



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。

提纲

- 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



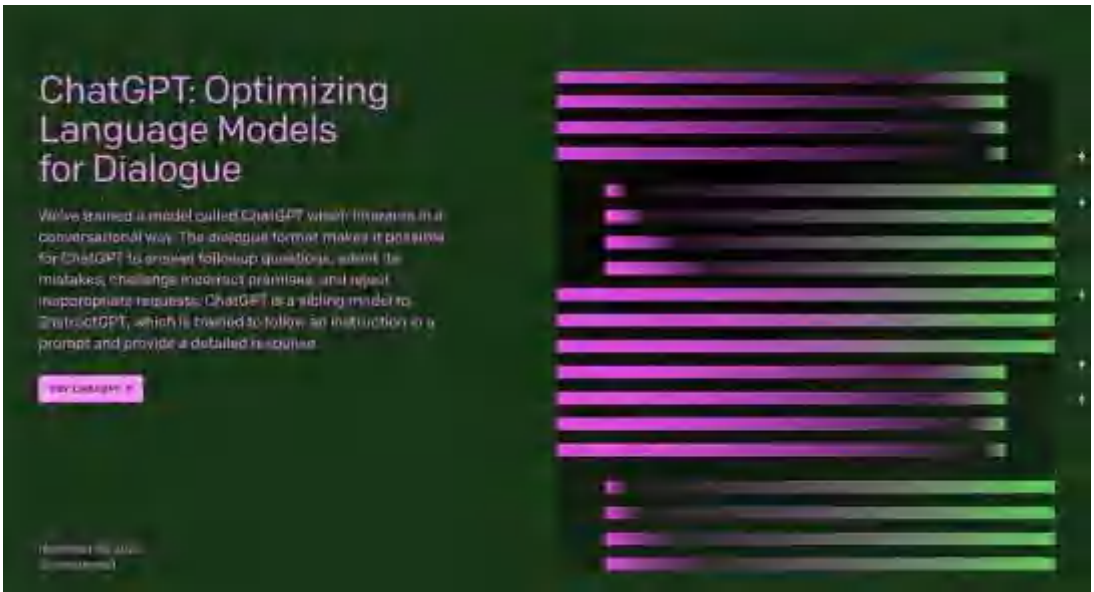
提纲

- **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

- 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



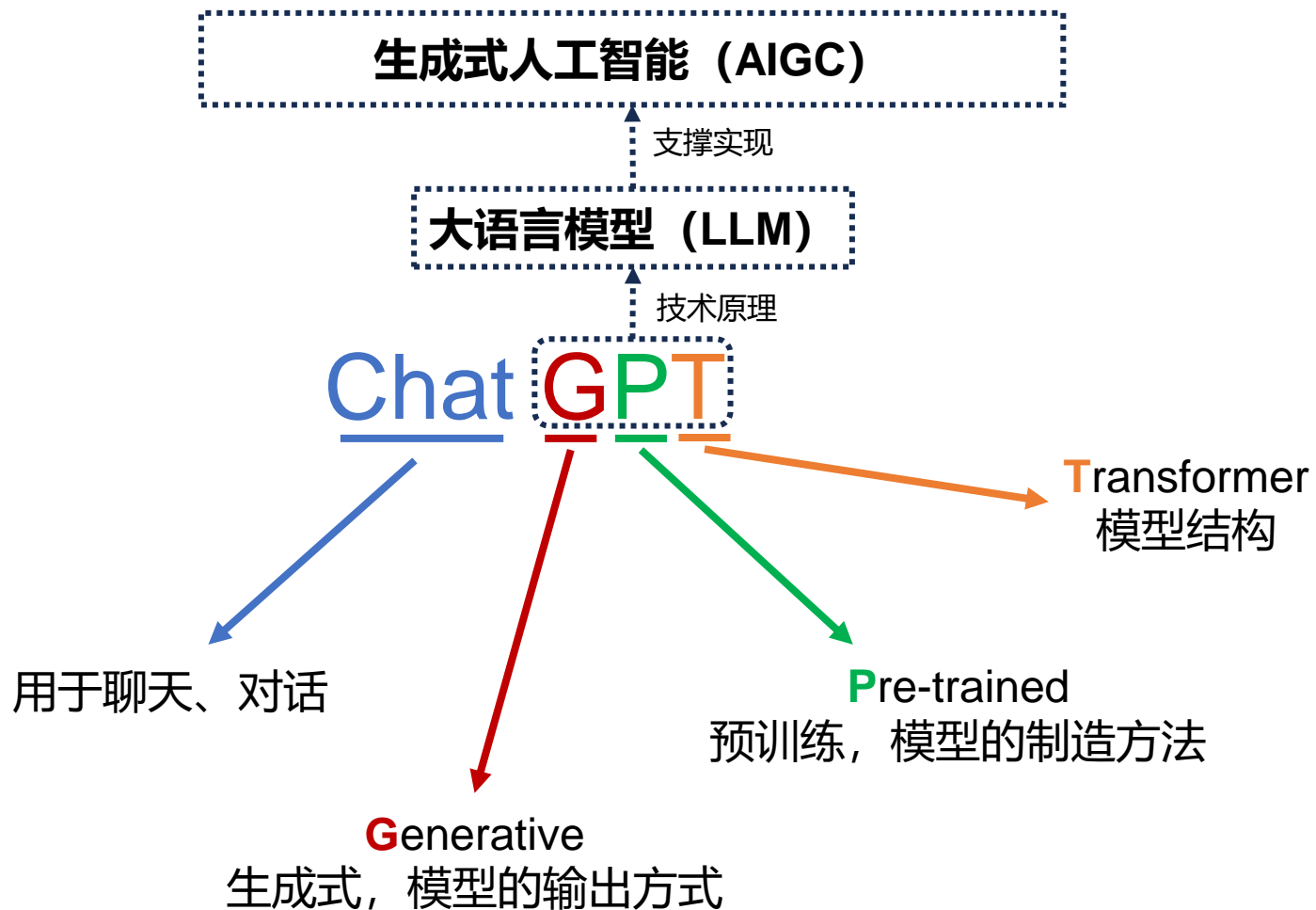
2022年11月30日
OpenAI发布对话式AI模型ChatGPT



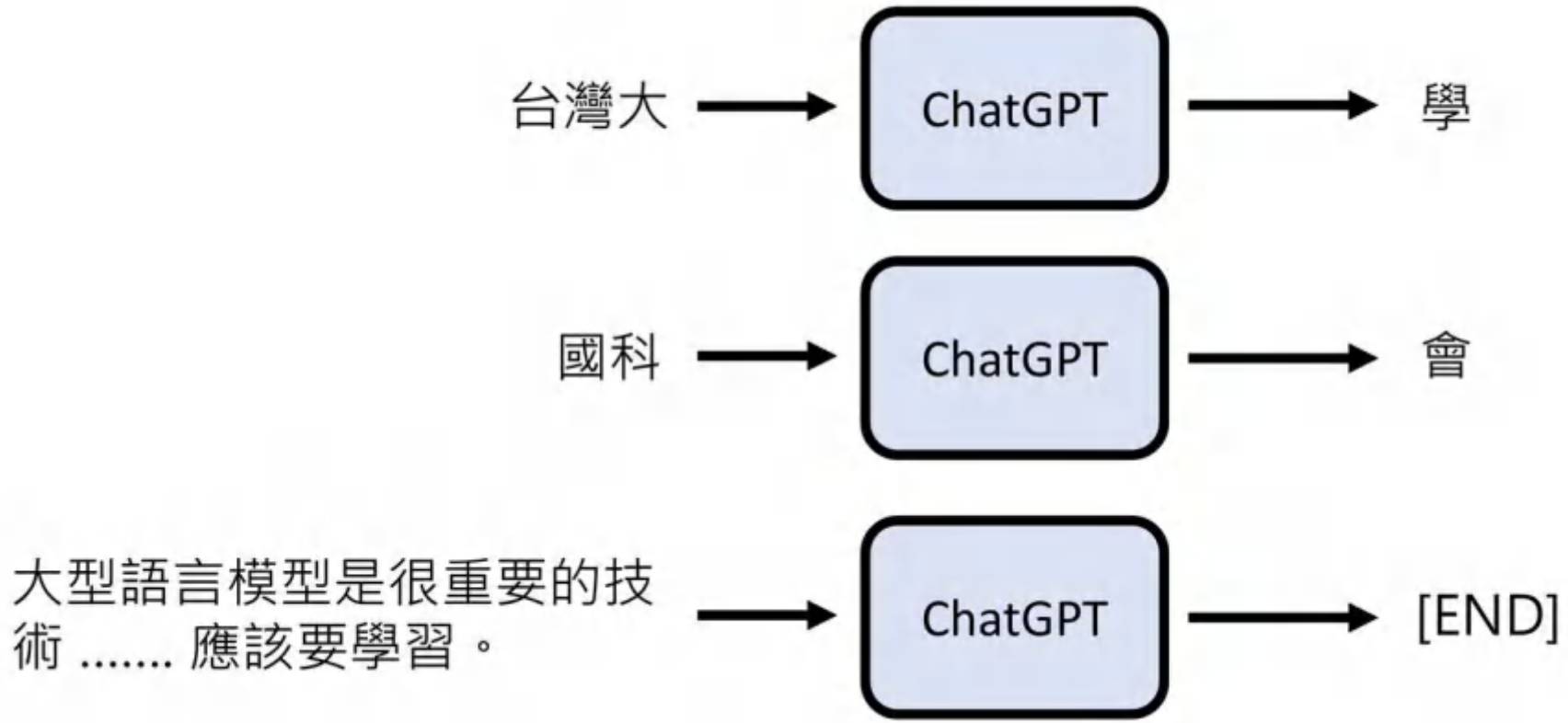
两个月
用户破亿

ChatGPT背后的方法

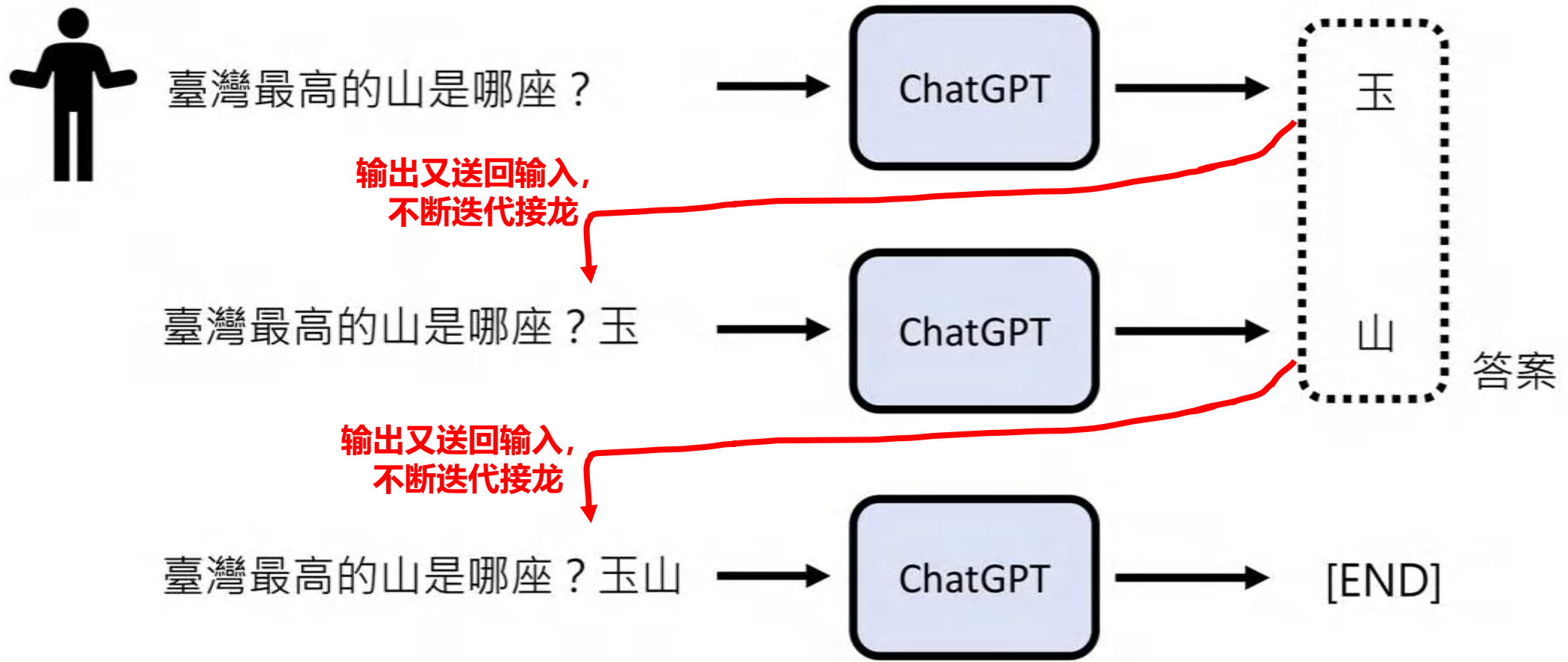
- 生成式人工智能 (AIGC) 和大语言模型 (LLM, 也简称大模型)



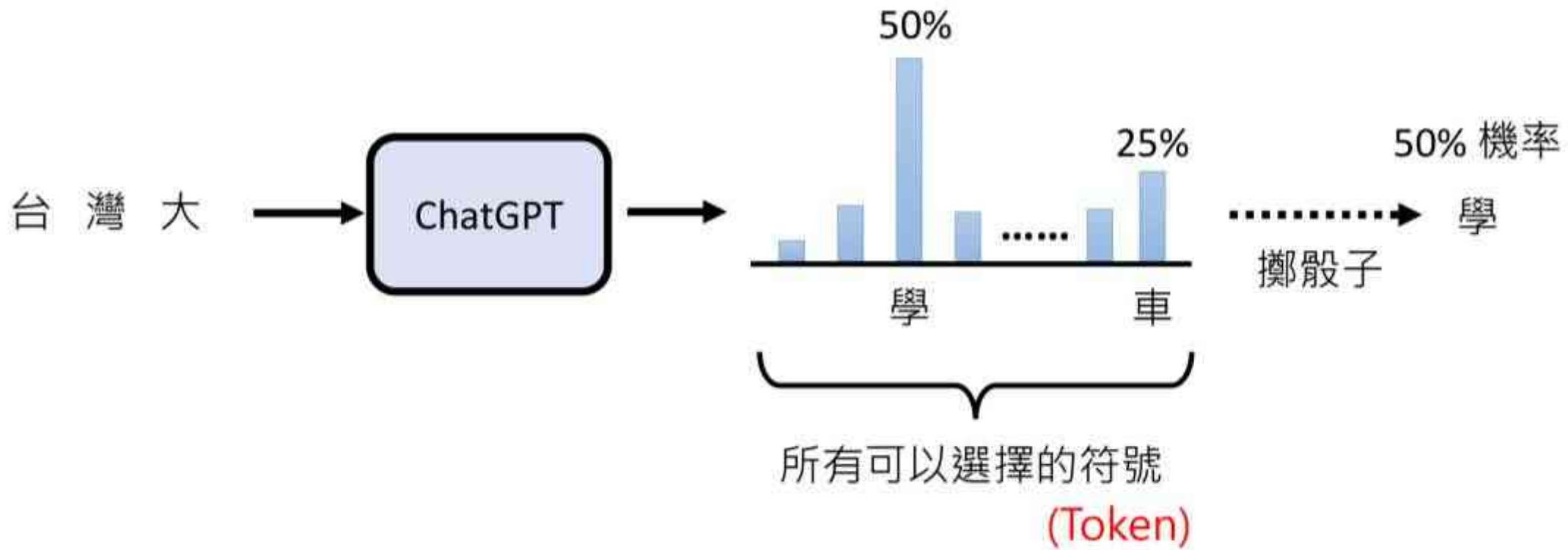
ChatGPT真正做的事：文字接龍



ChatGPT真正做的事：文字接龍



ChatGPT真正做的事：文字接龍



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/roses/>

這個網址並不存在!

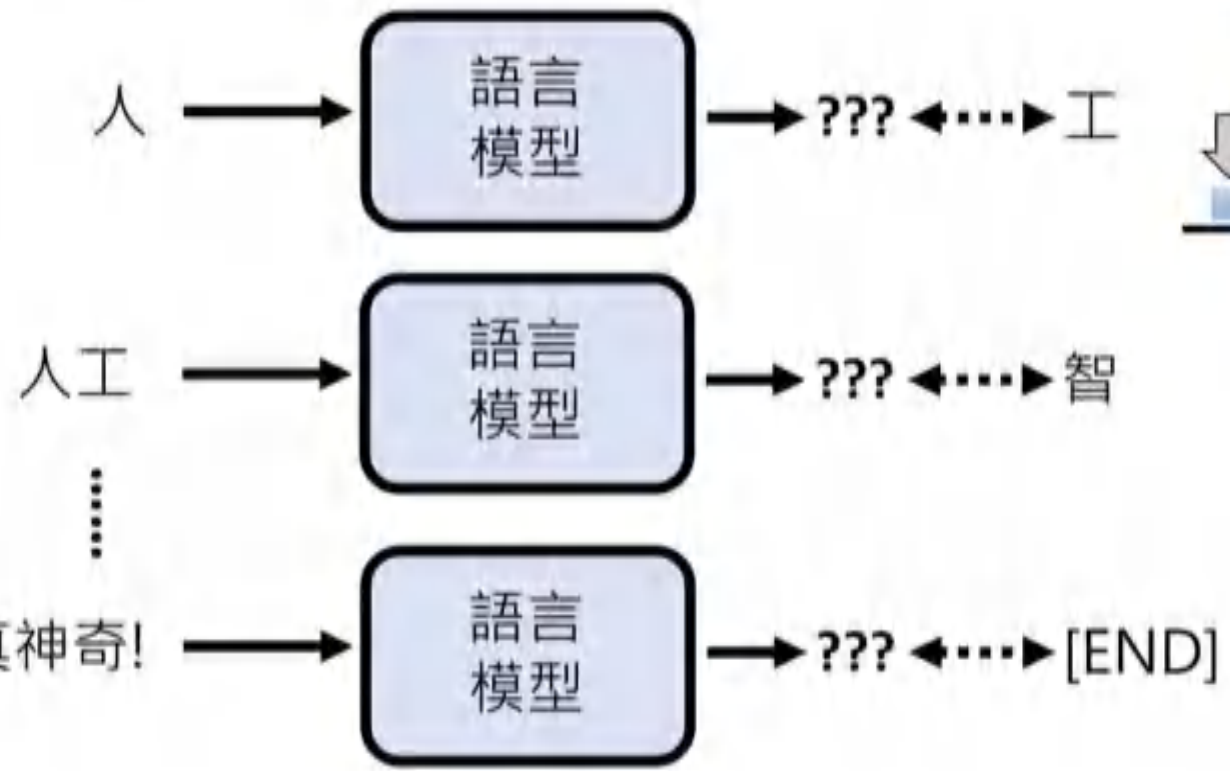
LLM怎么学习文字接龙?

任何文句都可以是教材!



網路上有無窮無盡的文句

人工智慧真神奇!

