

浙江大学DS系列专题

DeepSeek技术溯源及前沿探索

主讲人：朱强

浙江大学计算机科学与技术学院

人工智能省部共建协同创新中心（浙江大学）

<https://person.zju.edu.cn/zhuq>

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

一、语言模型

二、Transformer

三、ChatGPT

四、DeepSeek

五、新一代智能体

Language Modeling

对于任意的词序列，计算出这个序列是一句话的概率

我们每天都和语言模型打交道：

I saw a cat|

I saw a cat on the chair

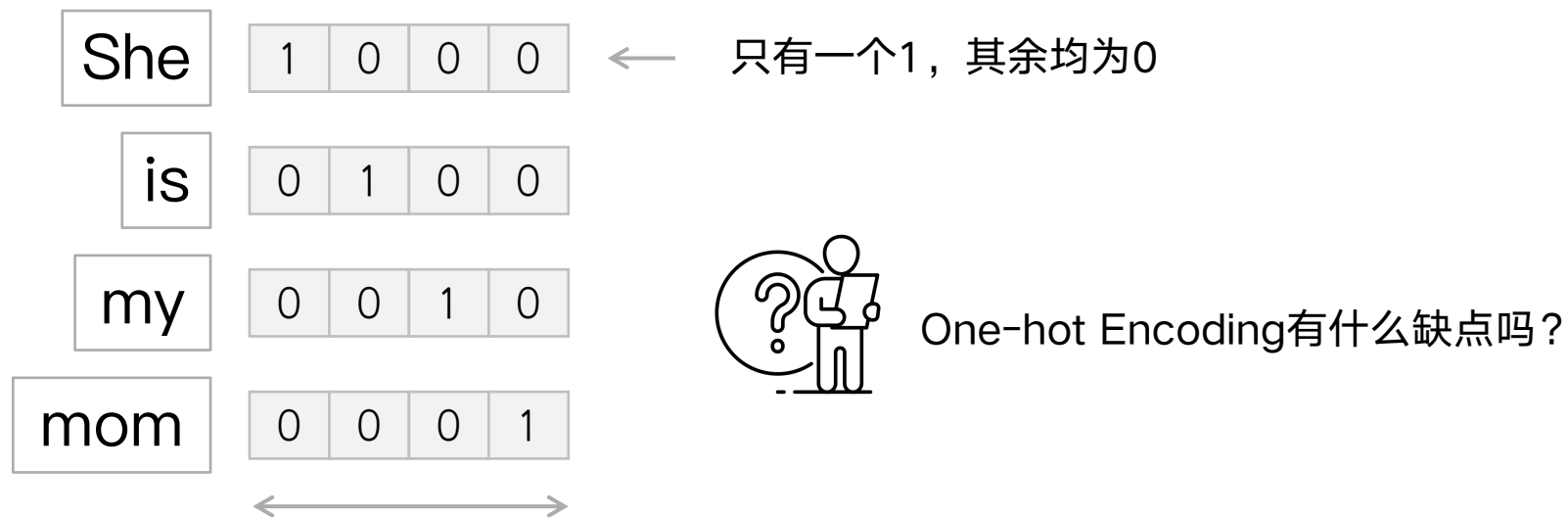
I saw a cat running after a
dog

I saw a cat in my dream

I saw a ca|

car ←

编码：让计算机理解人类语言



One-hot Encoding

Word Embedding

用一个**低维**的词向量表示一个词
能使距离**相近的向量**对应的物体有**相近的含义**

	游泳			飞翔	
鲸鱼	0.99	0.99	0.05	0.1	...
海豚	0.99	0.05	0.93	0.09	...
鹦鹉	0.02	0.01	0.99	0.98	...
企鹅	0.98	0.02	0.94	0.3	...



20维的向量用one-hot和word embedding的方法分别可以表示多少单词？

Word Embedding

A bottle of **tezgüino** is on the table.
Everyone likes **tezgüino**.
Tezgüino makes you drunk.
We make **tezgüino** out of corn.

