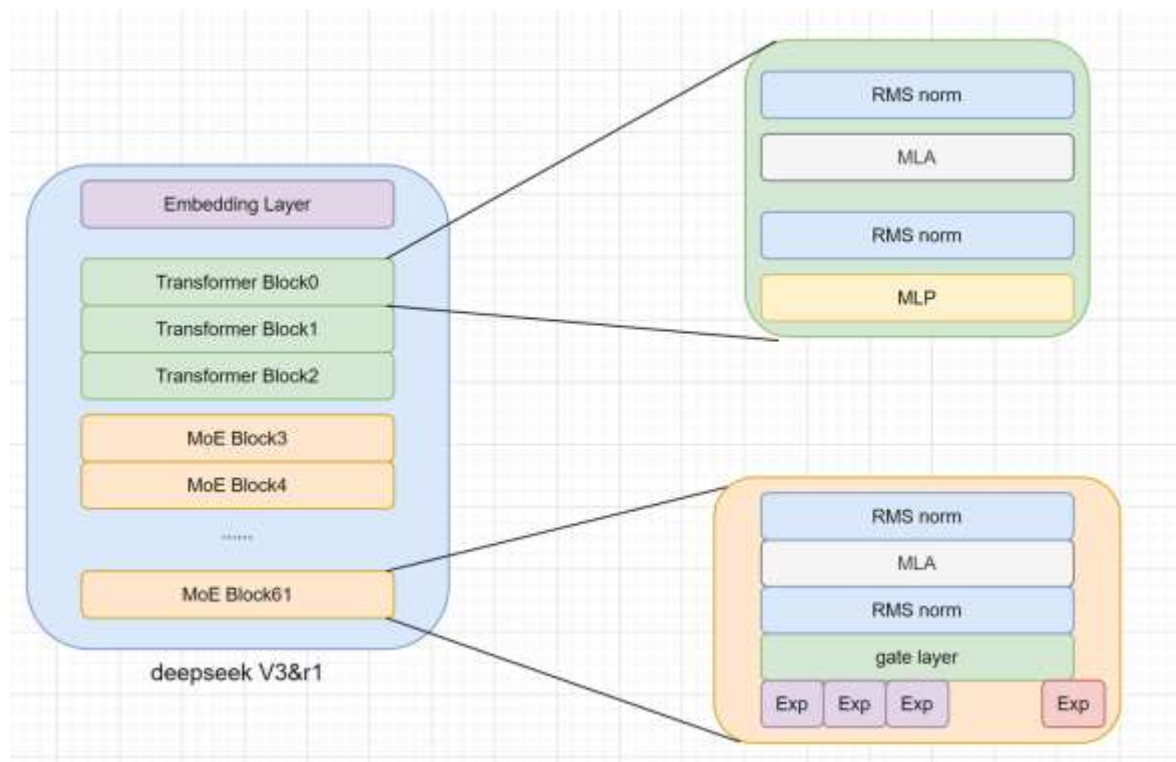
各项指标达到世界第一

DeepSeek大模型之路

- 2025年1月20日：开源推理模型 DeepSeek-R1
 - 继续创新，勇于探索OpenAI说不行的路

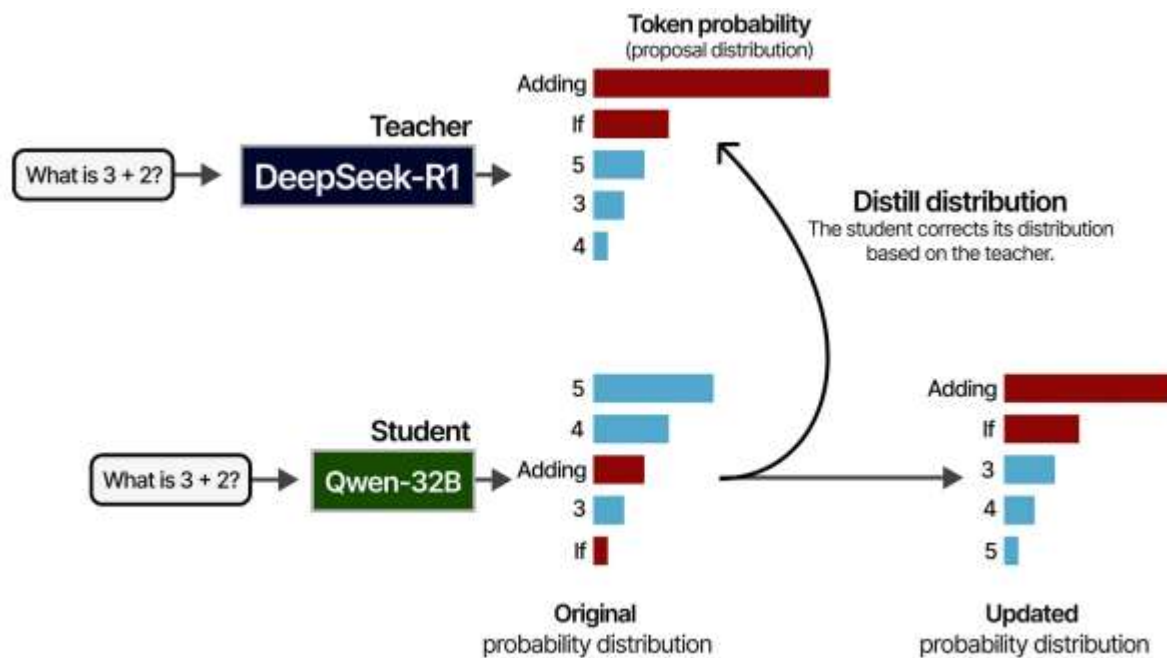


DeepSeek R1的模型结构



R1/V3 模型结构

1个Embedding, 3个普通Transformer, 59个MoE Transformer
671B (6710亿参数), 每次激活37B



开源了很多蒸馏版本

Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, Qwen2.5-14B, Qwen2.5-32B, Llama-3.1-8B, and Llama-3.3-70B-Instruct

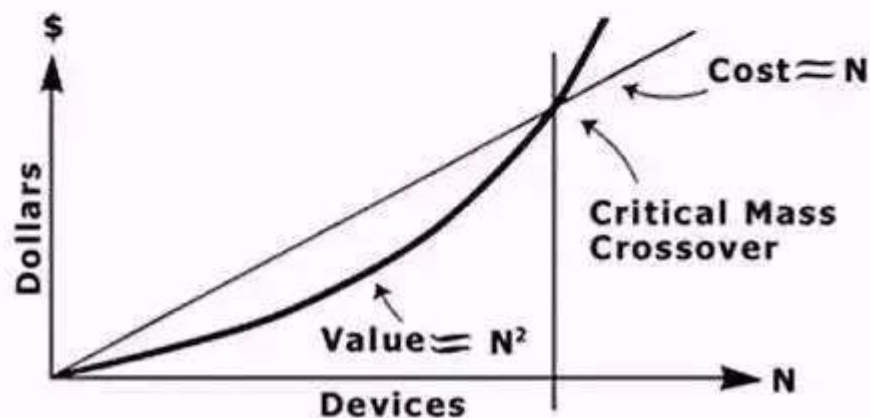