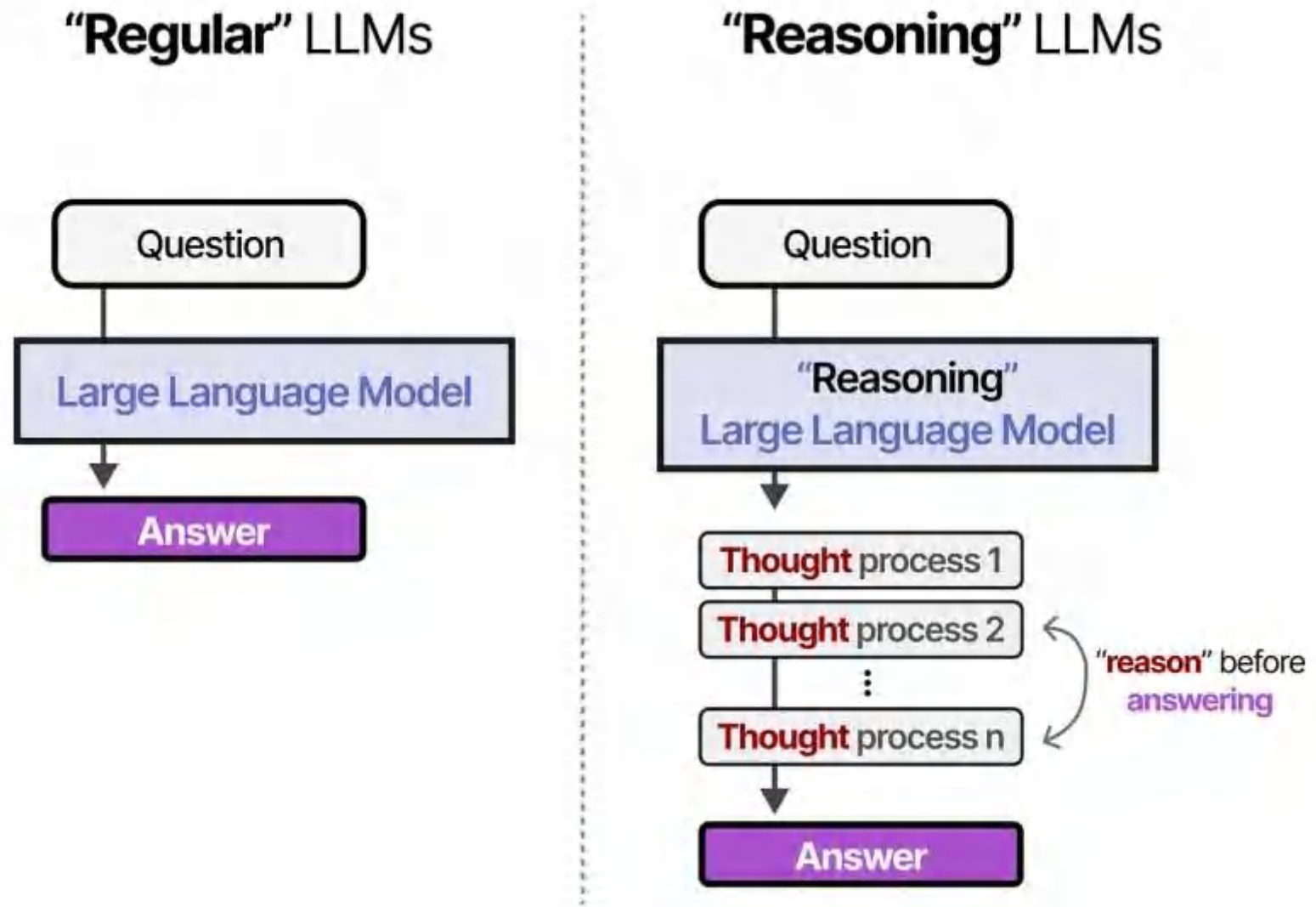


从ChatGPT到OpenAI O系列

推理大模型开始走入视野：OpenAI o1

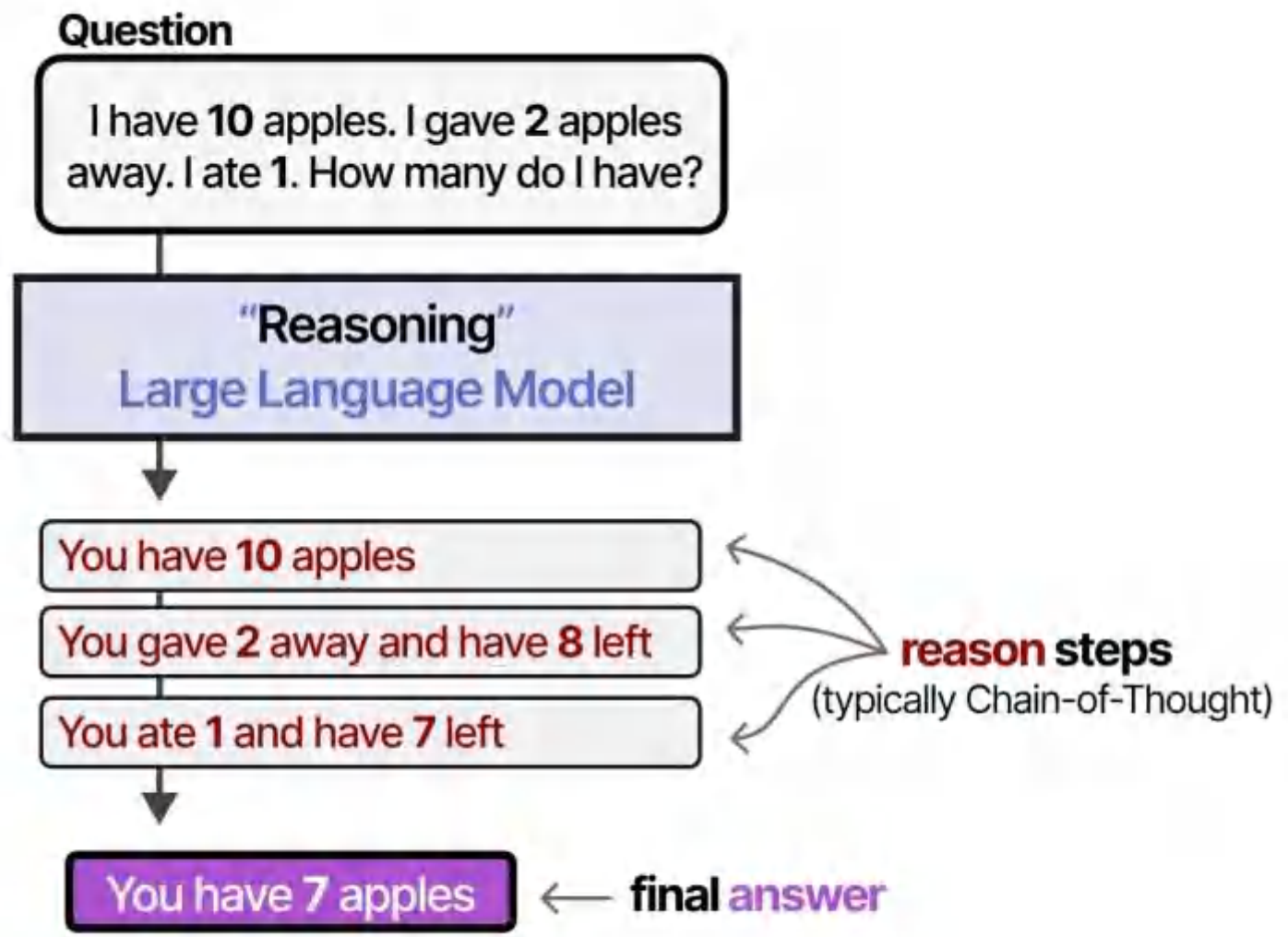
- 2024年9月12日，OpenAI官方宣布了OpenAI o1推理大模型。
- OpenAI宣称OpenAI o1大模型**推理（Reasoning）能力**相比较当前的大语言模型（GPT-4o）有了大幅提升。由于OpenAI o1模型的训练使用了一种新的AI训练方法，强调了“**思维链**”过程和强化学习的重要性，最终导致它在数学逻辑和推理方面有了大幅提升，大家开始称这里有思考过程的大模型为**推理大模型**。

什么是推理模型



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

2024年10月11日 11:17

嘿，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的省会，直线距离大约是360公里左右。所以知道直线距离后，可以预估一下时间。然后，高速可能会更快。楚天高速我记得是从长沙到武汉的。另外，highways like G0421可能也是一条路线，然后转G42沪蓉高速。这个路线可能会稍微长一点，但可能更快。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可能会比理论值低。然后，路况的好坏。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。有些驾驶员开得快，而有些则比较慢，这也会影响总时间。

其他因素还有：是否有休息的时间，比如中途停车上厕所、吃饭。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，长沙到武汉的车程时间可能在4到5小时之间。如果是自驾游或者比较赶时间，可能会更快。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信息可能会有更新。

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。



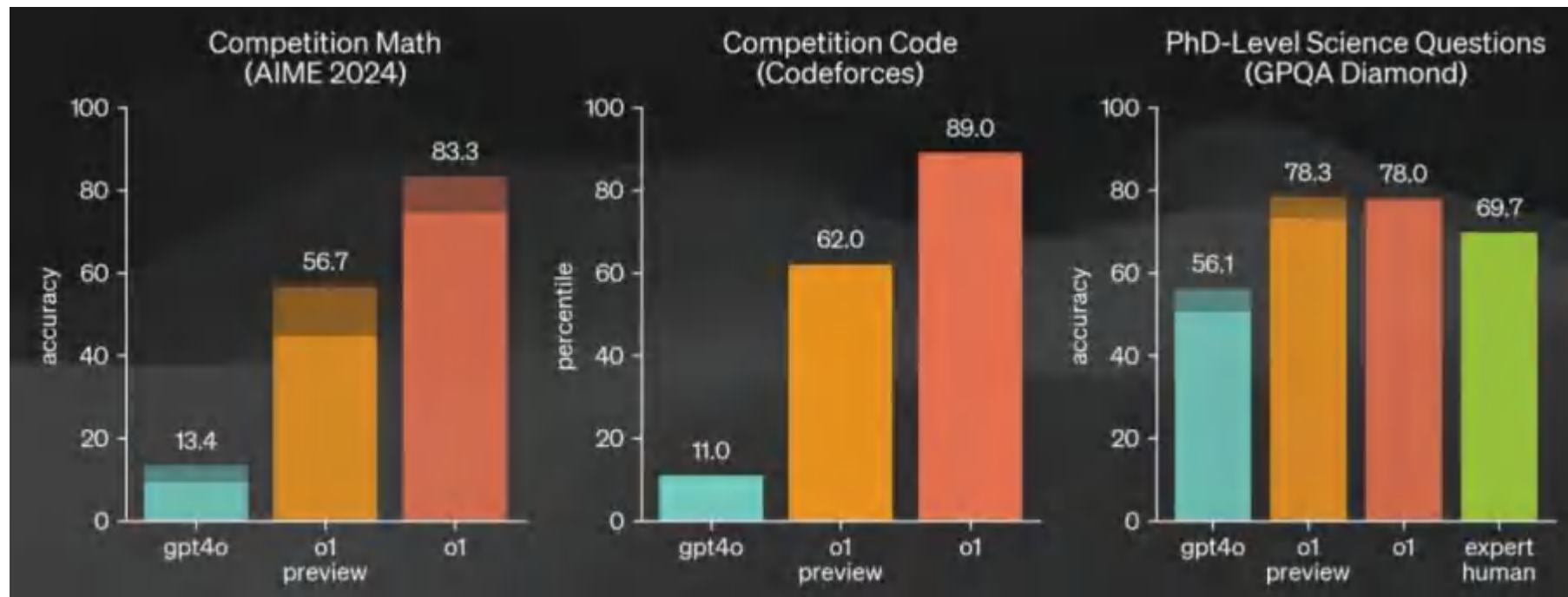
推理模型在一些领域优势明显

■ 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

韩氏家园迎岁新，氏族欢腾贺弄璋。家喜明珠辉玉牒，园添锦绣映红妆。

在诗中再加入祝贺喜得千金的意思

1. 首个展示思维链过程的推理模型



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”。现在要在其中加入喜得千金，可能需要看原诗中的意象。喜得千金通常涉及新生儿，特别是女儿，所以可以考虑在原诗的“园添锦绣映红妆”这句中加入。原诗中的“园”可能指家业，所以可以考虑在原诗的“家喜明珠辉玉牒”这句中加入。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”、“璋”、“牒”、“妆”，但可能需要改变部分内容。比如第三句末尾的“园四海”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插入。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

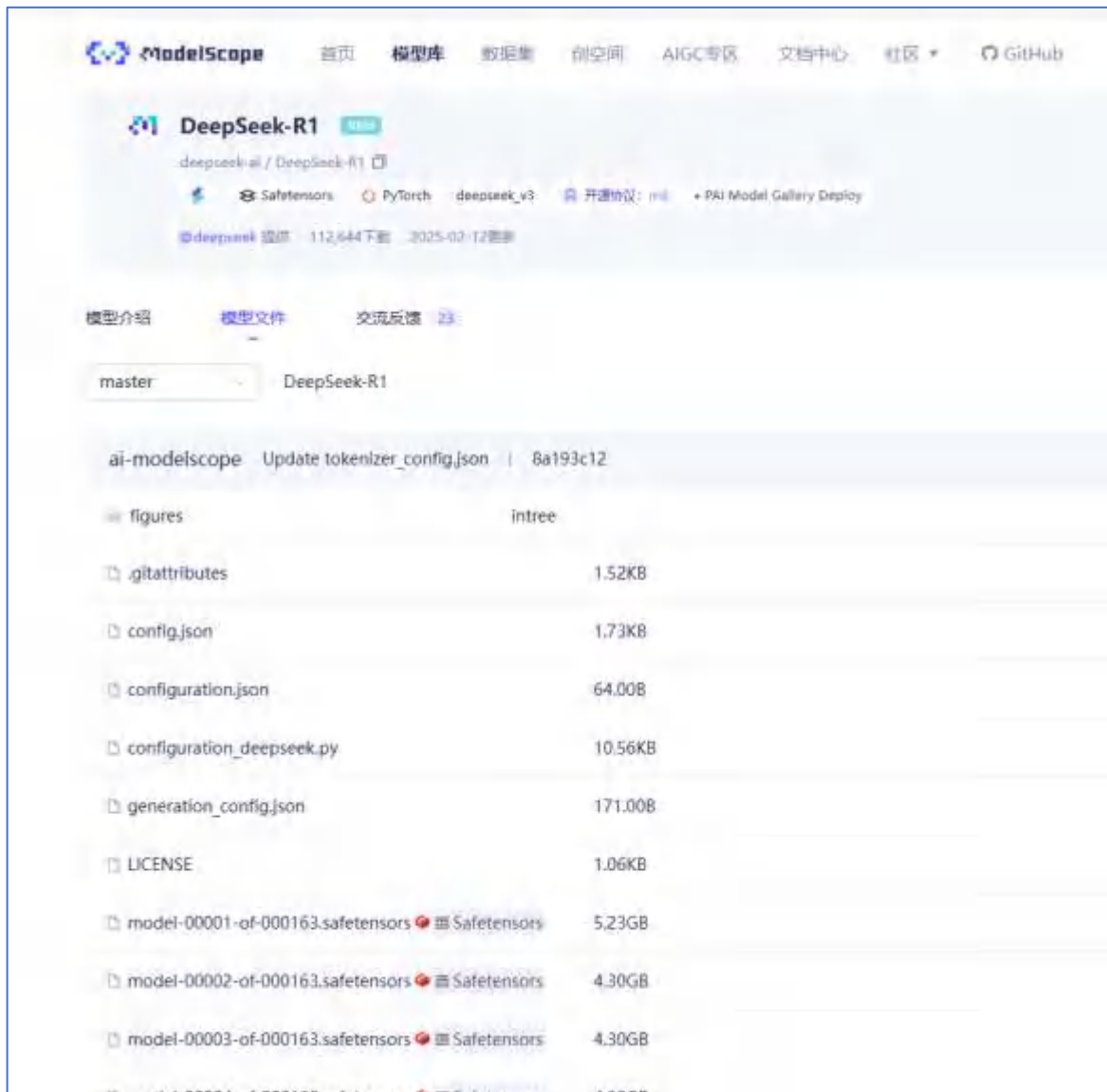
模型 ⁽¹⁾	上下文长度	最大思维链长度 ⁽²⁾	最大输出长度 ⁽³⁾	百万tokens 输入价格 (缓存命中) ⁽⁴⁾	百万tokens 输入价格 (缓存未命中)	百万tokens 输出价格 输出价格
deepseek-chat	64K	-	8K	0.07美元	0.27美元	1.10美元
deepseek-reasoner	64K	32K	8K	0.14美元	0.55美元	2.19美元 ⁽⁵⁾

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!



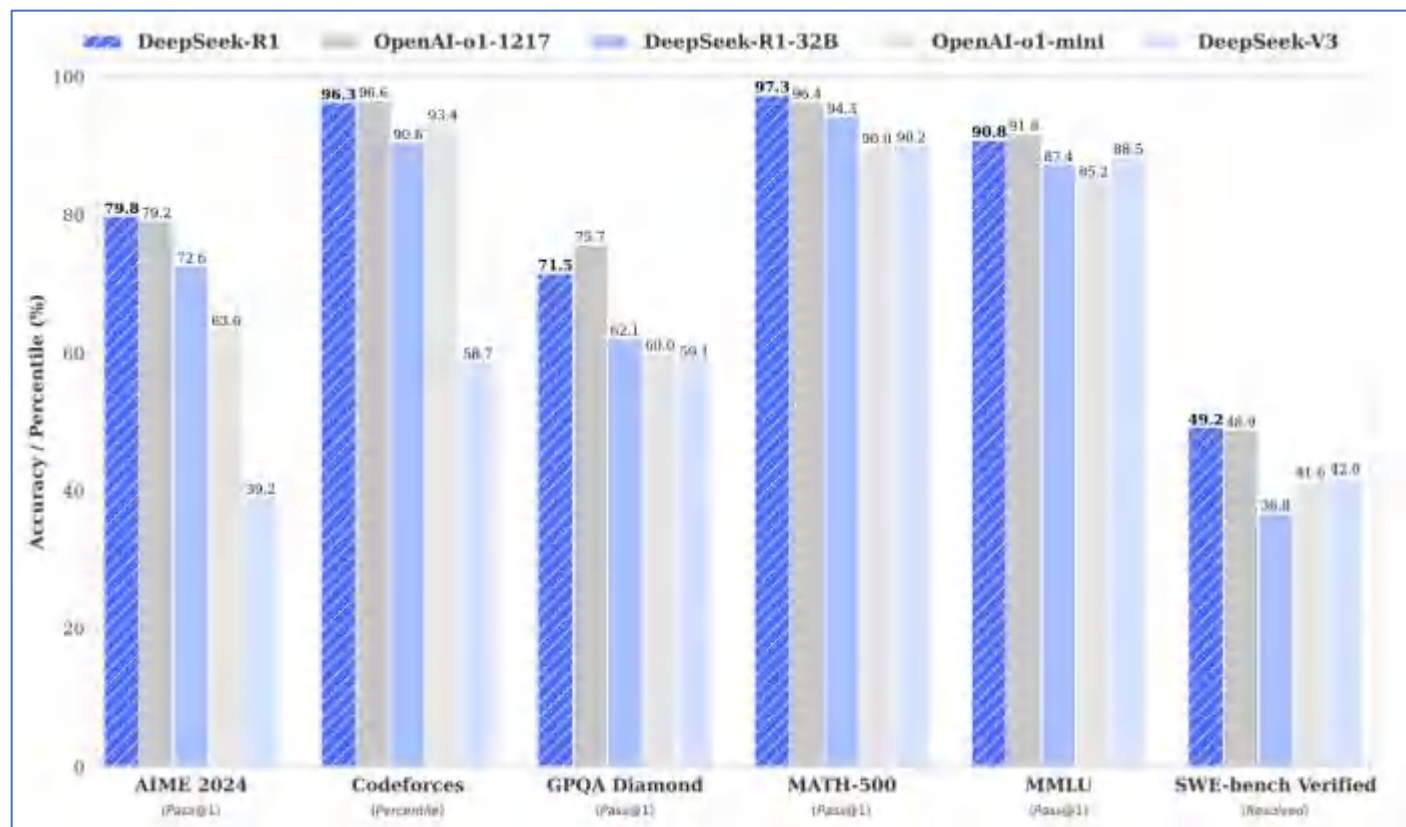
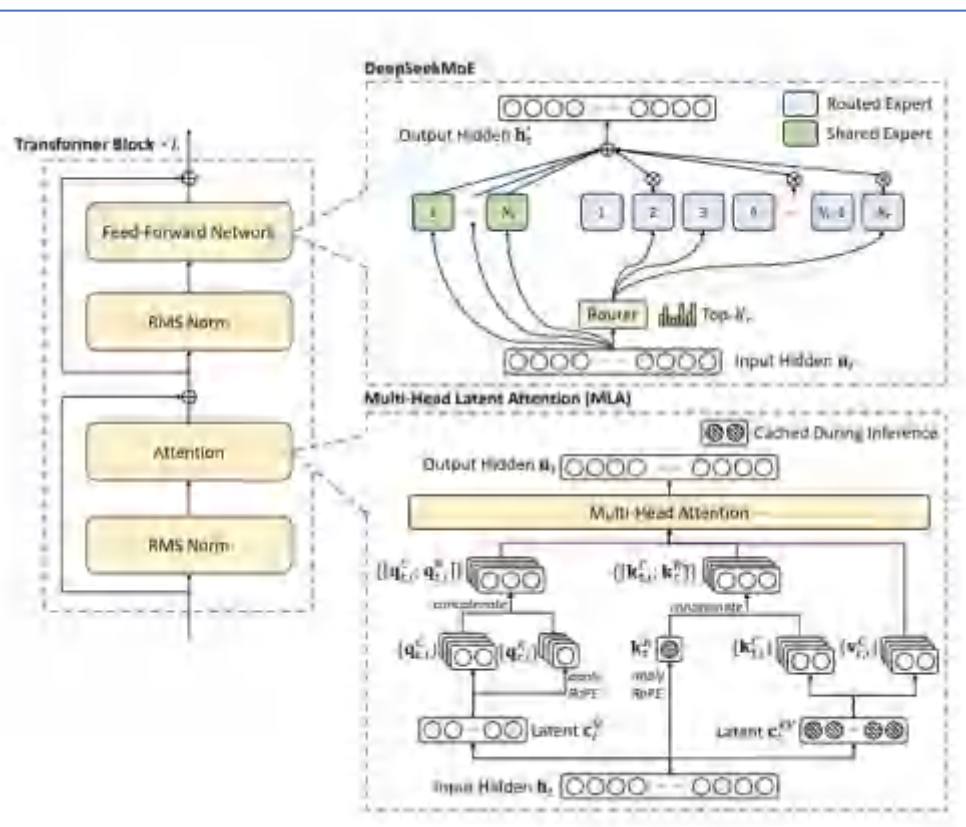
下载模型，可以本地安装，本地使用!