结合句子语境我们可以猜测：

tezgüino是一种由玉米制作的酒精类饮料

- (1) A bottle of _____ is on the table.
(2) Everyone likes _____.
(3) _____ makes you drunk.
(4) We make _____ out of corn.



		(1)	(2)	(3)	(4)
	tezgüino	1	1	1	1
	motor oil	1	0	0	0
	tortillas	0	1	0	1
	wine	1	1	1	0

两行内容十分相近



两个单词含义相近

语言模型：技术演化

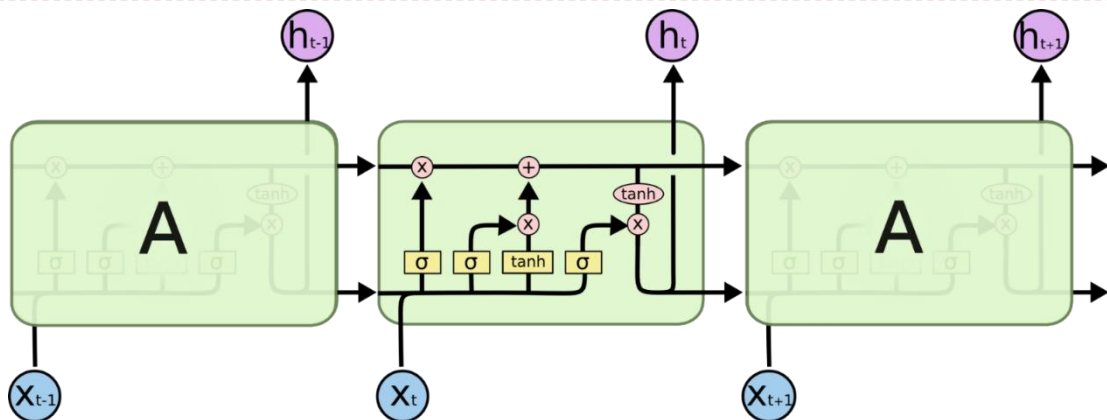
■ 基于统计的N-gram (1970 after)

Before: $P(\text{小}) \cdot P(\text{猫}|\text{小}) \cdot P(\text{抓}|\text{小猫}) \cdot P(\text{老}|\text{小猫抓}) \cdot P(\text{鼠}|\text{小猫抓老})$

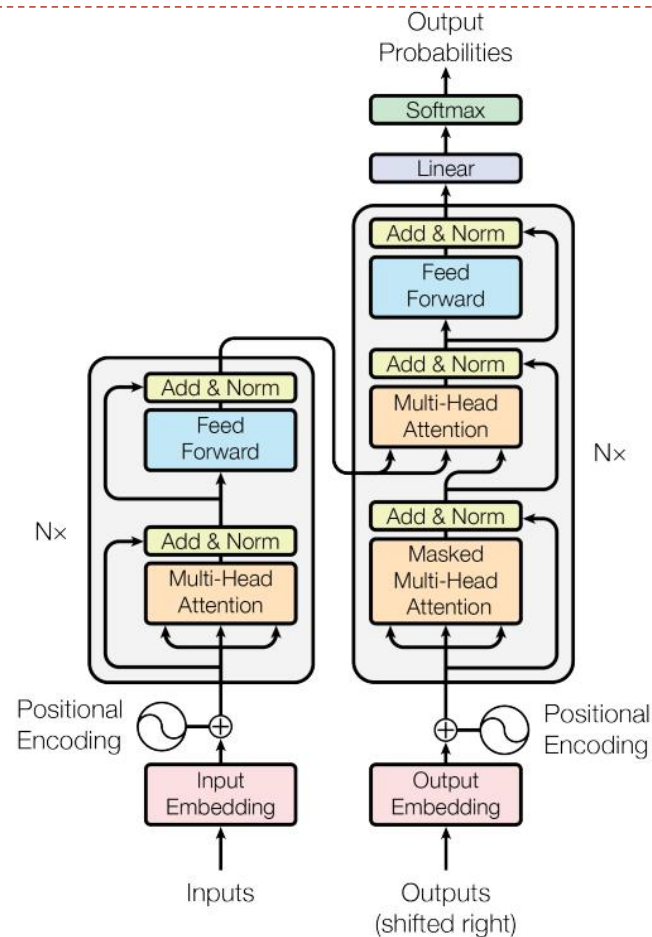
2-gram: $P(\text{小}) \cdot P(\text{猫}|\text{小}) \cdot P(\text{抓}|\text{猫}) \cdot P(\text{老}|\text{抓}) \cdot P(\text{鼠}|\text{老})$