- What is it: DeepSeek是什么
 - 从ChatGPT到DeepSeek-R1, TA到底厉害在哪里?
 - DeepSeek基本概念 (用户角度)
- How to use it: 我能用DeepSeek干什么
 - 以小见大, 掌握思维方法
 - 正确理解, 打开广阔天地
- Why it works: DeepSeek背后的原理
 - Transformer——大模型基础
 - DeepSeek模型的发展历程
- **Next: 下一步要关注什么**
 - **生态的爆发就在眼前, 整个链条上哪些方面值得关注**

为什么我认为生态马上会有真正的爆发？

Metcalfe's Law:

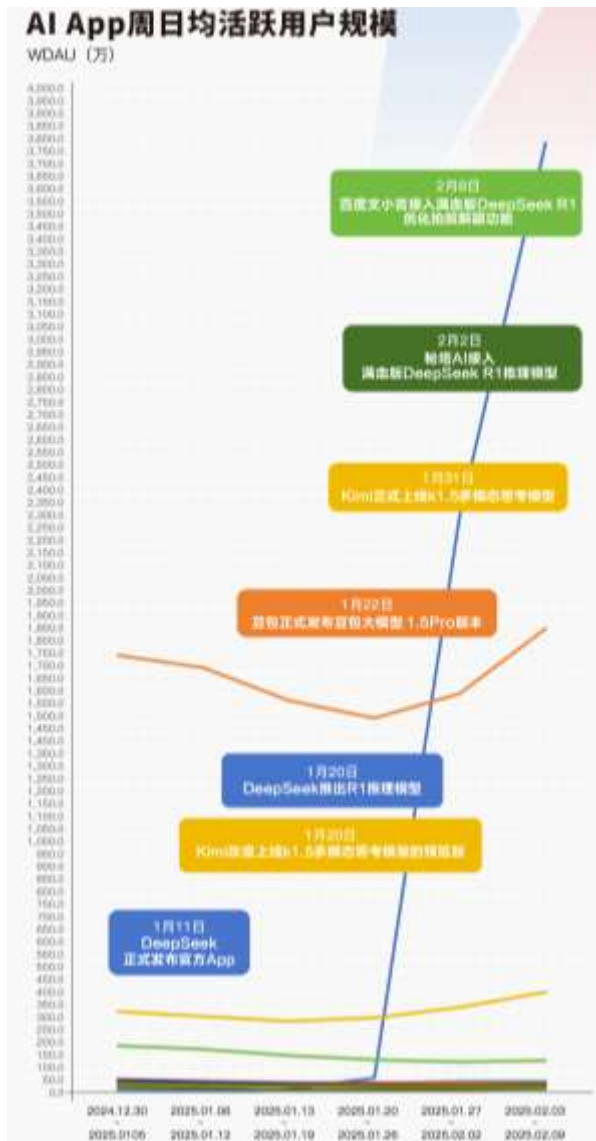
$$V=A*N^2$$

The Systemic Value of Compatibly Communicating Devices Grows as the Square of Their Number:



梅特卡夫定律

系统的价值是系统中节点数量的平方关系



<https://deepseek.hnu.edu.cn/>



DeepSeek使AI飞入寻常百姓家

量变引起质变，很可能马上（甚至已经）出现

哪些环节，我们因该关心啥？

行业应用

IT、教育、医疗、交通、城市治理。。。

公共平台

模型云服务、API接口汇聚服务、知识库服务。。。

模型算法

DeepSeek R1/V3、Kimi 1.5、Step-Video。。。

系统软件

推理引擎、训练框架、集群管理。。。

算力底座

算力中心、AI芯片（GPU）、高性能网络。。。

行业应用

公共平台

模型算法

系统软件

算力底座

■ 国产AI芯片（或称为GPU）

- DeepSeek让不少国产AI芯片公司焕发新生
 - 开源：我们都能安装；高效：我们能力弱一点也能上
 - 市场火爆：甲方嘎嘎需要，我们供不应求
- 但是不足还很明显，值得重点关注的至少有
 - 大规模训练是短板，还是无法撼动NV，**非常痛！**
 - 训练是创新算法和模型（至少现在还是）的源头
 - 推理效率还比较低，**比较痛**
 - FP8、显存等等，现在有一点点杀鸡用牛刀的意思

■ 高性能互连

- 多机多卡互连协同
 - 训练必须要；大模型推理也得要（比如R1 671B）
- 目前NV一家独大，国内不知何时能有挑战者，**非常痛！**
 - 主要是机间RDMA网络、机内GPU网络等
 - 国内技术研究进入前沿（如我组就做这个），产品还任重道远

行业应用

公共平台

模型算法

系统软件

算力底座

■ 国内状态还比较乐观

□ 普遍有开源软件

➤ PyTorch、vLLM、K8S。。。

□ 国内实力还不错，人才梯队也有

➤ 很多企业都有参与开源或自己研制训练框架、推理引擎等

■ 技术更新非常快！

□ 对中小企业等本地部署的玩家提出高要求

➤ 目前主流推理引擎的更新发版速度以天记

➤ 不求研发进去，至少要能看得懂、跟得上、用得会

□ 要大力培养这方面的人才

➤ DeepSeek的成功很大程度得益于这部分人

➤ 我省现有这方面的高端人才，想办法聚合

• 如HNU DeepSeek服务技术支撑小组

个人浅见，仅供参考

行业应用

公共平台

模型算法

系统软件

算力底座

■ 现状比较乐观

- DeepSeek等已经证明，我们已处于第一梯队

■ 忧患未曾远离

- NV的禁令，短期对模型算法的创新还是影响较大
- 福祸相依：DeepSeek为了规避硬件限制，降本增效，逼出了各种创新
 - MLA、NSA、MoE。。。

■ 下一步注重啥

- 开放很重要！
 - 模型开放、算法开放、训练数据开放、推理部署开放
- 用阳谋对抗阴谋，用全中国全世界的智慧一起创新

行业应用

公共平台

模型算法

系统软件

算力底座

■ 目前处于比较混战的阶段

- 技术含量相对下面三层较底，也没有绝对统一的标准、规范、形式
- 极大量的需求，都需要通过这一层接入

■ 下一步

- 这里可能成为创业的集中赛道
- 我看好知识库服务平台（点到为止，多的不能再说了☺）
 - 行业需求千变万化，归总形式主要是这个
 - 这块有一定的技术门槛和资源门槛
 - 对用户体验影响极大

行业应用

公共平台

模型算法

系统软件

算力底座

- 目前形势一片大好，但扎实落地是要务
 - 具备专业技能的人，目前已感受到LLM的巨大帮助
 - 如何能让更多普通人也感受到？
 - 功能边界、用户体验、智能体、具身智能。。。
- 实现我们下面的目标，得靠这一块！

新一代大模型



帮助大部分的普通人，
摆脱一部分中级甚至
是高级脑力劳动

欢迎进一步交流!



全国**第三家**、中西部**第一家**国家超级计算中心



中心拥有“天河”系列超级计算机、“天河·天马”计算集群等多个计算平台
通用算力达206P Flops (FP64), 人工智能专用算力达1060P Flops (FP16), 磁盘总容量达50PB。



中心建设运行“全球互联网域名根镜像节点”和“中国国家顶级域名解析节点”。
中心数据网络同时接入电信、移动、联通和教育网四大运营商, 关键流量通过BGP专线保障

谢谢!

陈果

湖南大学

邮箱: guochen@hnu.edu.cn

个人主页: <https://grzy.hnu.edu.cn/site/index/chenguo>