<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)



2. 秘塔搜索+: metaso.cn (切换为长思考-R1模式即可)

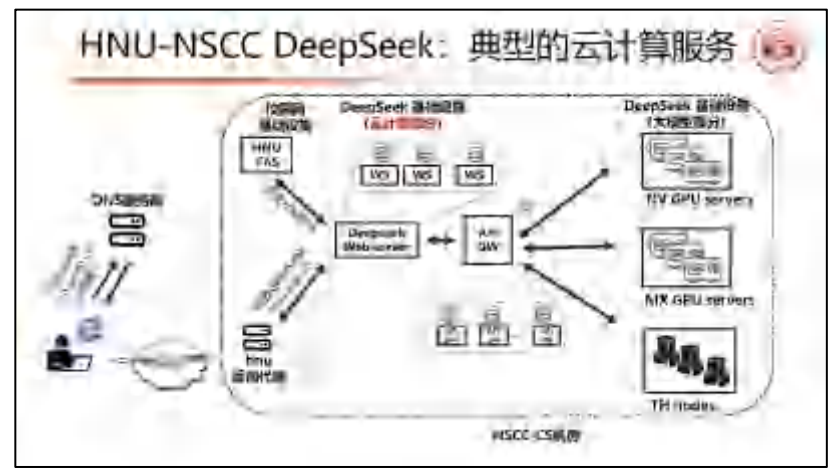
3. 360纳米AI搜索+: n.cn (需要提问后再点击右下角才能切换为Deepseek-R1)



还有很多, 不一一列举。。。

本地自己搭一套!

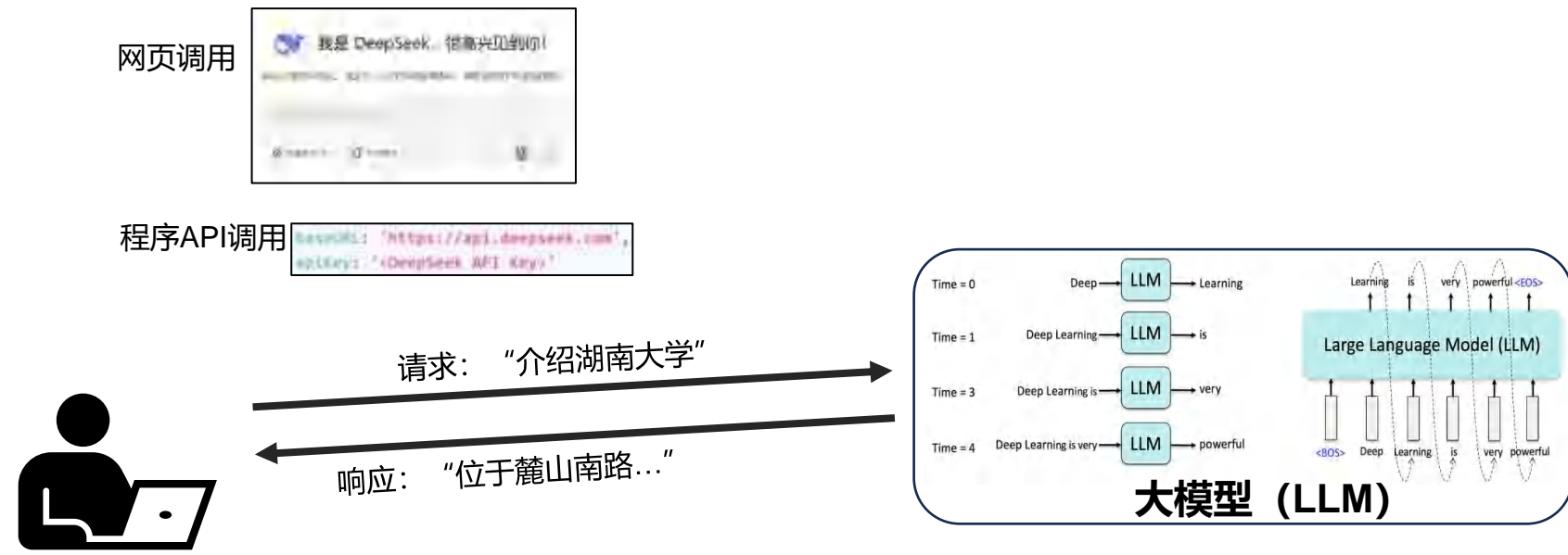
<https://deepseek.hnu.edu.cn/>



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!

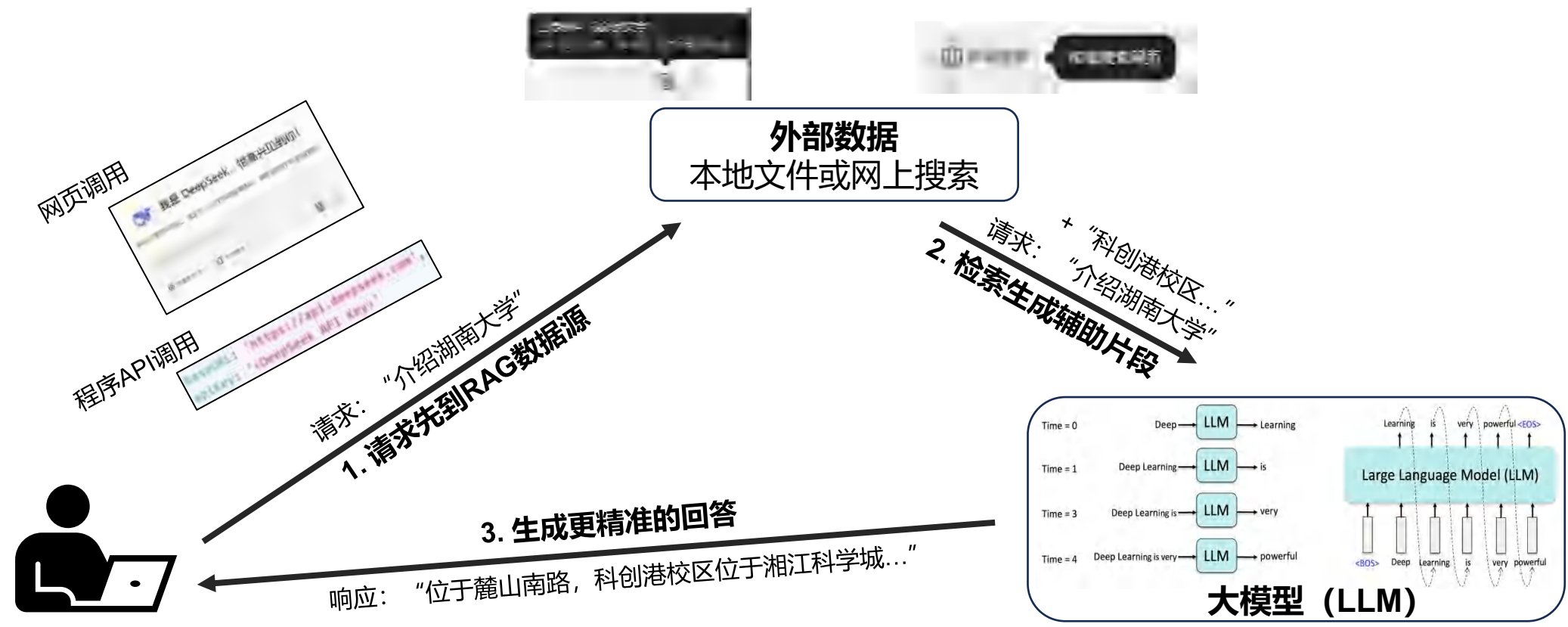
调用DeepSeek服务的流程：普通调用

- 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



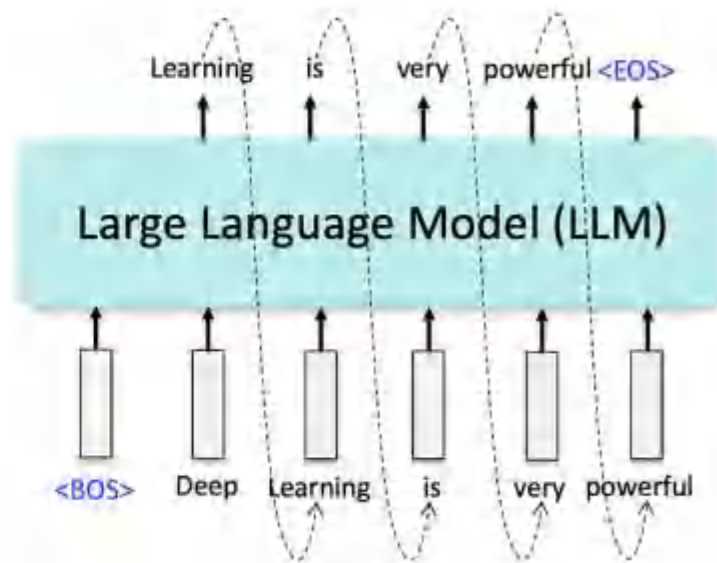
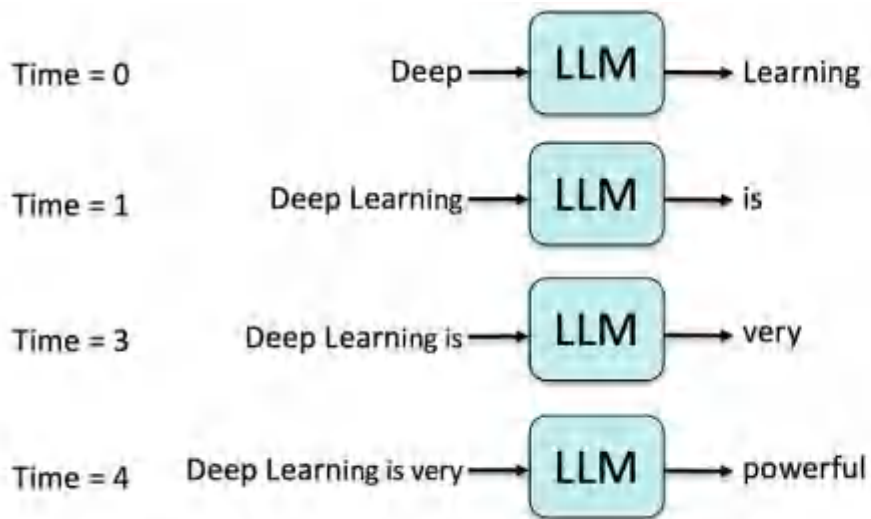
调用DeepSeek服务的流程：文件和联网搜索（RAG）

- 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

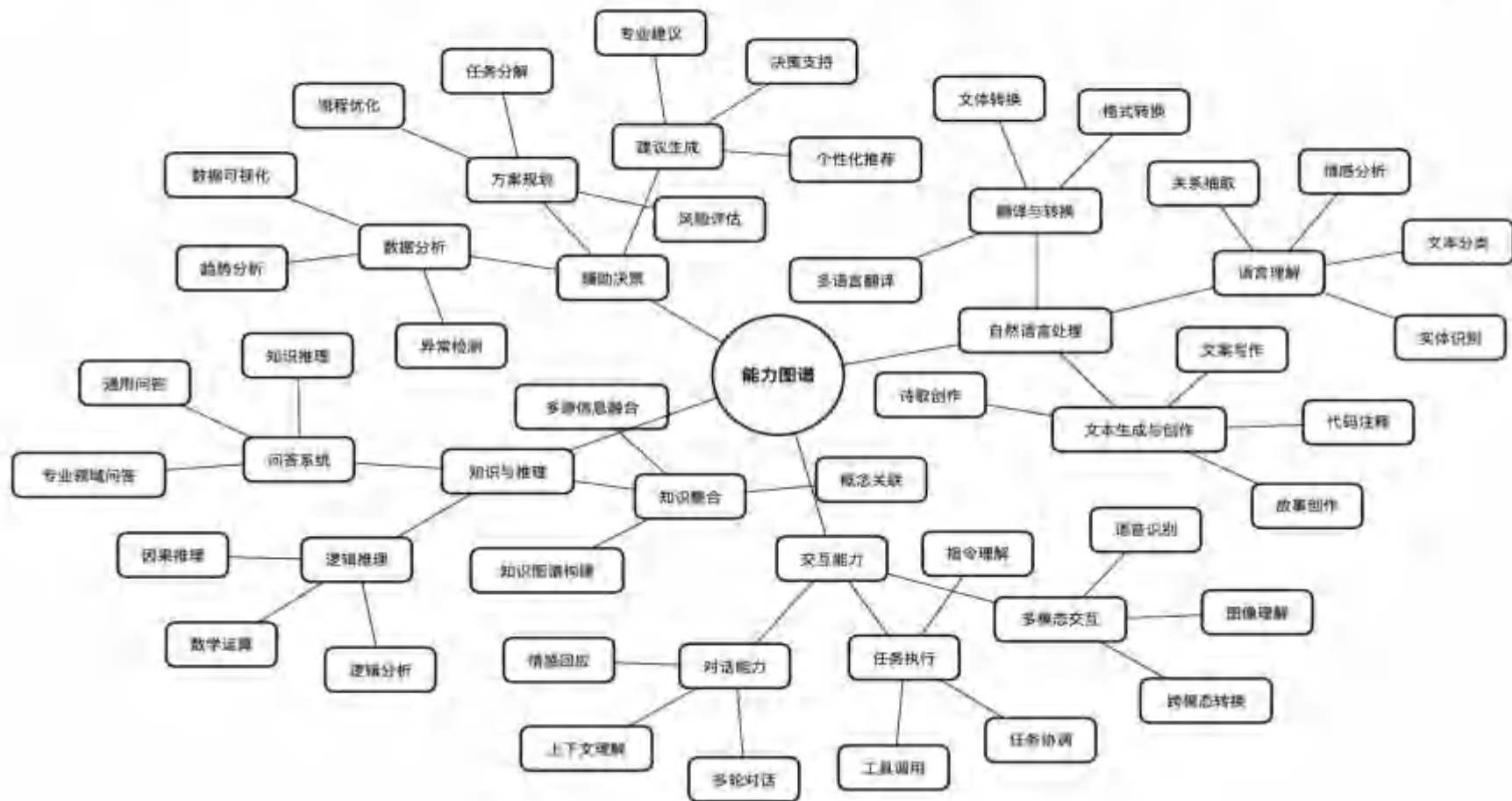


提纲

- 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

构造一个中译英的例子，对应编译器

88 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

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】

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

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++)`

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 (类型检查)
- “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、函数参数数

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

信息梳理

■ 案例：整理deepseek出来之前gpt o1的收费情况做对比



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

数据分析

■ 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 文学硕士	应届毕业生
							大学本科 工学学士	



输出结果：

序号	姓名	招录单位	学历学位
1		湘西自治州老干部活动中心	大学专科
2		湘西自治州州委移民事务中心	大学专科
3		湘西自治州州委组织部	大学专科
4		湘西自治州州委组织部	大学专科
5		湘西自治州州委组织部	大学专科
6		湘西自治州州委组织部	大学专科
7		湘西自治州州委组织部	大学专科
8		湘西自治州州委组织部	大学专科
9		湘西自治州州委组织部	大学专科
10		湘西自治州州委组织部	大学专科
11		湘西自治州州委组织部	大学专科
12		湘西自治州州委组织部	大学专科
13		湘西自治州州委组织部	大学专科
14		湘西自治州州委组织部	大学专科
15		湘西自治州州委组织部	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	高中
19		湘西自治州老干部活动中心	高中
20		湘西自治州州委移民事务中心	高中

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. 特别说明：
 - 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。



咨询分析

案例：分析某专家研究特长，给出研究方向建议咨询

引用4个网页, 5篇公众号文章

已深度思考 (用时25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:

- 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。
- 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、CON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。
- 学术与荣誉**
 - 发表论文40余篇 (含NSDL, 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。
- 研究理念**
他倡导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”, 通过校企合作新决实际技术瓶颈。其团队与联想、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。

引用4个网页, 7篇公众号文章

已深度思考 (用时41秒)

陈果教授作为国家超算长沙中心常务副主任, **非合适** 向教育厅领导介绍DeepSeek的)

- 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实践模型推理(文献5)。其团队清晰展示AI技术在教育场景的可行性。
- 教育应用场景适配性**
 - 教学辅助: DeepSeek已用于开发课程;
 - 科研支持: 湖南省科技厅教学案例显示, 可辅助论文写作、数据分析;
 - 算力普惠: 校园版服务实现异构异构的算力手。
- 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, 覆盖更多场景。

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:

- 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。
- 李德毅教授 (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。
- 王德胜教授 (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。
- 张松海教授 (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, 优先选择陈果教授或李德毅教授;
- 若需突出国家级平台背书与产业转化经验, 则张松海教授或王德胜教授更合适。