3-gram: $P(\text{小}) \cdot P(\text{猫}|\text{小}) \cdot P(\text{抓}|\text{小猫}) \cdot P(\text{老}|\text{猫抓}) \cdot P(\text{鼠}|\text{猫抓老})$

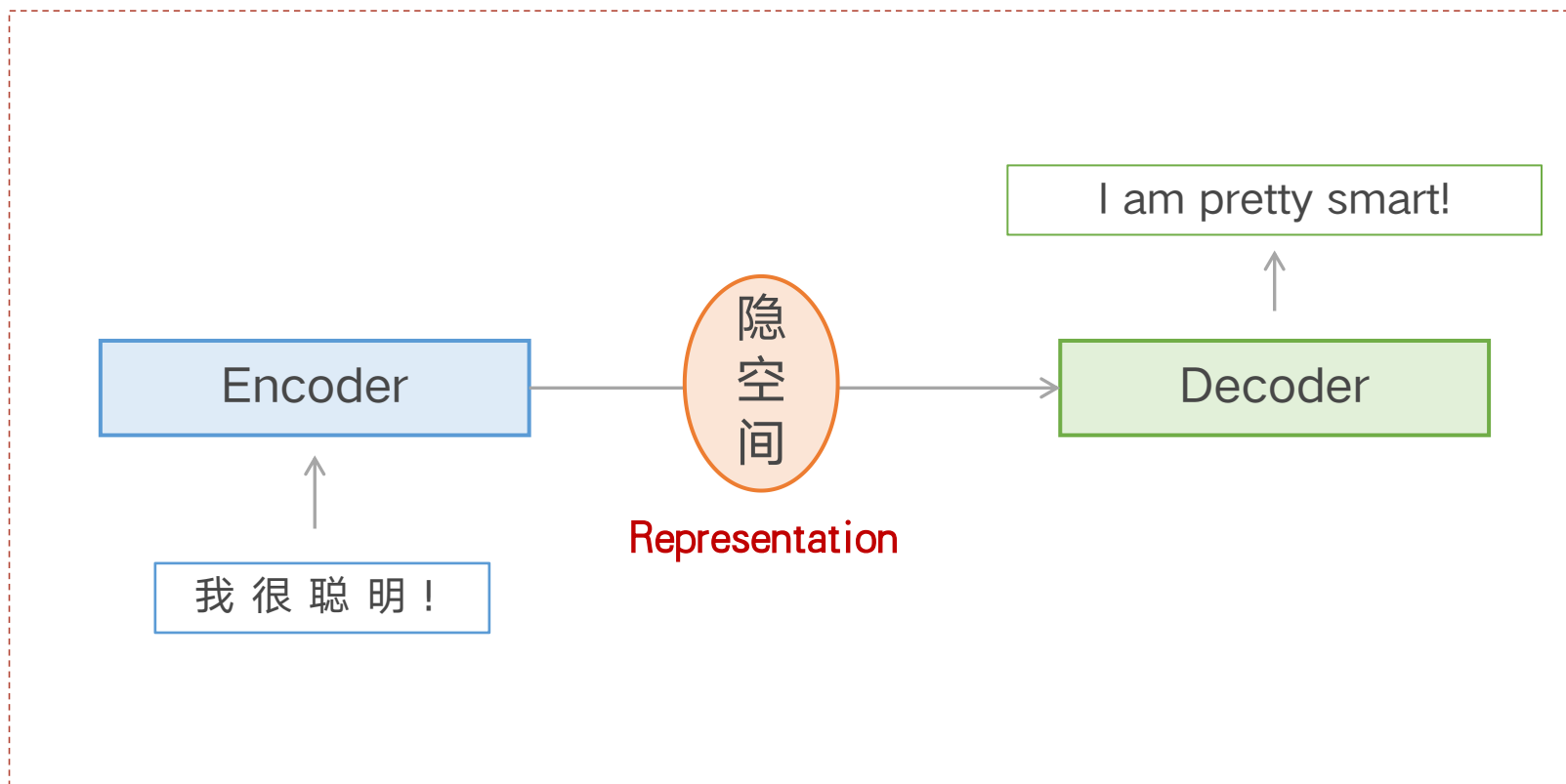
■ 基于神经网络的LSTM/GRU (2000 after)