写程序

■ 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描述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;
}
```

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滚动名单
- 使用 CSS 动画实现无级滚动效果
- 滚动速度恒定
- 抽签过程
- 点击“开始抽签”按钮后名单加速滚动
- 再次点击“停止抽签”后暂停滚动结果
- 结果显示在页面下方

使用方法

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



做教辅

案例：设计一个针对幼儿园小朋友的科普讲稿

已深度思考
好的！以下是为幼儿园小朋友设计的科普讲稿。用户让我设计一个科普讲稿，用户让我设计一个科普讲稿。首先，我应该从孩子的角度出发，所以需要用比喻和故事来解释。接下来，功能性和互动性很重要。老师或家长可能会使用。时间控制也很重要。小朋友有直观的教具。还要考虑用户的安全。所以讲稿里要有互动环节。最后，结构要清晰，避免枯燥。可能

好的！以下是为幼儿园小朋友设计的科普讲稿。用户让我设计一个科普讲稿。首先，我应该从孩子的角度出发，所以需要用比喻和故事来解释。接下来，功能性和互动性很重要。老师或家长可能会使用。时间控制也很重要。小朋友有直观的教具。还要考虑用户的安全。所以讲稿里要有互动环节。最后，结构要清晰，避免枯燥。可能

标题：神奇的计算机小精灵

（老师拿出一个玩具电脑或平板电脑）

小朋友们早上好！今天我们要认识一位神奇的“小精灵”哦！你们见过它吗？在家里、教室里、商店里，它无处不在。

第一部分：计算机像什么？

计算机呀，长得像一本厚厚的书，像一台桌子，像一台电视机。有的计算机像一本厚厚的书，有的像一台桌子，有的像一台电视机。

第二部分：计算机的“身体”在哪里？

计算机小精灵的身体藏在很多地方：键盘像一排排小牙齿，鼠标像一只小老鼠，摄像头是计算机的小眼睛。

第三部分：计算机能做什么？

计算机小精灵可厉害了！它能帮我们做很多事情：它能帮我们算数，它能帮我们画画，它能帮我们讲故事。

标题：超级计算机——一万个大脑

（老师举起一只手，假装变魔术）

小朋友们，如果老师有一只超级计算机，它能帮我们做很多事情：它能帮我们算数，它能帮我们画画，它能帮我们讲故事。

第一部分：超级计算机有多厉害？

普通计算机像一辆小自行车，超级计算机像一架超音速飞机。它能一秒就能绕地球7圈！

第二部分：超级计算机的秘密武器

它的秘密就是——好多好多的大脑！普通计算机只有1个大脑，超级计算机有10000个大脑！

第三部分：超级计算机能做什么？

它能帮我们做很多事情：它能帮我们算数，它能帮我们画画，它能帮我们讲故事。

标题：超级计算机——一万个大脑

（老师举起一只手，假装变魔术）

小朋友们，如果老师有一只超级计算机，它能帮我们做很多事情：它能帮我们算数，它能帮我们画画，它能帮我们讲故事。

第一部分：超级计算机有多厉害？

普通计算机像一辆小自行车，超级计算机像一架超音速飞机。它能一秒就能绕地球7圈！

第二部分：超级计算机的秘密武器

它的秘密就是——好多好多的大脑！普通计算机只有1个大脑，超级计算机有10000个大脑！

第三部分：超级计算机能做什么？

它能帮我们做很多事情：它能帮我们算数，它能帮我们画画，它能帮我们讲故事。

标题：超级计算机——一万个大脑

（老师举起一只手，假装变魔术）

小朋友们，如果老师有一只超级计算机，它能帮我们做很多事情：它能帮我们算数，它能帮我们画画，它能帮我们讲故事。

第一部分：超级计算机有多厉害？

普通计算机像一辆小自行车，超级计算机像一架超音速飞机。它能一秒就能绕地球7圈！

第二部分：超级计算机的秘密武器

它的秘密就是——好多好多的大脑！普通计算机只有1个大脑，超级计算机有10000个大脑！

第三部分：超级计算机能做什么？

它能帮我们做很多事情：它能帮我们算数，它能帮我们画画，它能帮我们讲故事。

正确理解，打开广阔天地

正确理解DeepSeek的**不能**

DeepSeek (R1或V3) 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**

以"GuoLab"这个英文词做拜年诗一首，可以中英文夹杂，加入科研主题

COM的韵脚和中文的抗字相同，可以尝试和中文押韵

已深度思考 (用时 21 秒)

《GuoLab科研贺岁》
Genome解码旧年功，
Unlock新意觅迹踪。
Open智识迎瑞鹤，
Lab灯火映春鸿。
Analyze数据千重浪，
Bench前贺岁万家丰。

赏析：这首作品以“GuoLab”分析等科研场景与鹤兆祥瑞错，暗合实验室跨年攻关的丰年”的新春祝愿，理性与

对不起，我还没有学

以“GuoLab”

已深度思考 (用时 14)

《GuoLab新春志》
Guo融瑞雪润琼枝，
Lab点梅香赋岁辞。
数据织成桃符锦，
荧屏守岁创新时。

赏析：此诗以“GuoLab”“Lab点梅香”将荧屏守岁”既写科研人员春符号，在保留拜年

《GuoLab网路春芳》
Guo启新元覆网康，
Lab燃红梅映顶红。
拥塞推开SIGCOMM，
流量织就IEEE筐。
千帆校验瓣纹烫，
万兆带宽锦鲤翔。
最是论文除夕宴，
路由春酒醉SIGBED。

注：尾韵调整为ang韵（康/刊/筐/翔/宴/BED），其中：
• IEEE筐：借IEEE期刊集群喻学术丰收，暨谐音COM且押ang韵
• SIGBED：嵌入式系统顶会，尾音“BED”与“宴”形成中英混搭趣味

赏析：此诗以计算机网络架构化为祈福代码。中英术语如Fire的科技浪漫，又置“带宽纳财”的守岁，于协议栈上贺春。展现

《GuoLab网路春芳》
Guo启新元覆网康，
Lab燃红梅映顶红。
拥塞推开SIGCOMM，
流量织就IEEE筐。
千帆校验瓣纹烫，
万兆带宽锦鲤翔。
最是论文除夕宴，
路由春酒醉SIGBED。

注：尾韵调整为ang韵（康/刊/筐/翔/宴/BED），其中：
• IEEE筐：借IEEE期刊集群喻学术丰收，暨谐音COM且押ang韵
• SIGBED：嵌入式系统顶会，尾音“BED”与“宴”形成中英混搭趣味

赏析：此诗将计算机网络术语与科研与数据流量。中段“路由撞报”以颠覆传统烟火意象，将除夕代码逐基与碳基的交响中，完成对科研人

是代码除夕夜，
论文焰火照AI开。

赏析：此诗将“Fire”是代码除夕夜，论文焰火照AI开。

是代码除夕夜，
论文焰火照AI开。

赏析：此诗将“Fire”是代码除夕夜，论文焰火照AI开。

《GuoLab网路春芳》
Guo启新元覆网康，
Lab燃红梅映顶红。
拥塞推开SIGCOMM，
流量织就IEEE筐。
千帆校验瓣纹烫，
万兆带宽锦鲤翔。
最是论文除夕宴，
路由春酒醉SIGBED。

赏析：此诗以“拓补”替代IPv6，基能合同结构而又研位新年万象更新之态。嵌联直透SIGCOMM、INFOCOM两大顶刊为科研位位理理，指联“瓣纹”喻论文产研印迹，“锦鲤”取关美学会奖项与数据流流，尾联“焰火春酒”将实验烟火化作除夕酒。 “春酒”既指往日智慧又借吉“春酒”，喻新年刊成果如喜讯广布。全诗以网路协议为经纬，及刊名为经纬，在诗律流中流流出“格律”、“顶刊”的科研新春图景。

嵌入计算机网络顶刊群，首联以IPv6（下一代互联网协议）画新年新象，SIGCOMM（网络系统顶会）喻学术顶会，颈联“瓣纹”喻论文产研印迹，INFOCOM（通信领域顶会）喻学术顶会，尾联“焰火春酒”喻论文产研印迹，尾联将论文写作比作守岁烟火，将传统酒醉传统酒醉，在HTTP协议协议的平仄中，完成对学子“顶刊如雪片，奖金看寒来”的赛博朋克式祈愿。

赏析：此诗将“Fire”是代码除夕夜，论文焰火照AI开。

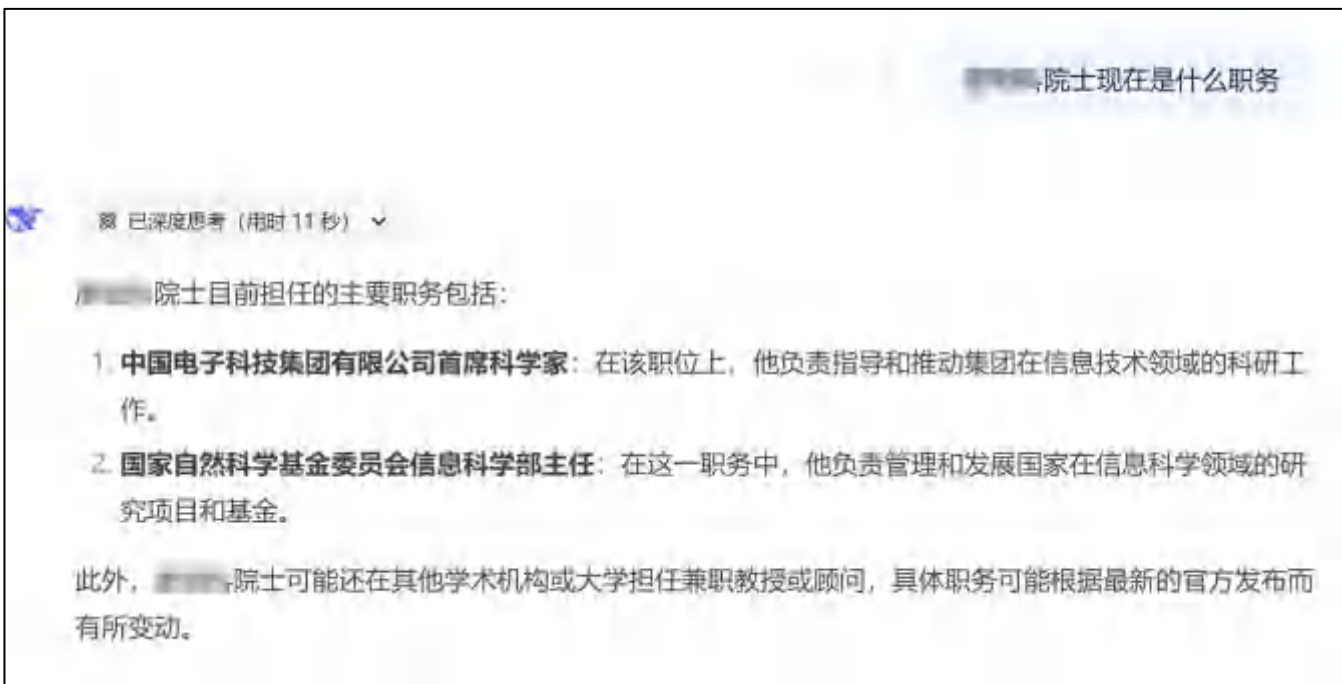
赏析：此诗将“Fire”是代码除夕夜，论文焰火照AI开。

赏析：此诗将“Fire”是代码除夕夜，论文焰火照AI开。

赏析：此诗将“Fire”是代码除夕夜，论文焰火照AI开。

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！

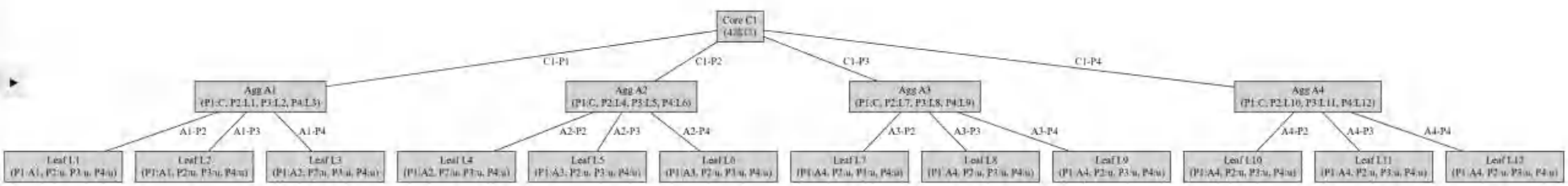


“尽信书不如无书”！要有判断筛选能力，擅用联网搜索和知识库！

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）

The screenshot shows a web-based network configuration tool on the left and a terminal window on the right. The terminal displays Graphviz DOT code for a network topology. The code defines a core layer with four switches (Core_C1 to Core_C4) and four aggregation layers (Agg_A1 to Agg_A4). Each aggregation layer is connected to the core and has three leaf nodes (Leaf_L1 to Leaf_L12). The code uses 'graph TD' and 'graph LR' to define the nodes and their connections.



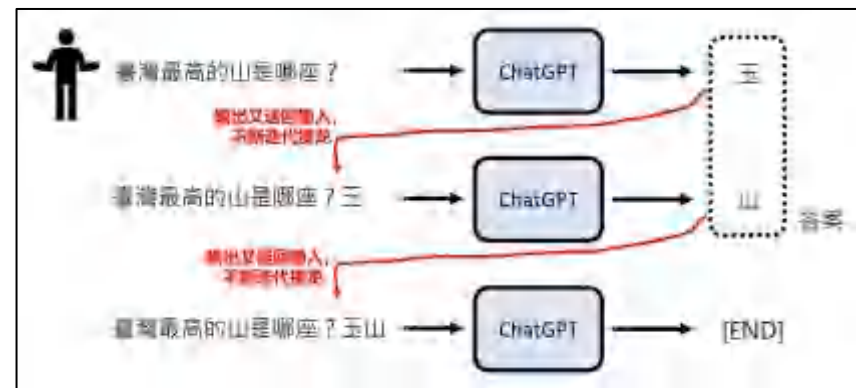
DeepSeek (R1或V3) 的不能

- 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

```

15  model = LLM(model='deepseek-chat', base_url='https://api.deepseek.com/')
16  # 输入提示词
17  all_messages = [{"role": "user", "content": "你好"}]
18
19  # 开始推理
20  while True:
21      # 添加消息到列表中
22      if (len(all_messages) - 1) % 2 == 0:
23          # 用户输入
24          user_input = input("请输入问题: ")
25          # 添加用户输入到消息列表
26          all_messages.append({"role": "user", "content": user_input})
27
28      # 开始推理
29      response = model.generate(prompt_tokens=all_messages[-1]['content'],
30                                max_tokens=1024, temperature=0.5)
31      # 打印推理结果
32      print(f"模型回答: {response}")
33
34      # 添加模型输出到消息列表
35      all_messages.append({"role": "assistant", "content": response})
36
37      # 生成推理指令
38      generate_prompt = f"请继续提问: "
39      # 生成推理指令
40      all_messages.append({"role": "user", "content": generate_prompt})
41
42      # 生成推理指令
43      generate_prompt = "\n\n"
44      all_messages.append({"role": "user", "content": generate_prompt})

```



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)

DeepSeek (R1或V3) 的不能

■ **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

行业大模型



通用大模型



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)

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!

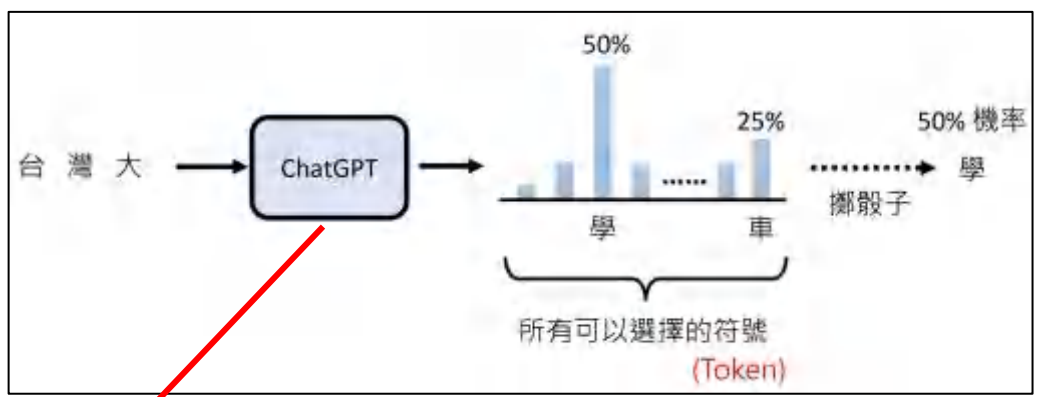
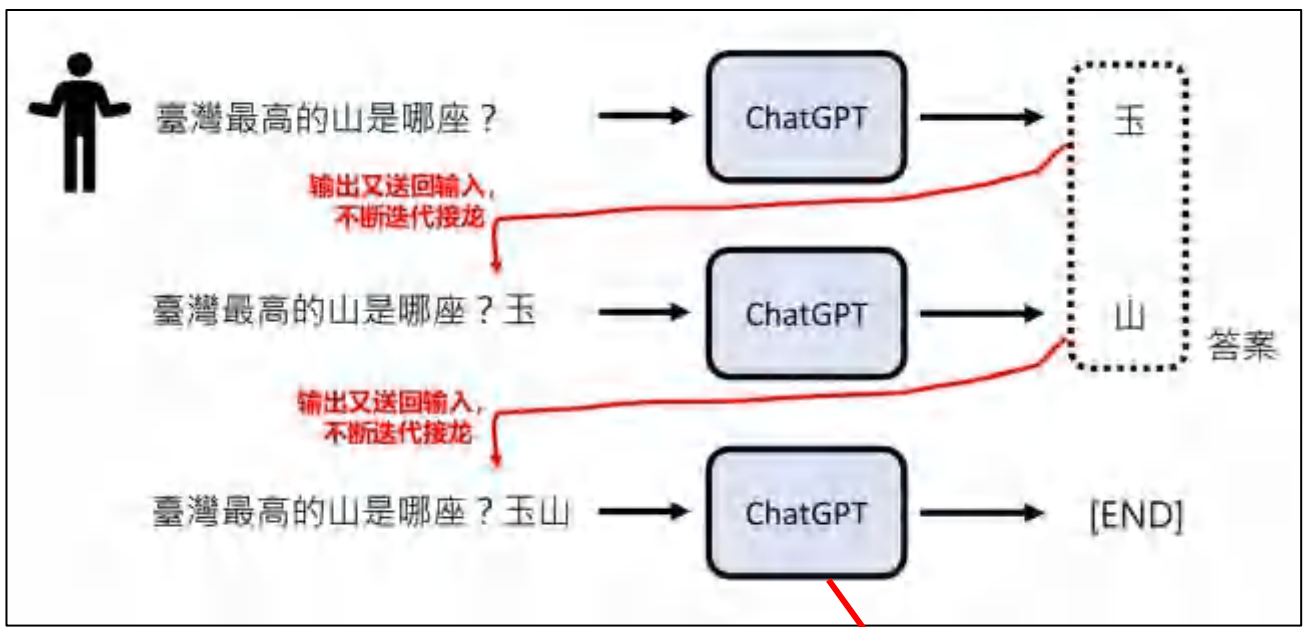
**会不会用，即将成为现代社会生产效率的分水岭！
会的人或组织，会远远甩开那些不会用的！
发挥你的创造力和能动性，赶紧用起来吧！**

提纲

- 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

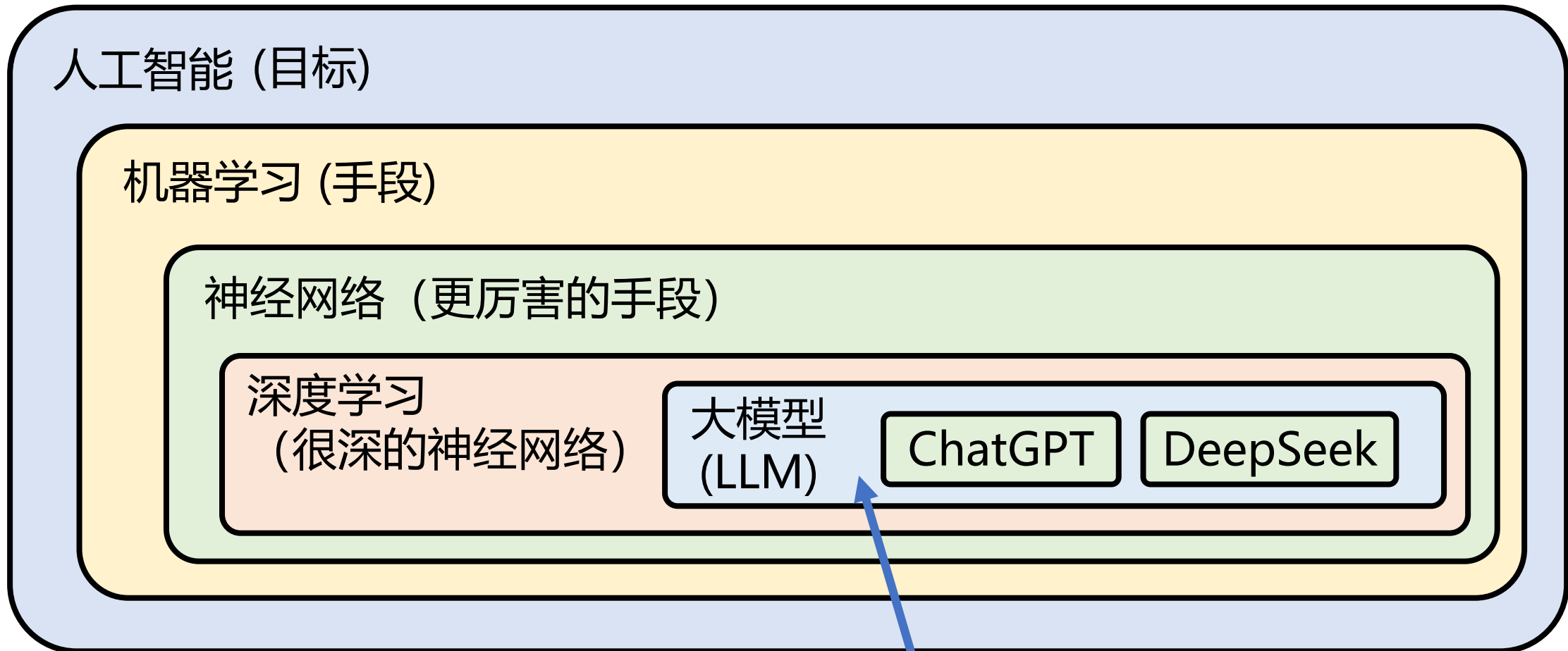
Transformer——大模型基础

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？

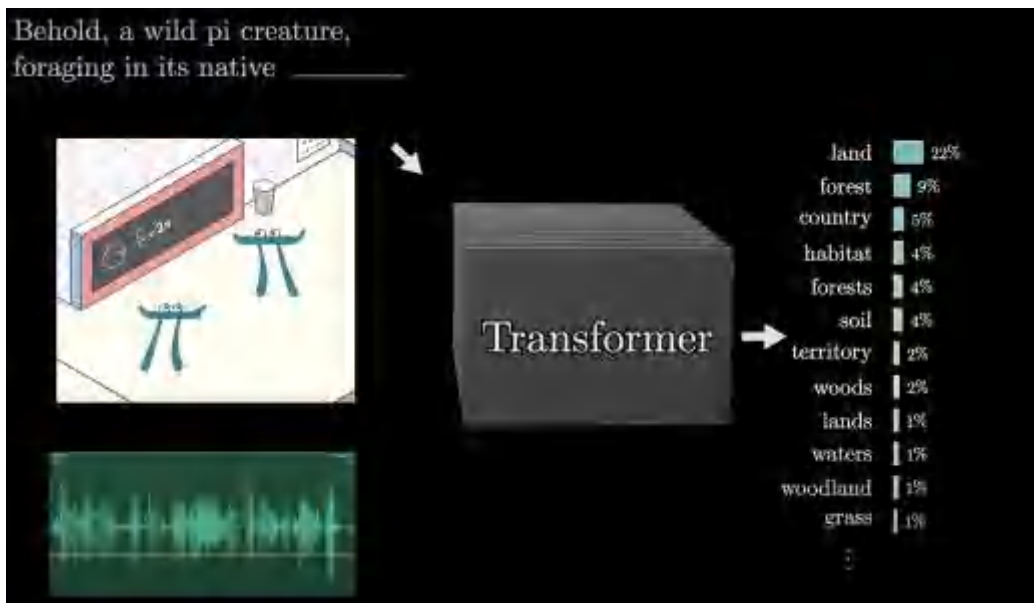
几个必须澄清的概念



Transformer
大模型常用的一种神经网络

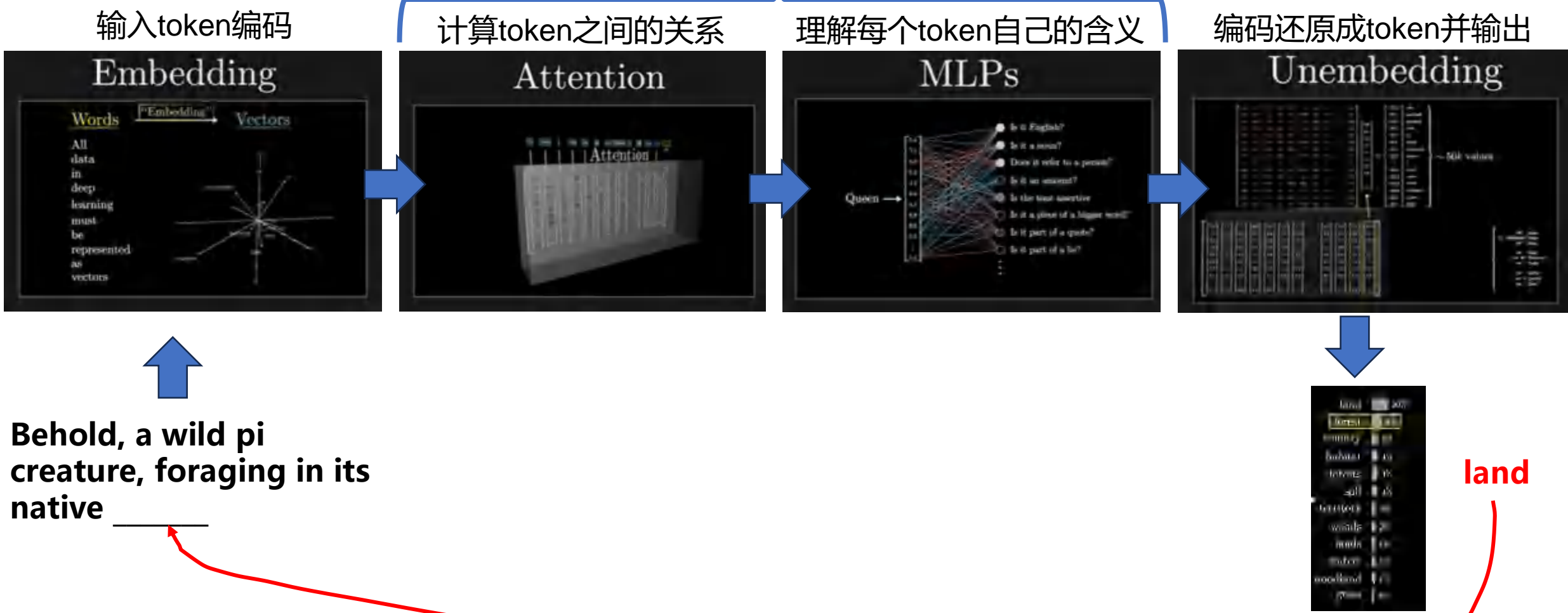
Transformer是什么

- 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

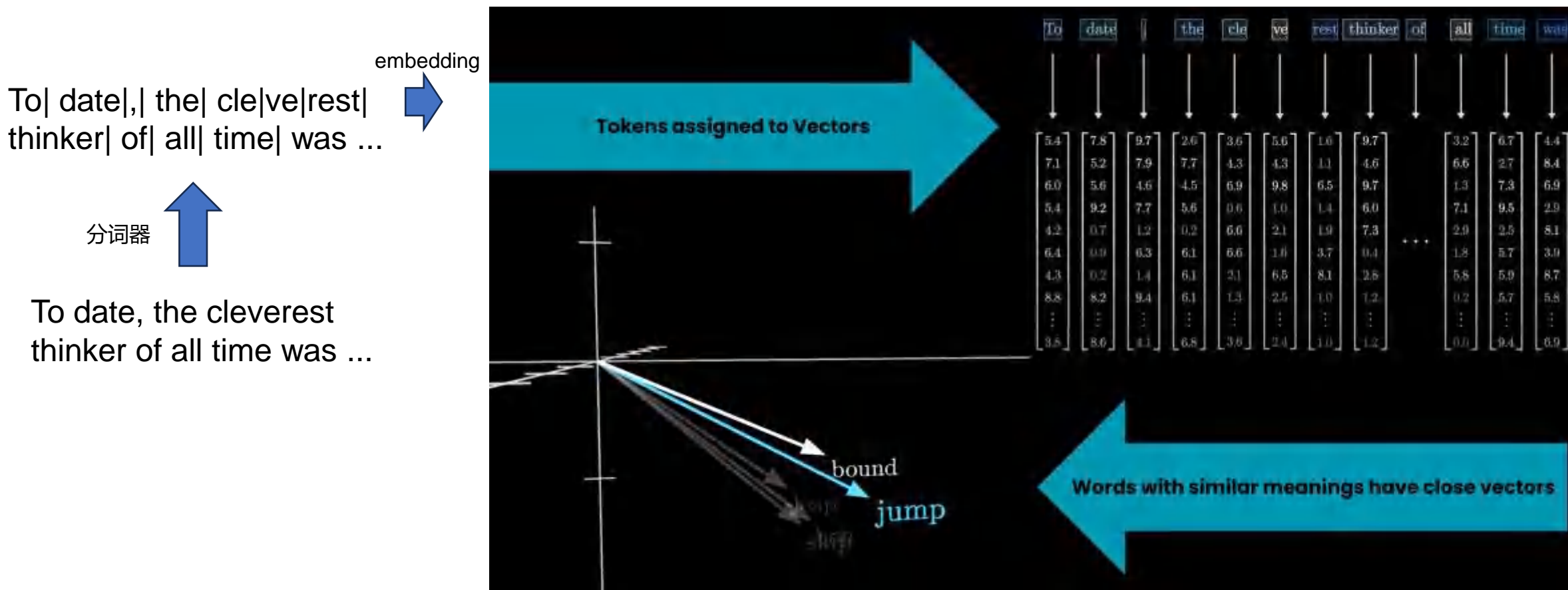
通常重复很多次