■ Transformer (2017 after)



常见的深度学习模型框架，可用于解决 Seq2Seq 问题



可以根据任务选择不同的编码器和解码器（LSTM/GRU/Transformer）

一、语言模型

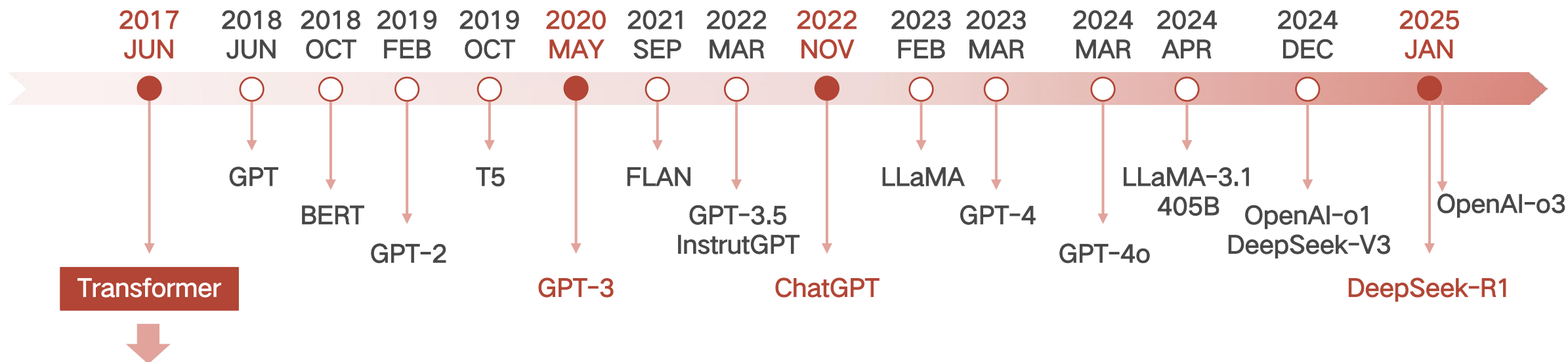
二、Transformer

三、ChatGPT

四、DeepSeek

五、新一代智能体

大型语言模型简史



Transformer: 理论架构创新

- 自注意力机制：支持并行计算/全局上下文的理解能力
- 多头注意力：从多个角度捕捉复杂的语义关系
- 前馈网络/位置编码/层归一化：解决了传统模型的诸多局限性

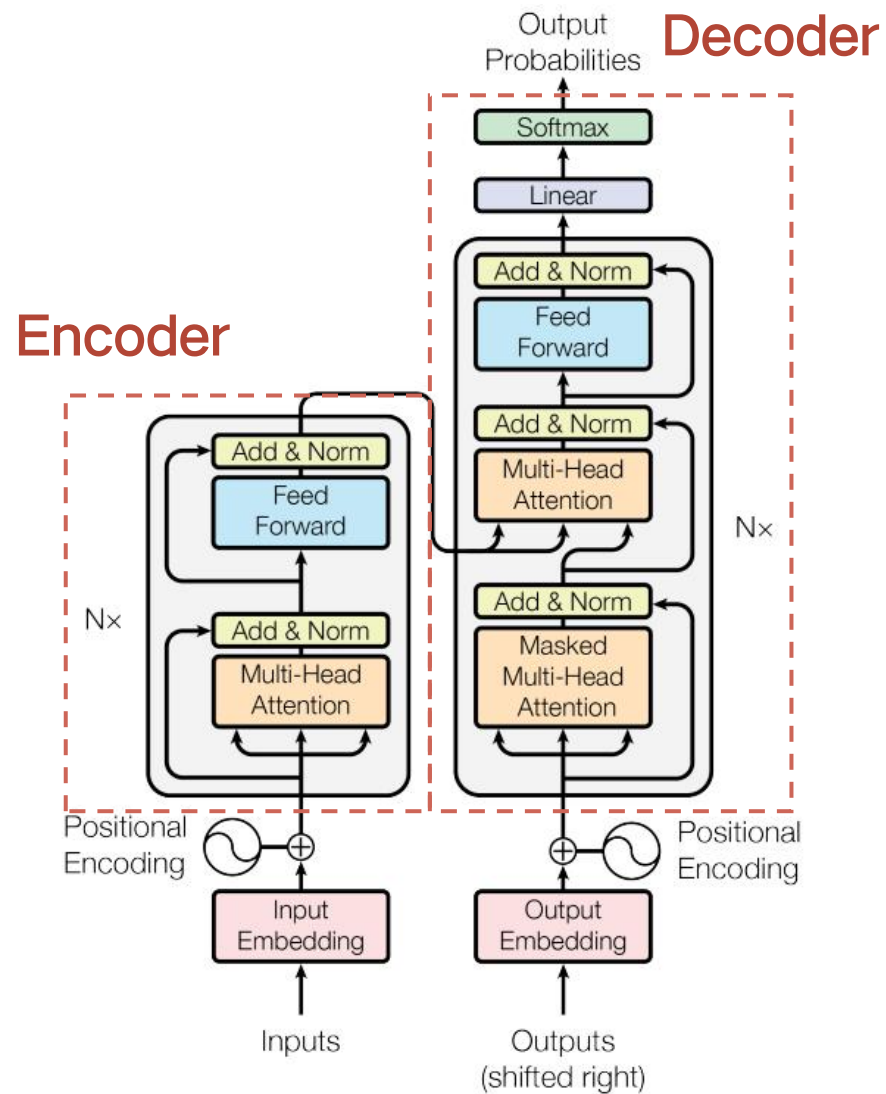
Transformer：大模型的技术基座

Attention Is All You Need

NIPS 2017, 引用量**15万+**

引入全新**注意力机制**，改变了深度学习模型的处理方式

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



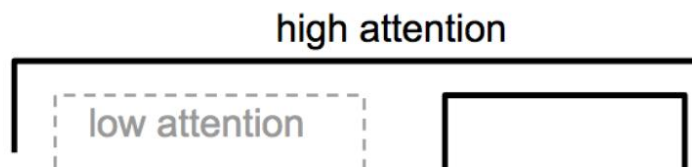
Transformer：（自）注意力机制

在理解语言任务时，Attention 机制本质上是捕捉单词间的关系

1 中国 南北 饮食文化 存在差异，豆花有 南甜北咸 之分。南方人 一般 喜欢 吃 甜豆花



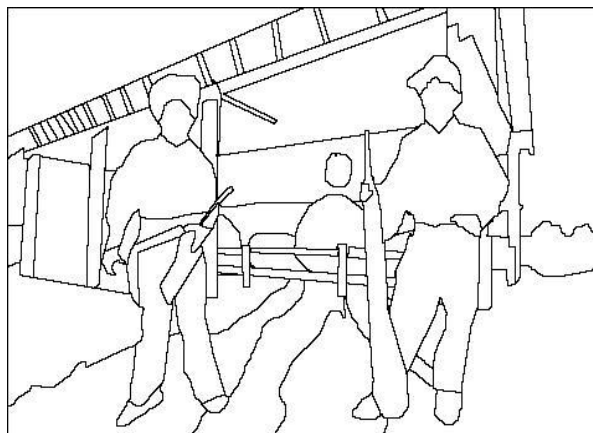
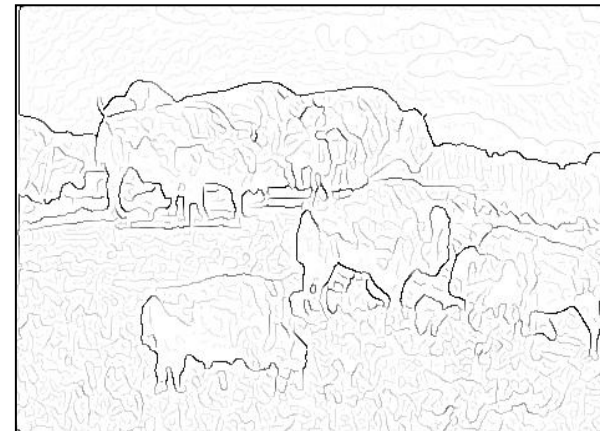
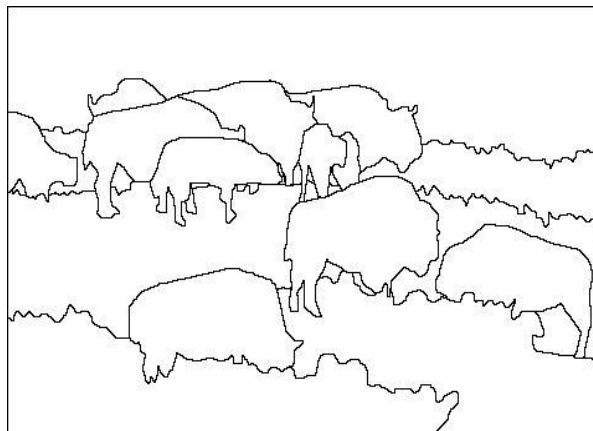
2 She is eating a green apple.