Behold, a wild pi creature, foraging in its native _____

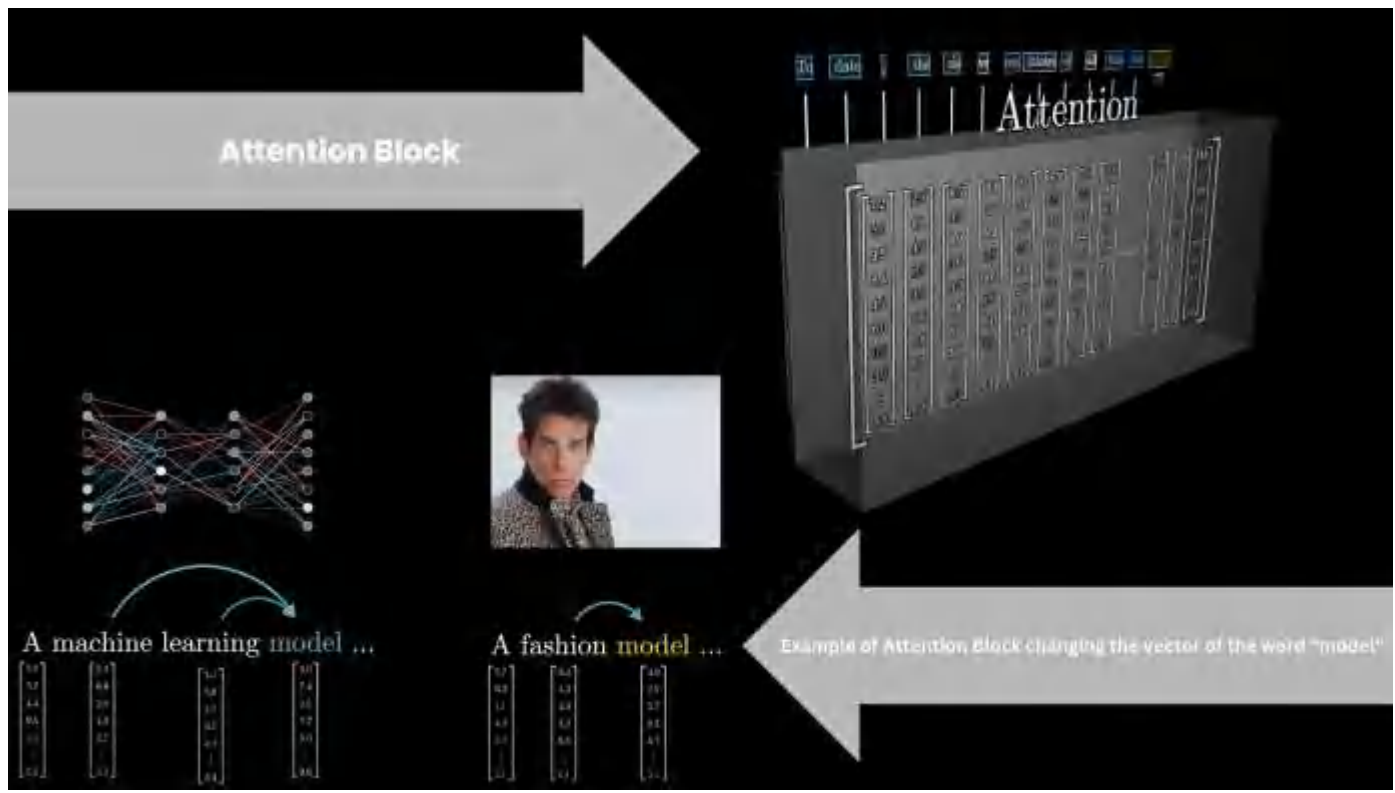
Embedding

- 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



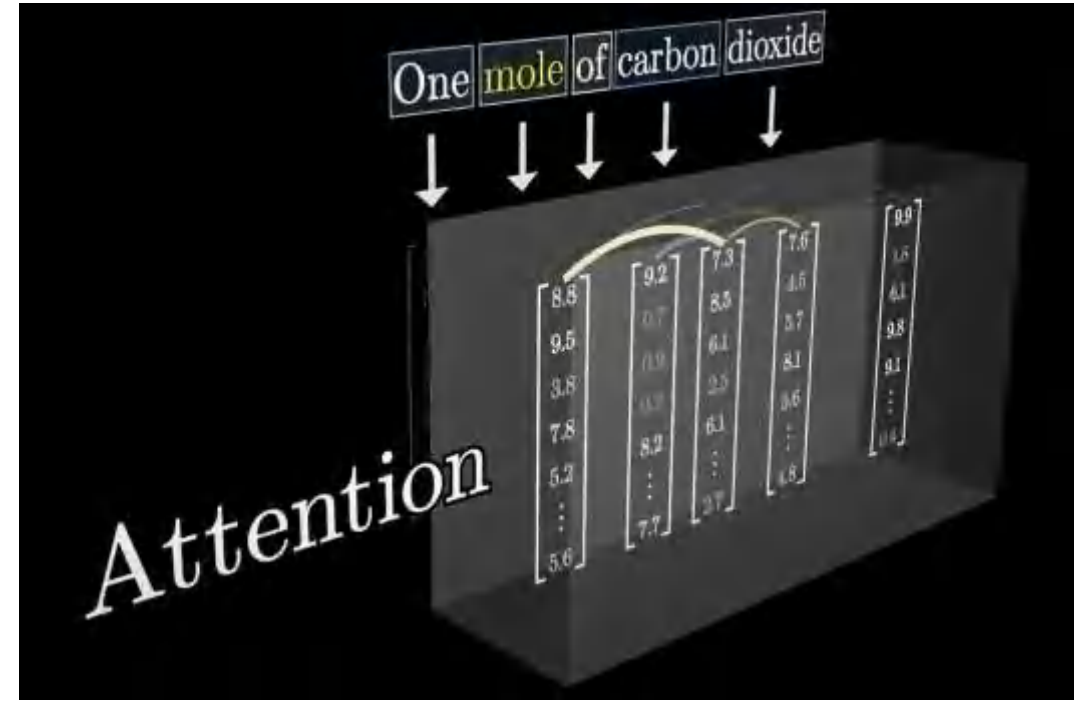
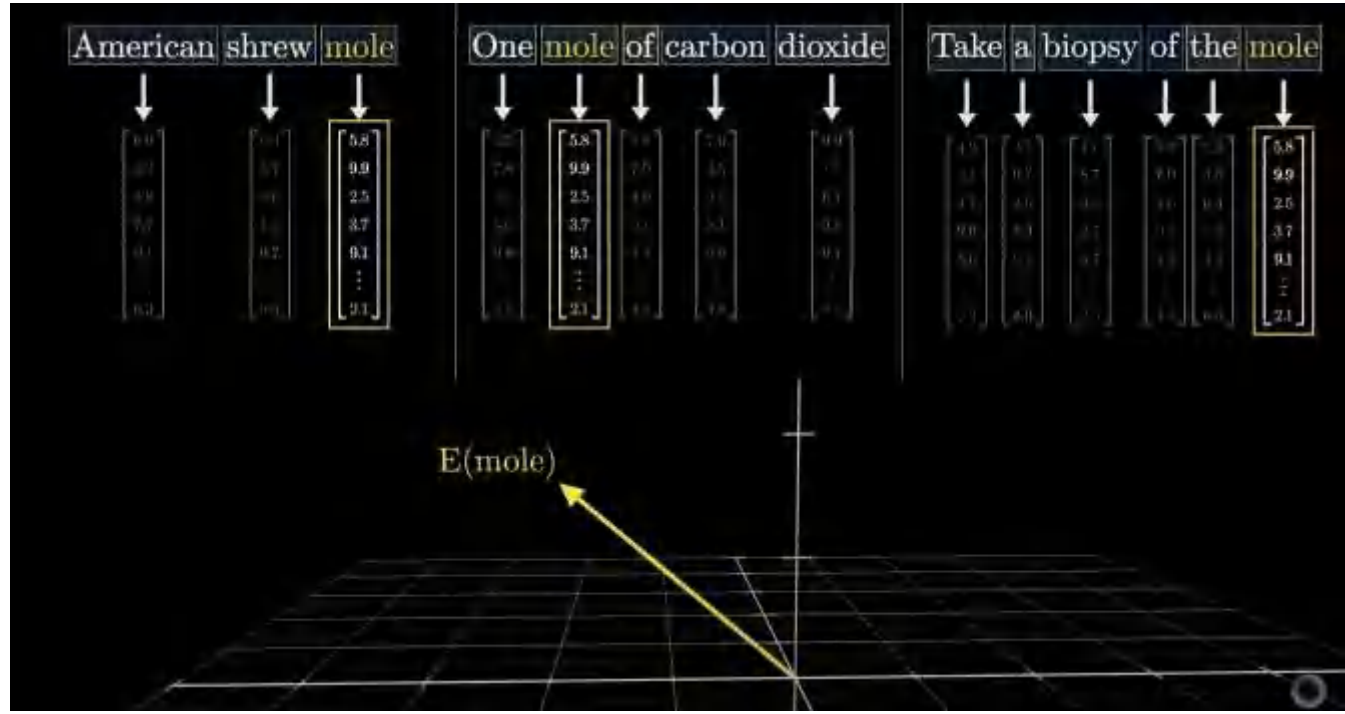
Attention

- 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



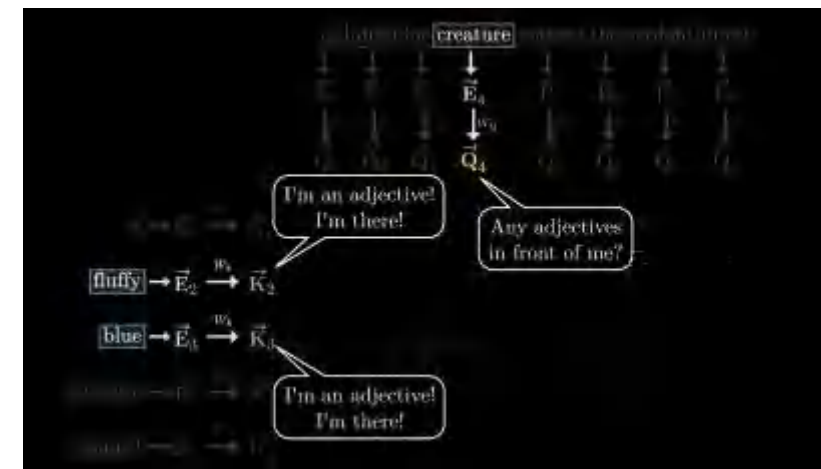
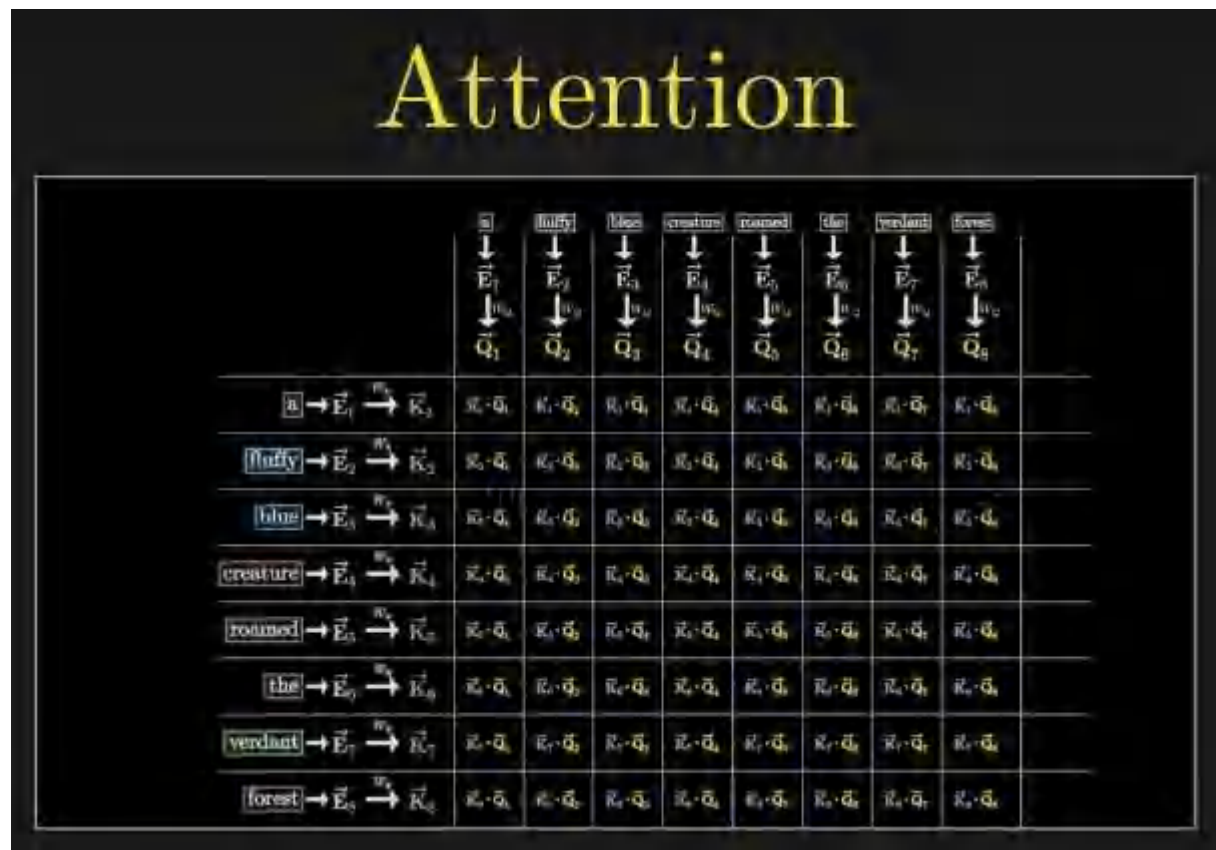
Attention: 多说两句

- 三句话都有mole这个词 (鼯鼠、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)

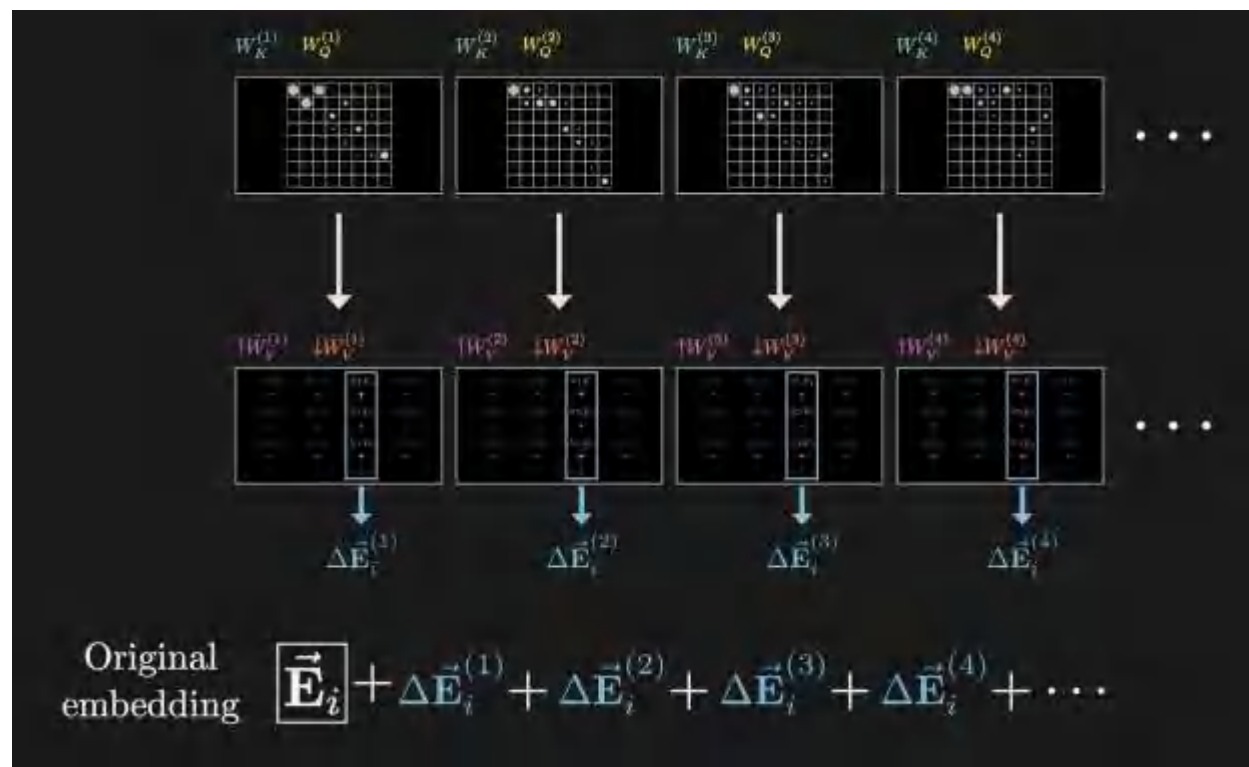
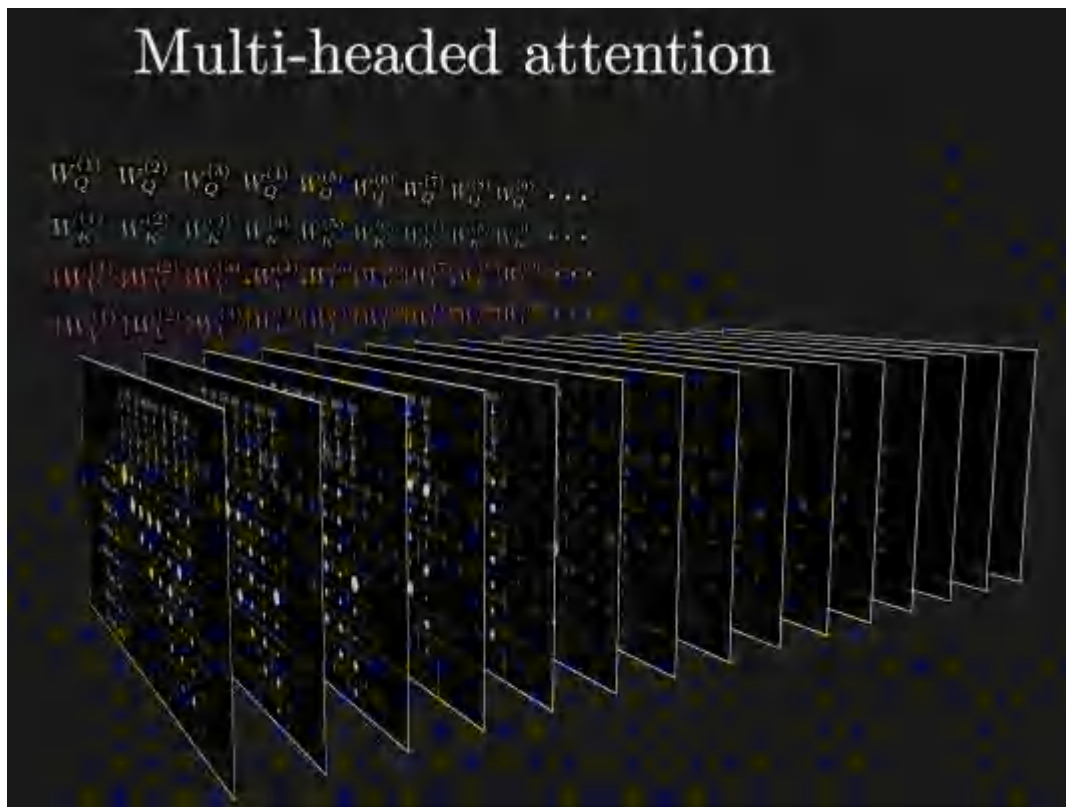


	a	fluffy	blue	creature	rounded	the	verdant	forest
a	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fluffy	0.00	1.00	0.00	0.42	0.00	0.00	0.00	0.00
blue	0.00	0.00	1.00	0.58	0.00	0.00	0.00	0.00
creature	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
rounded	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
the	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
verdant	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
forest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Attention: 多说两句

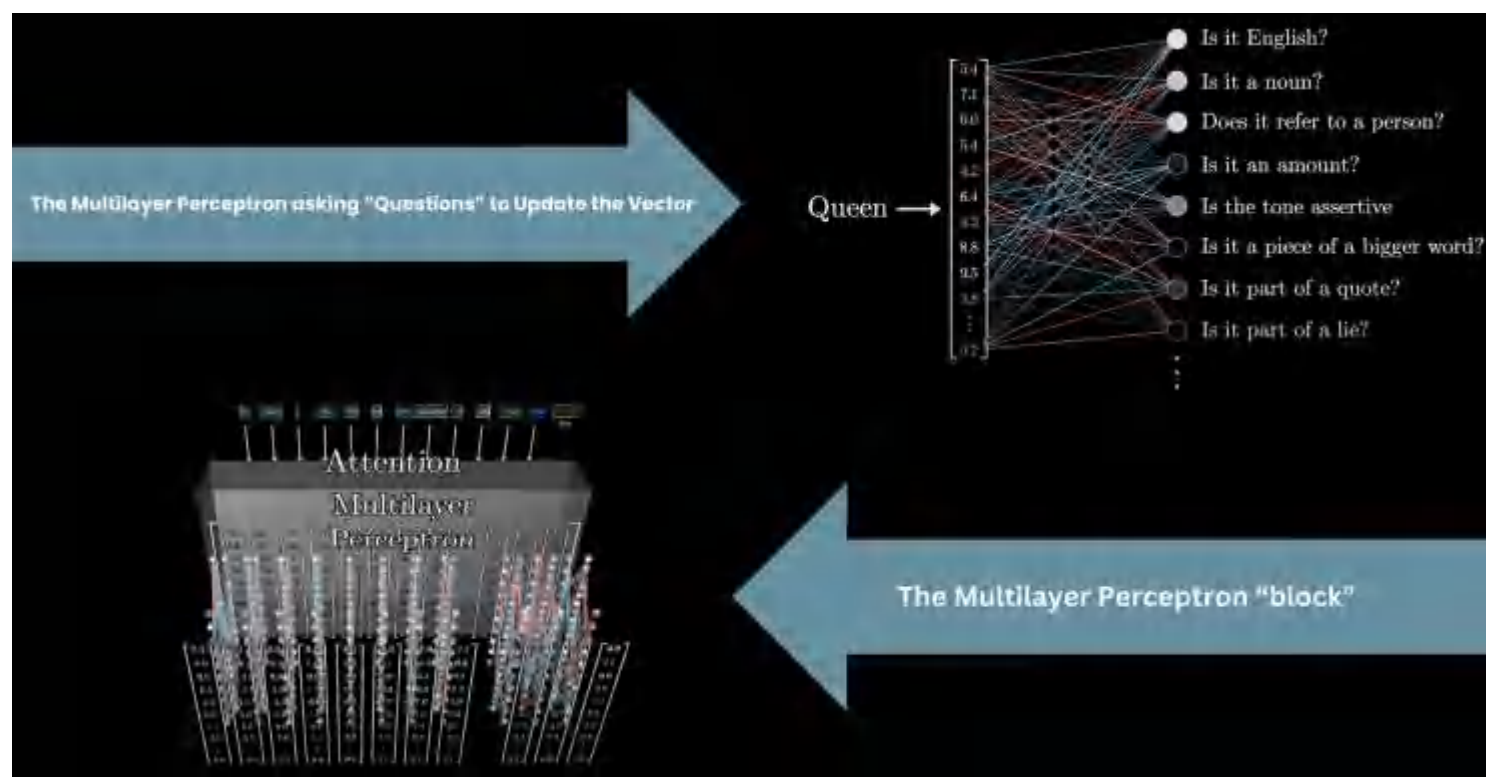
■ 多头注意力 (Multi-head Attention, MHA)

- 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



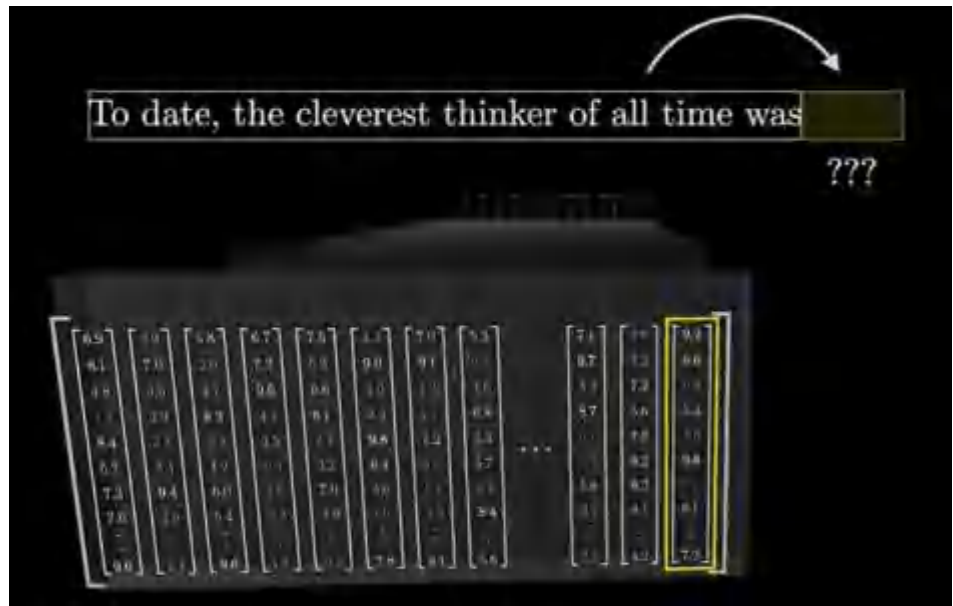
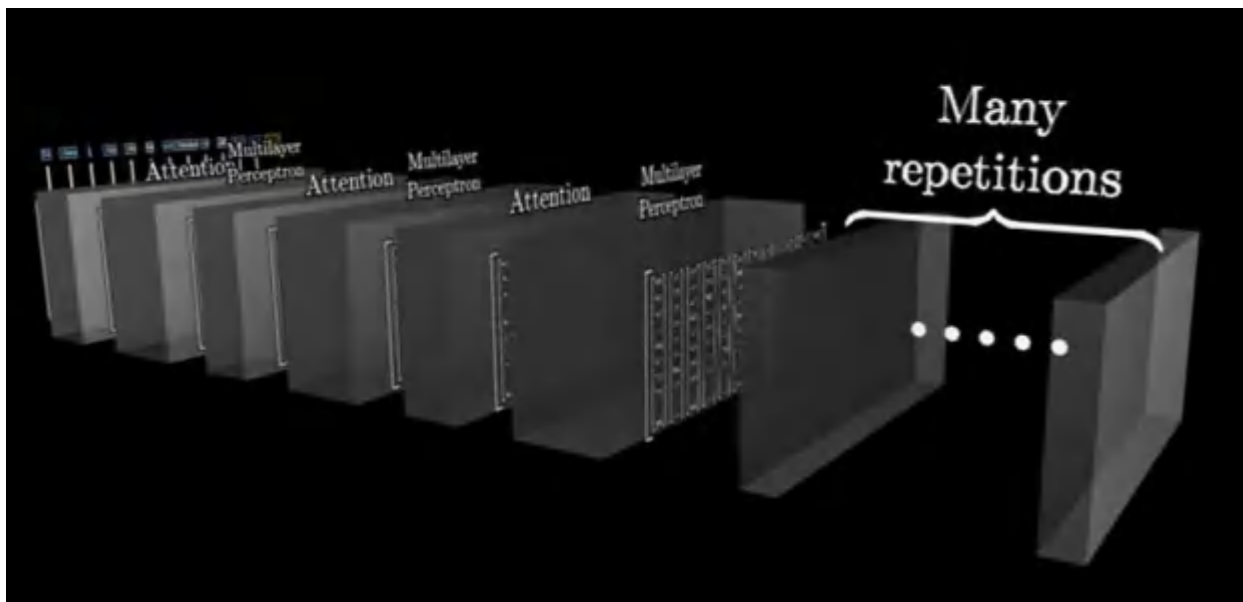
MLP(Multilayer Perceptron)

- 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

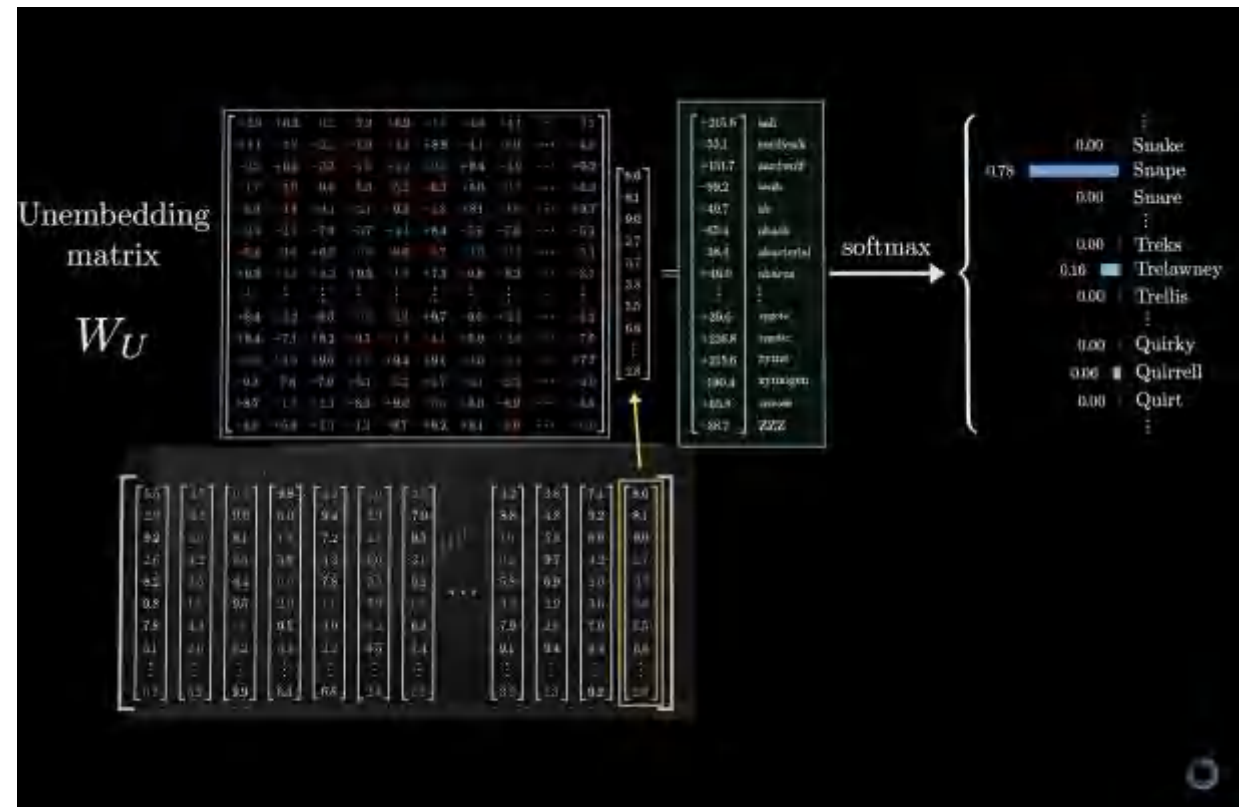
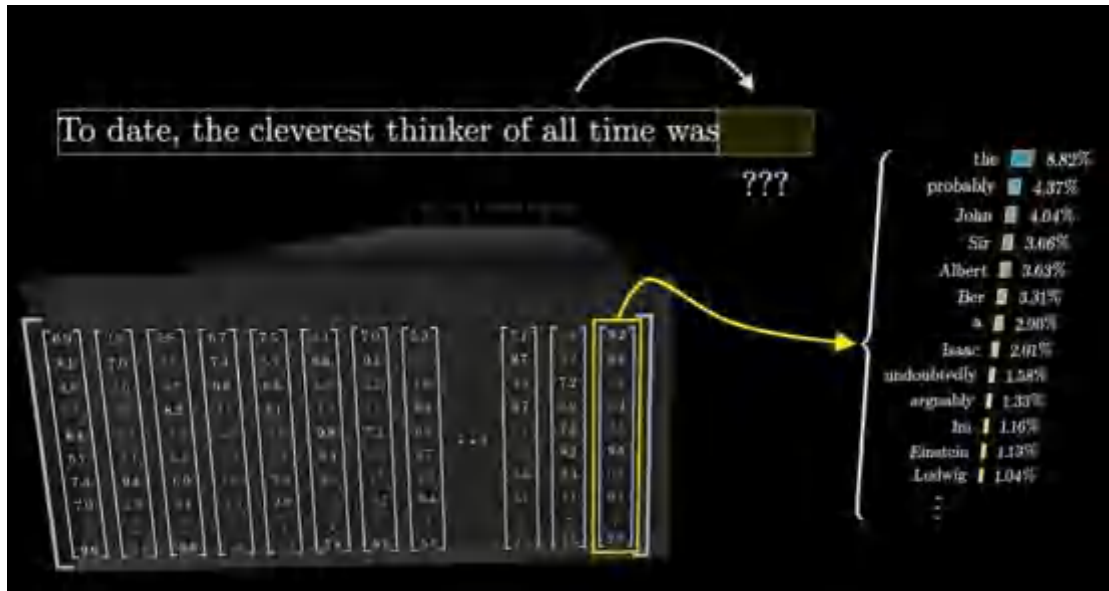
- 重复很多很多次Attention和MLP
 - 通常至少得几十次，每次都有**不同的权重矩阵**
 - 每个token之间和token自己的含义都被充分地加载到乘完的最后一个向量中



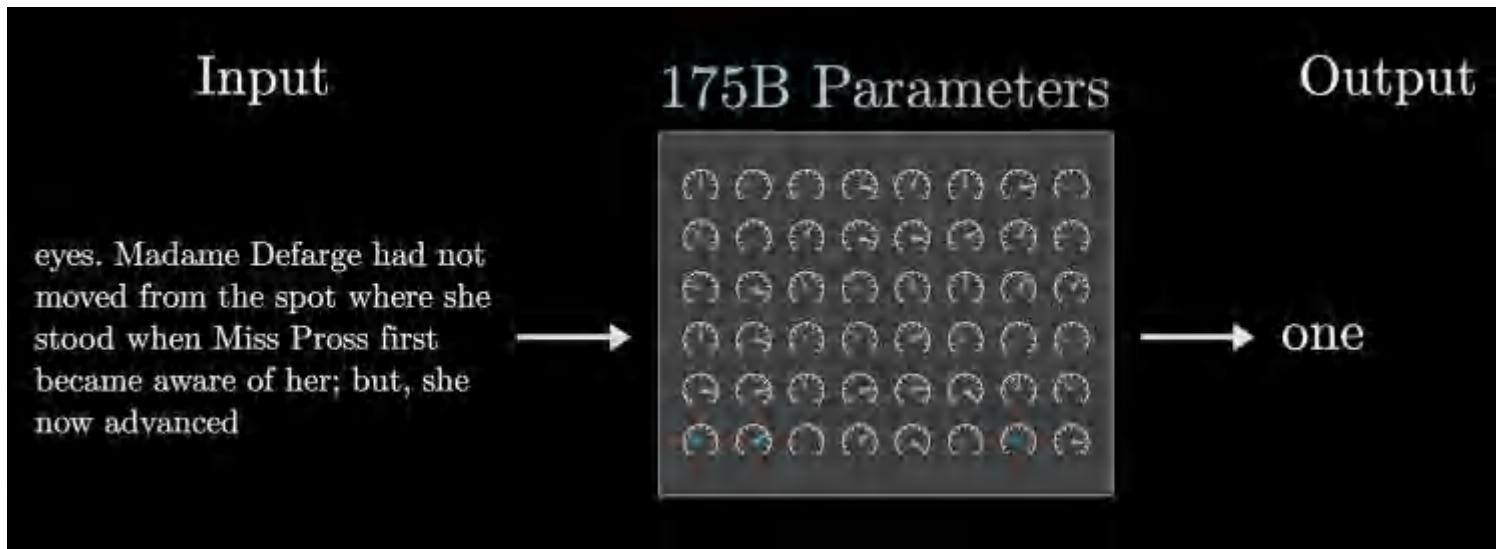
Unembedding

■ 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



上千亿个参数 (不断调整)



用大量的基础数据
预训练(Pre-trained)



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

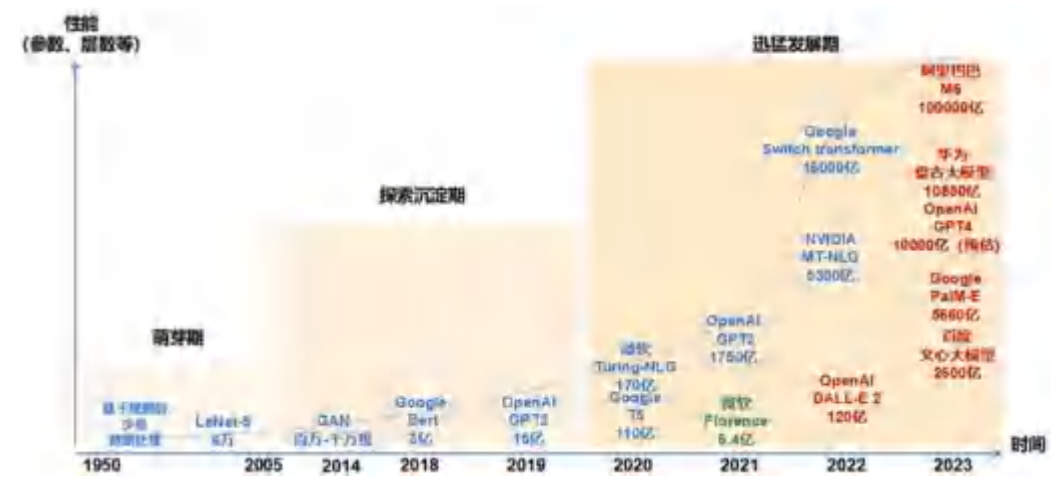
✗ 没调好, 再调!