3 The animal didn't cross the street because it was too **tired/wide**

Transformer：（自）注意力机制

在理解**图像任务**时，Attention机制本质上是一种**图像特征抽取**



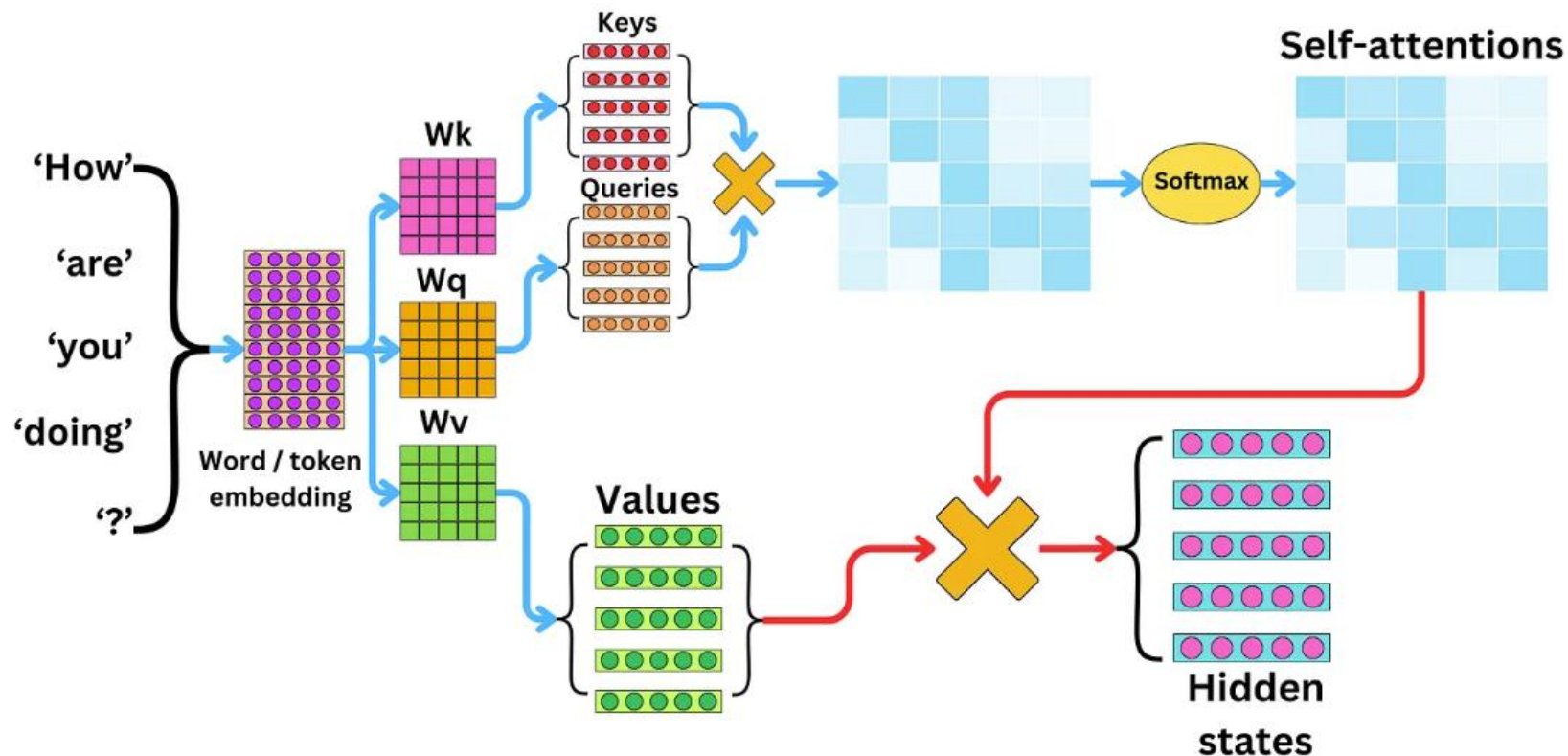
Image

Sketch

Gradient

Transformer：训练机制

场景：你在图书馆想找一本关于“机器学习基础”的书



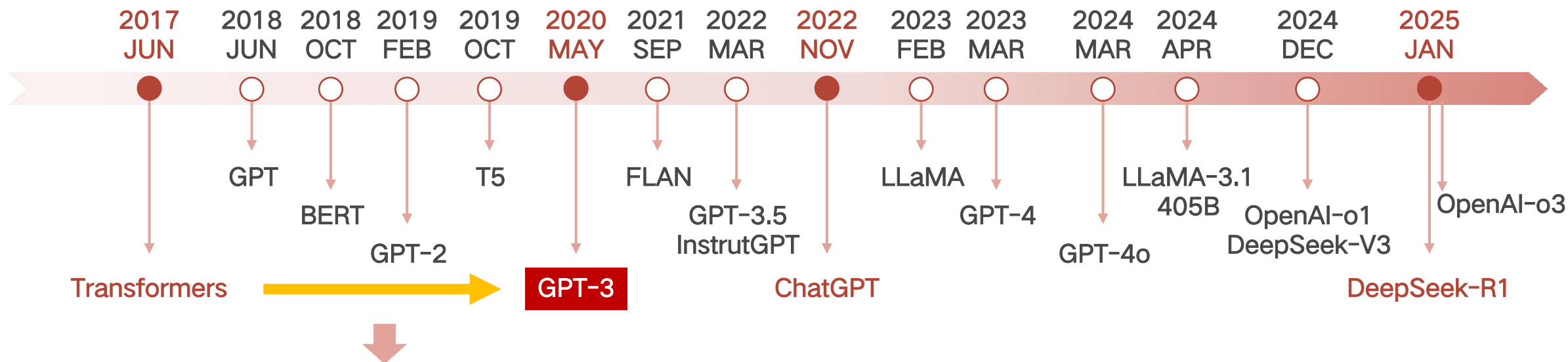
Query： 描述要找的书（精准的需求描述）

Key： 书的索引编号（高效的书籍定位）