不断生成内容(Generative)

数据和参数的规模越大，LLM就越聪明



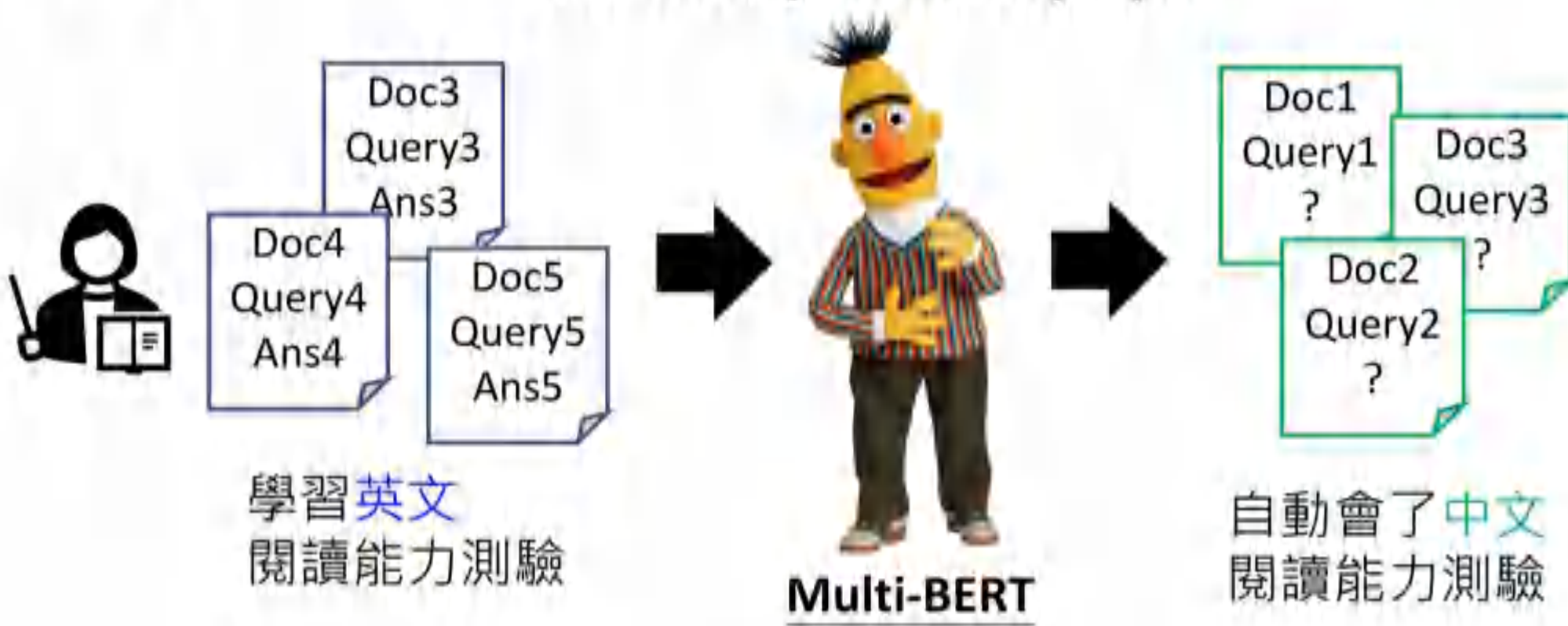
DeepSeek R1/V3
671B



大到一定規模，開始“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

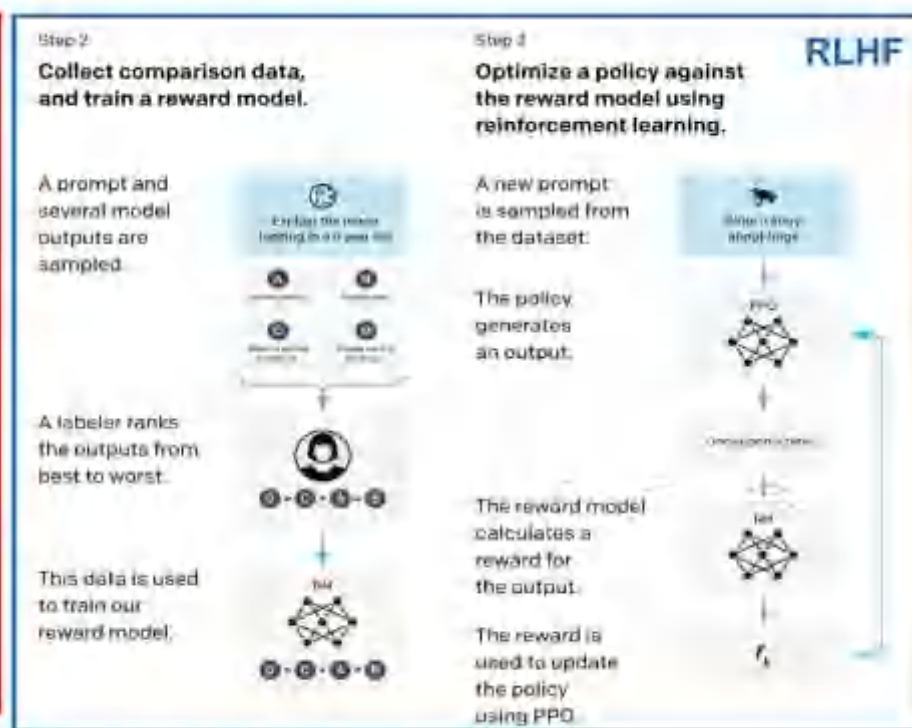
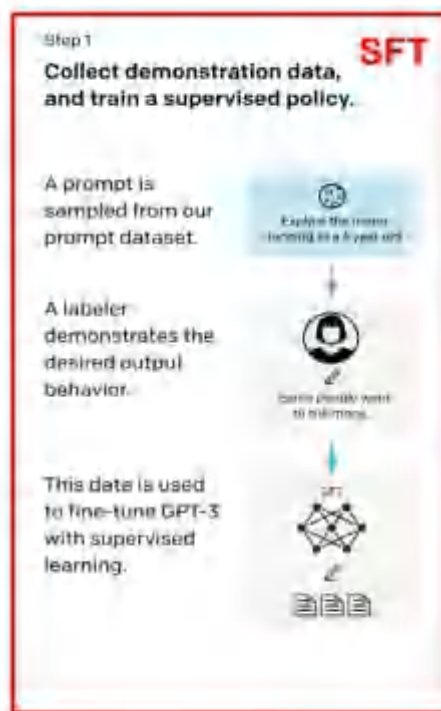
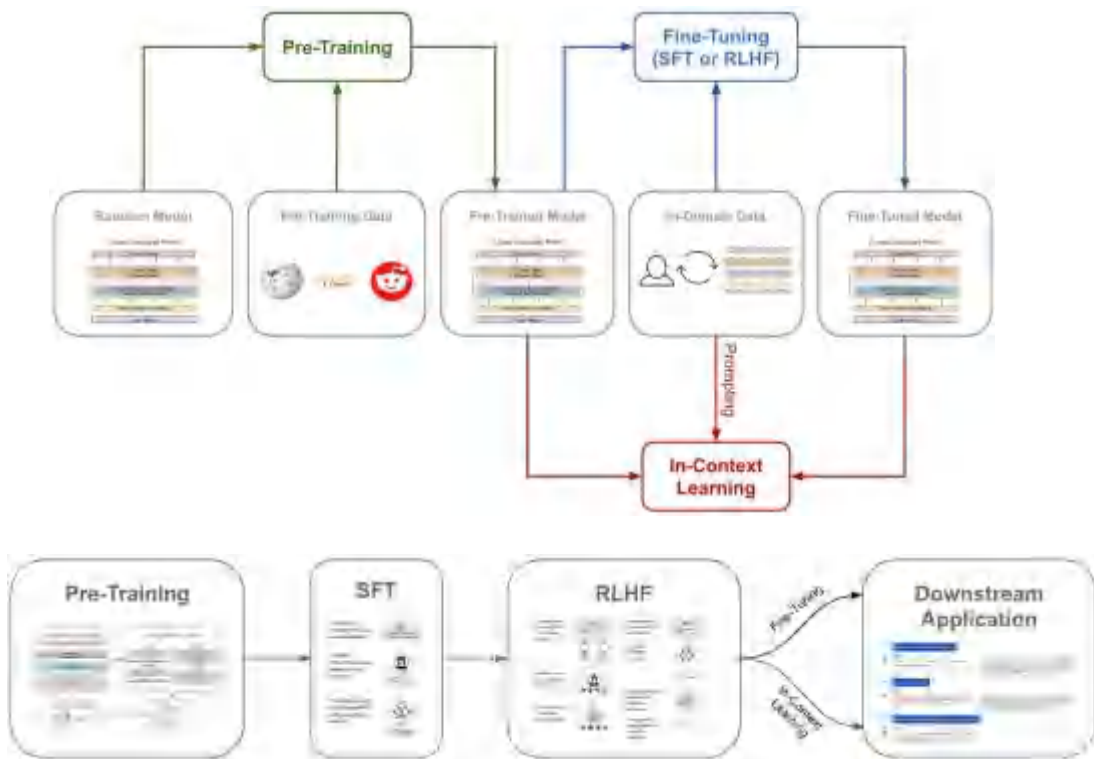
Pre-training on 104 languages



训练完发现针对特定领域不得劲怎么办？

■ 后训练对齐

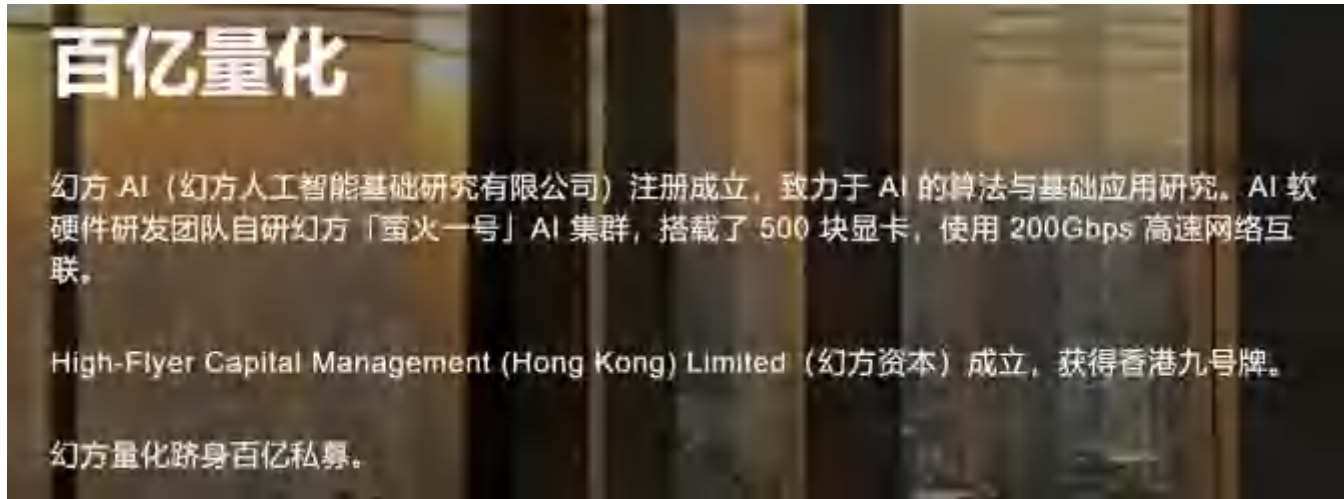
- 将预训练后的模型进一步对齐数据，防止“胡说八道”
- 通常有SFT（监督式微调）和RL（强化学习，约等于给个指导自己学习）



DeepSeek模型的发展历程

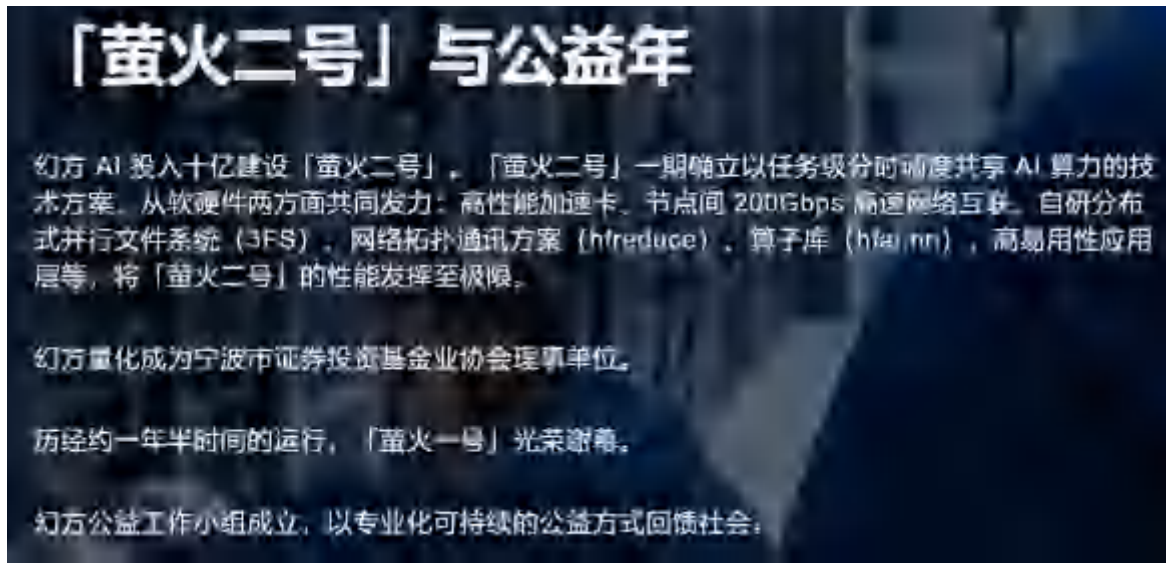
幻方公司早期对AI的投入

2019



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



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_{layer}	n_{head}	n_{head}	n_{head}	Context Length	Sequence Batch Size	Learning Rate	Tokens
7B	32	1024	32	12	1096	256	$4.3e-4$	2.0T
67B	95	1024	64	9	1096	1024	$3.3e-4$	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 attention design of DeepSeek LLM largely follows the design of LLaMA (Touvron et al., 2023a,b), adapting 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 $2d_{\text{model}}$. 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., 2022) instead of the traditional Multi-Head Attention (MHA).

However, in terms of network 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 circles (intermediate checkpoints of the base model from its training process. Please refer to the terms outlined in [License section](#). Commercial usage is per [License section](#).

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

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

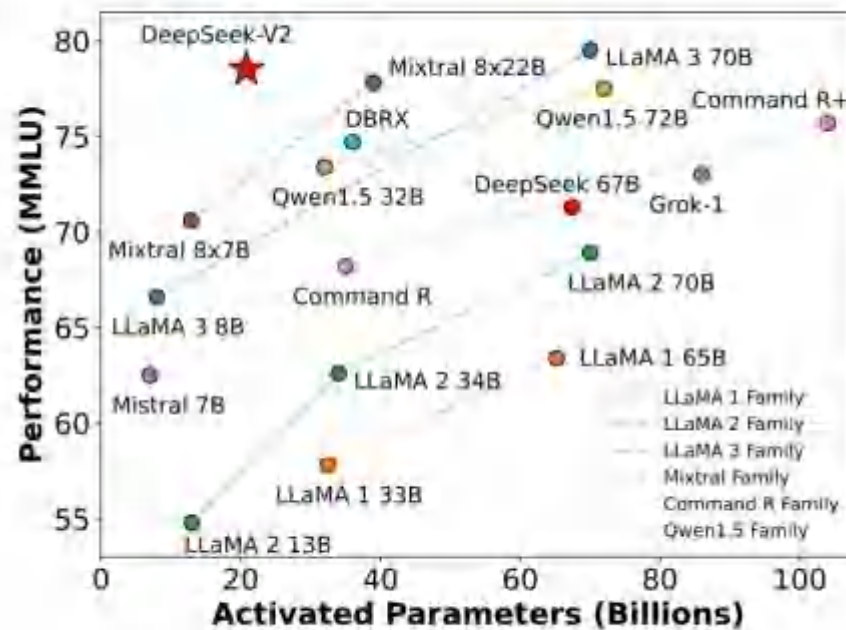
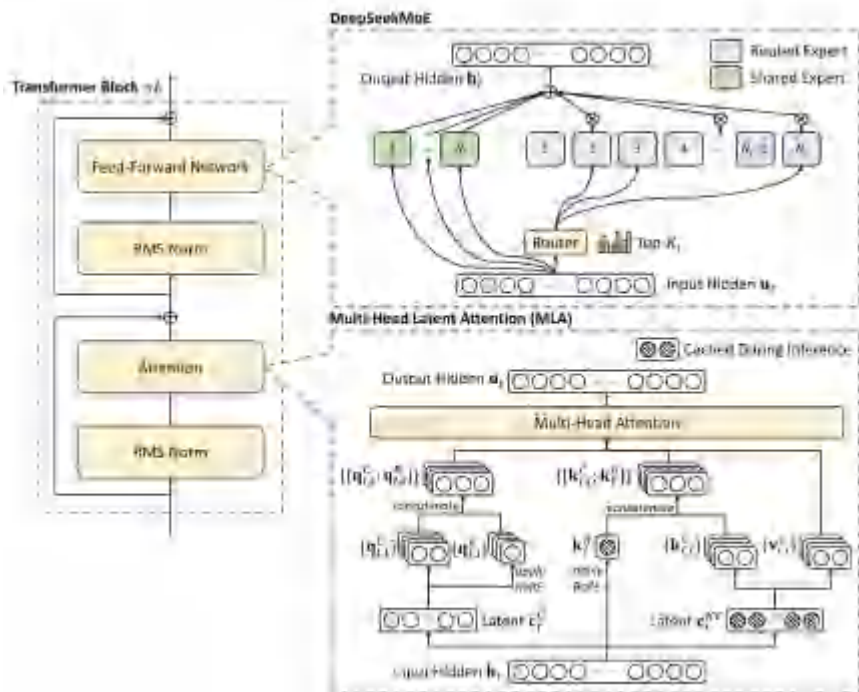
Approach	Coeff. α where $N_{\text{opt}} \propto M_{\text{opt}}^\alpha$	Coeff. β where $D_{\text{opt}} \propto C^\beta$
OpenAI (OpenWebText2)	0.33	0.27
Churchilla (Massa et al.)	0.49	0.51
Ours (Early Data)	0.450	0.360
Ours (Current Data)	0.624	0.475
Ours (OpenWebText2)	0.578	0.423

严谨地研究scaling law,
敢于质疑成名结论

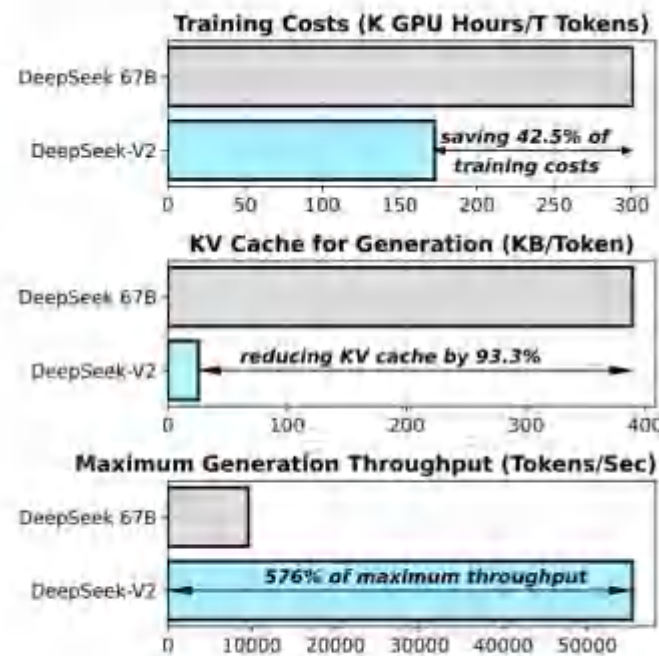
初露峥嵘：开放基因，严谨思维

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)



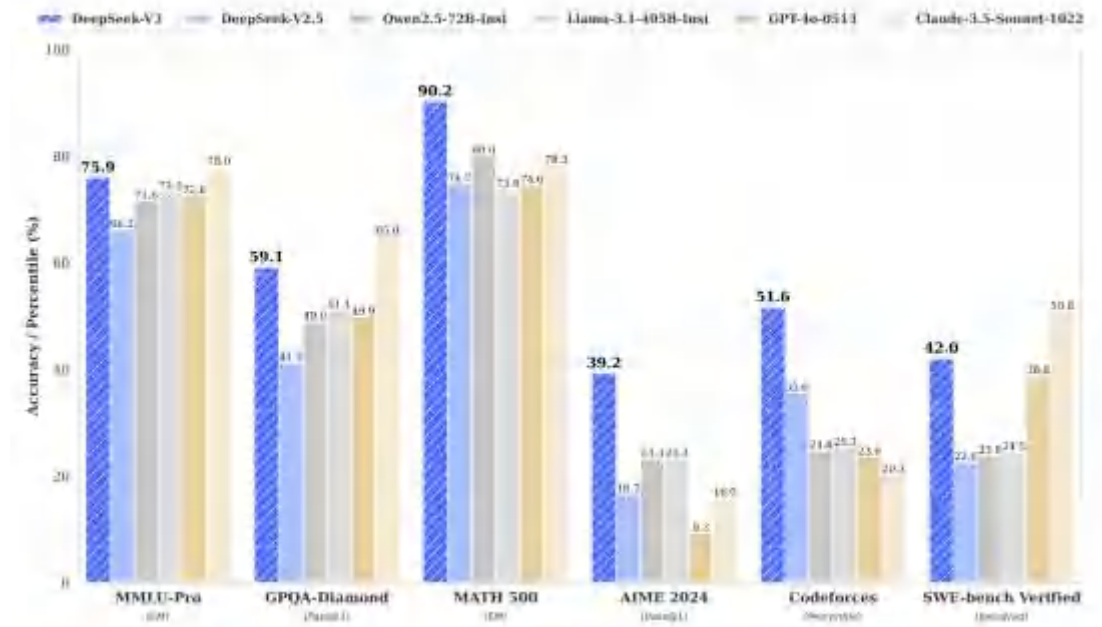
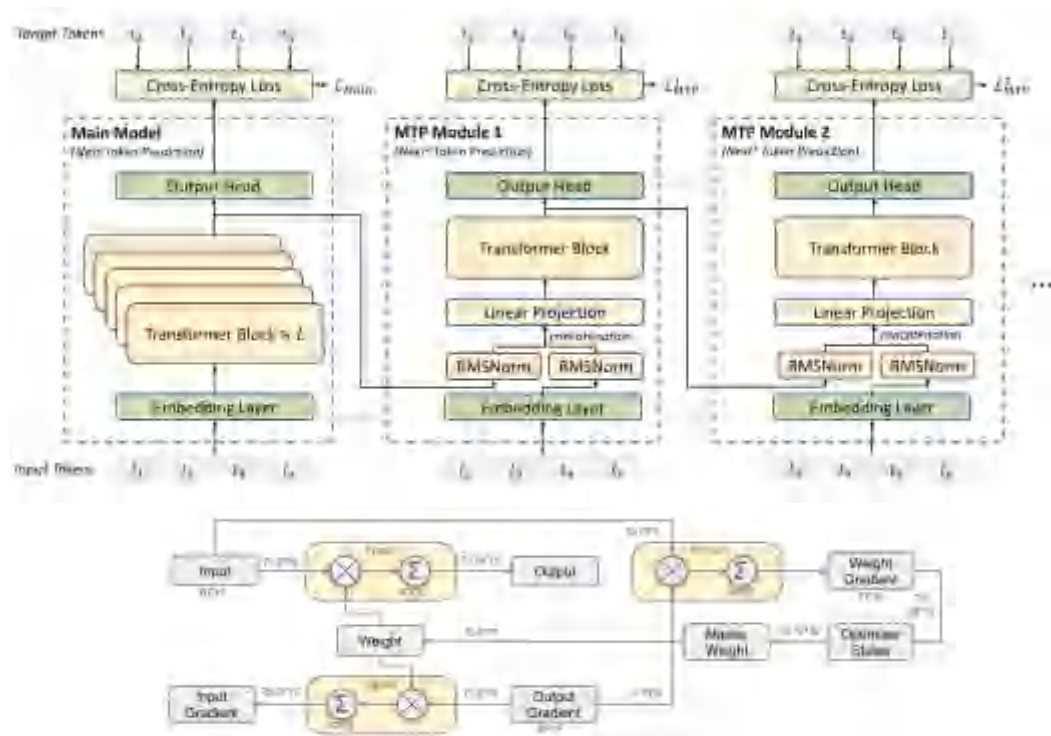
(b)

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

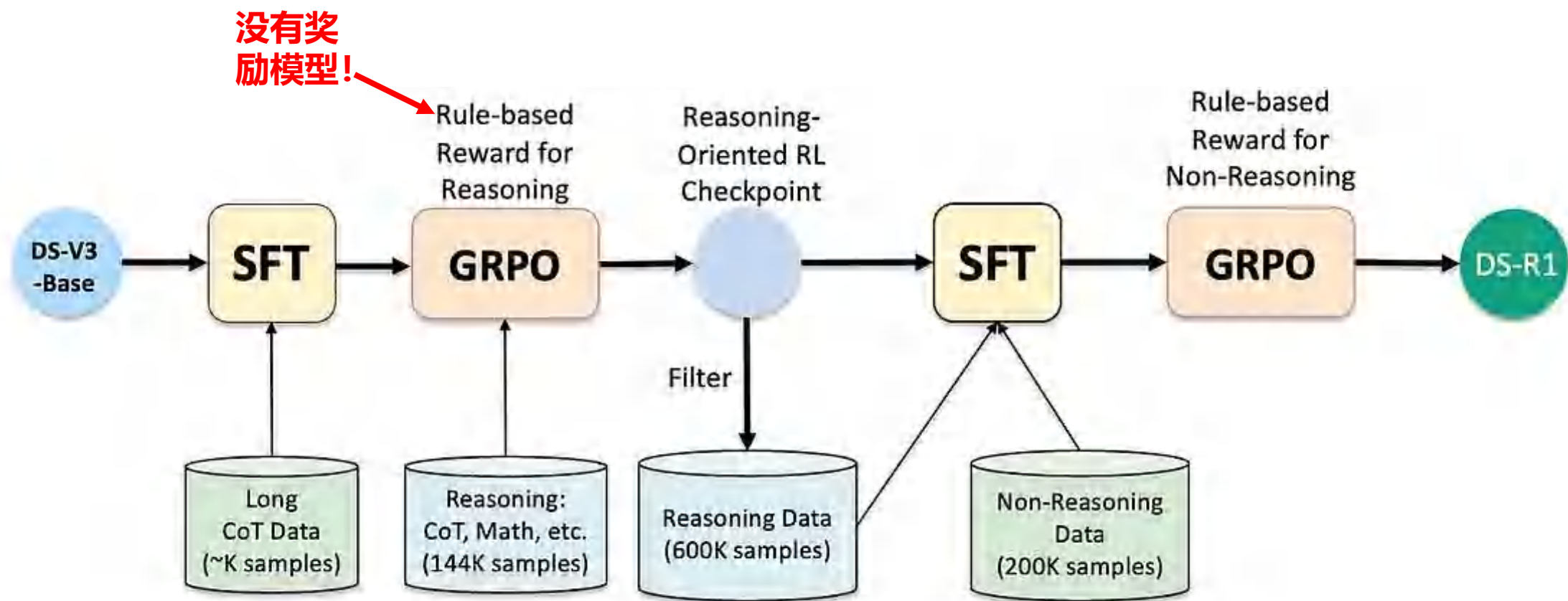
MTP, FP8训练, 继续增大MoE专家数量。。。

进入TOP梯队

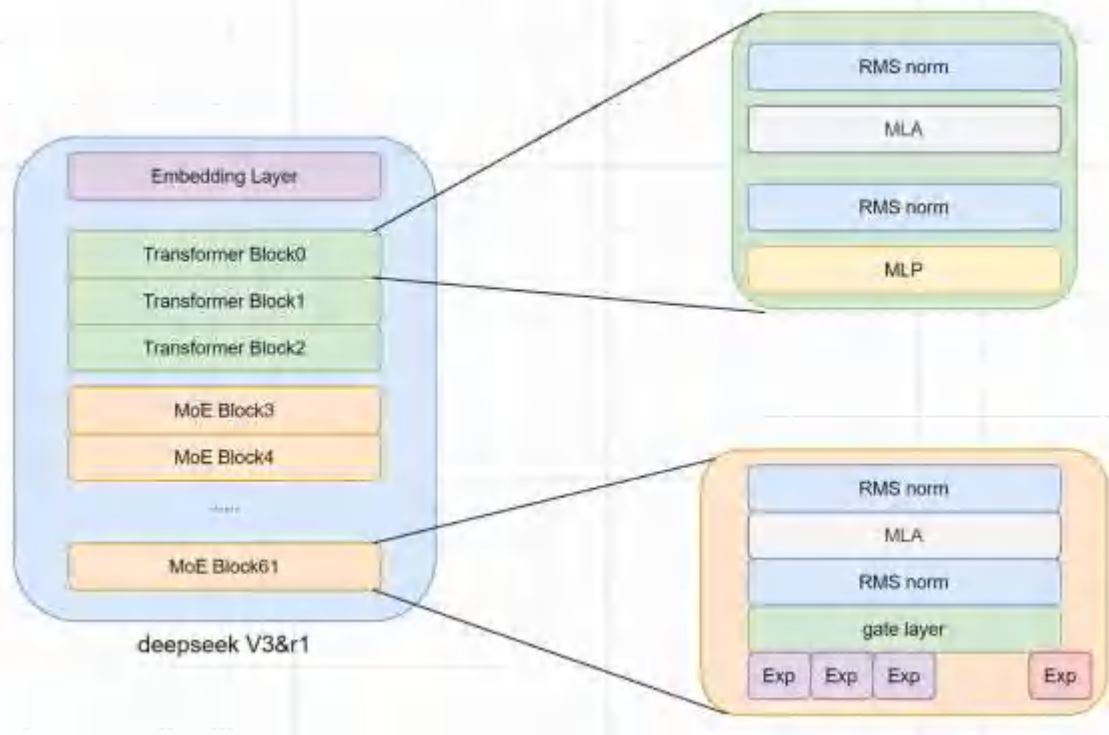
各项指标达到世界第一

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

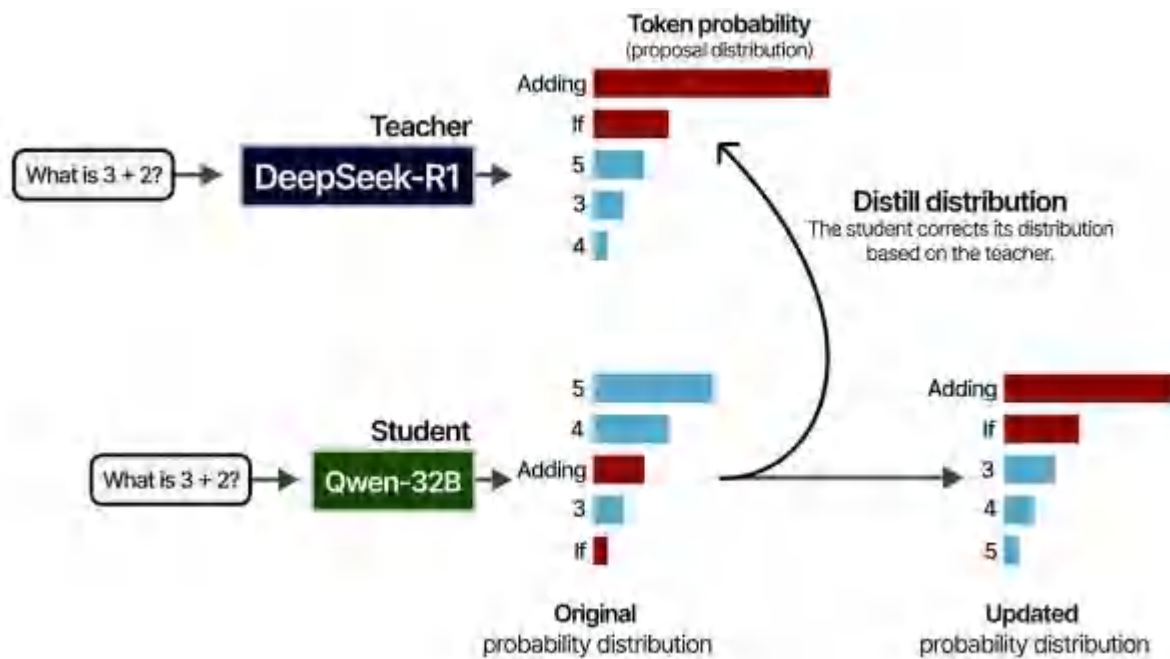


DeepSeek R1的模型结构



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

提纲

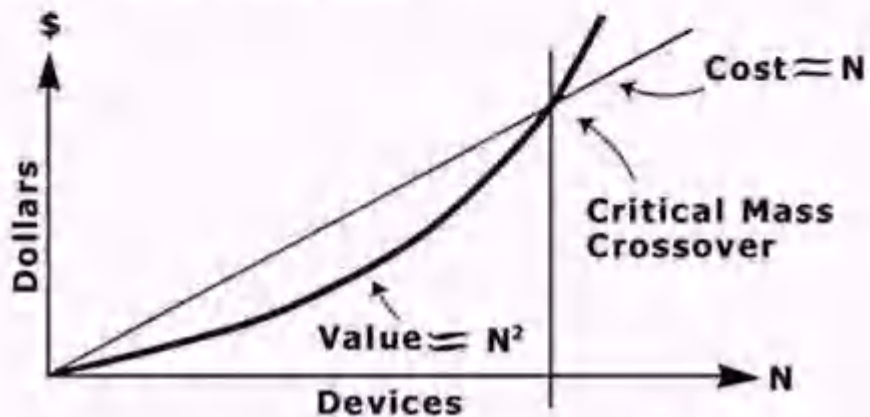
- What is it: DeepSeek是什么
 - 从ChatGPT到DeepSeek-R1, TA到底厉害在哪里?
 - DeepSeek基本概念 (用户角度)
- How to use it: 我能用DeepSeek干什么
 - 以小见大, 掌握思维方法
 - 正确理解, 打开广阔天地
- Why it works: DeepSeek背后的原理
 - Transformer——大模型基础
 - DeepSeek模型的发展历程
- **Next: 下一步要关注什么**
 - **生态的爆发就在眼前, 整个链条上哪些方面值得关注**

为什么我认为生态马上会有真正的爆发?

Metcalfe's Law:

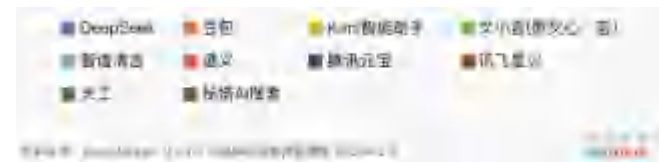
$$v=A*N^2$$

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。。。

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

■ 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

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

■ 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

个人浅见，仅供参考



- 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

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

- 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！





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

硬件资源 HARDWARE	中心拥有“天河”系列超级计算机、“天河·去马”计算集群等多个计算平台 峰值算力达100T FLOPS (FP64), 1.3亿核时/月算力达1000T FLOPS (FP32), 峰值存储20PB
网络环境 NETWORK	中心通过“全球互联网骨干网核心节点”和“中国国家互联网骨干网核心节点” 中心网络接入国际、国内、省内、市网、教育网、电信网、移动网、联通网、广电网、有线电视网

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>