Value： 内容的抽取（由目标任务驱动）

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

大型语言模型简史



预训练时代：大力出奇迹（“暴力美学”）

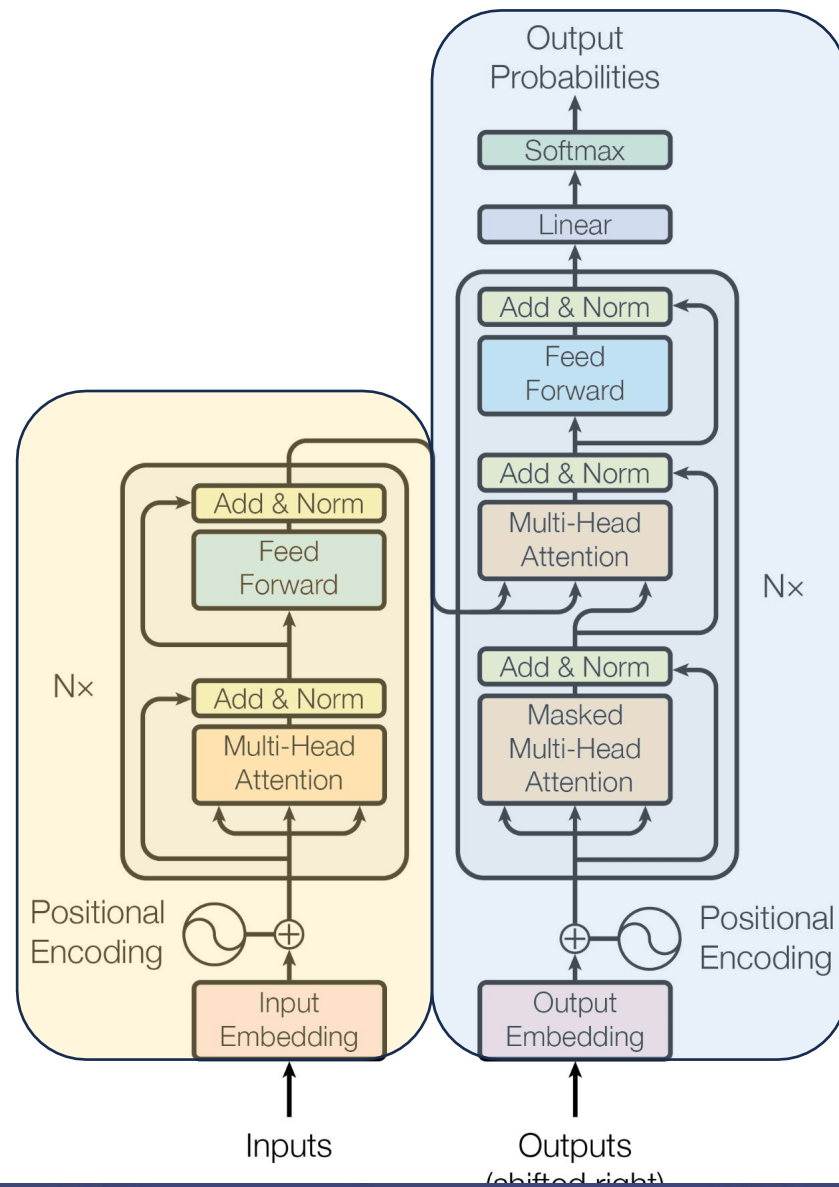
- BERT: Bidirectional Encoder Representations Transformers
- GPT: Generative Pertained Transformer
- 自监督算法: MLM/NTP/MAE解决海量数据标注问题



The LLM Era – Paradigm Shift in Machine Learning

BERT
Oct
2018

Representation



GPT
Jun
2018

Generation



The LLM Era – Paradigm Shift in Machine Learning

BERT – 2018
DistilBERT – 2019
RoBERTa – 2019
ALBERT – 2019
ELECTRA – 2020
Representation
2020
...



GPT – 2018
GPT-2 – 2019
GPT-3 – 2020
GPT-Neo – 2021
GPT-3.5 (ChatGPT) – 2022
LLaMA – 2023
GPT-4 – 2023
...
Generation

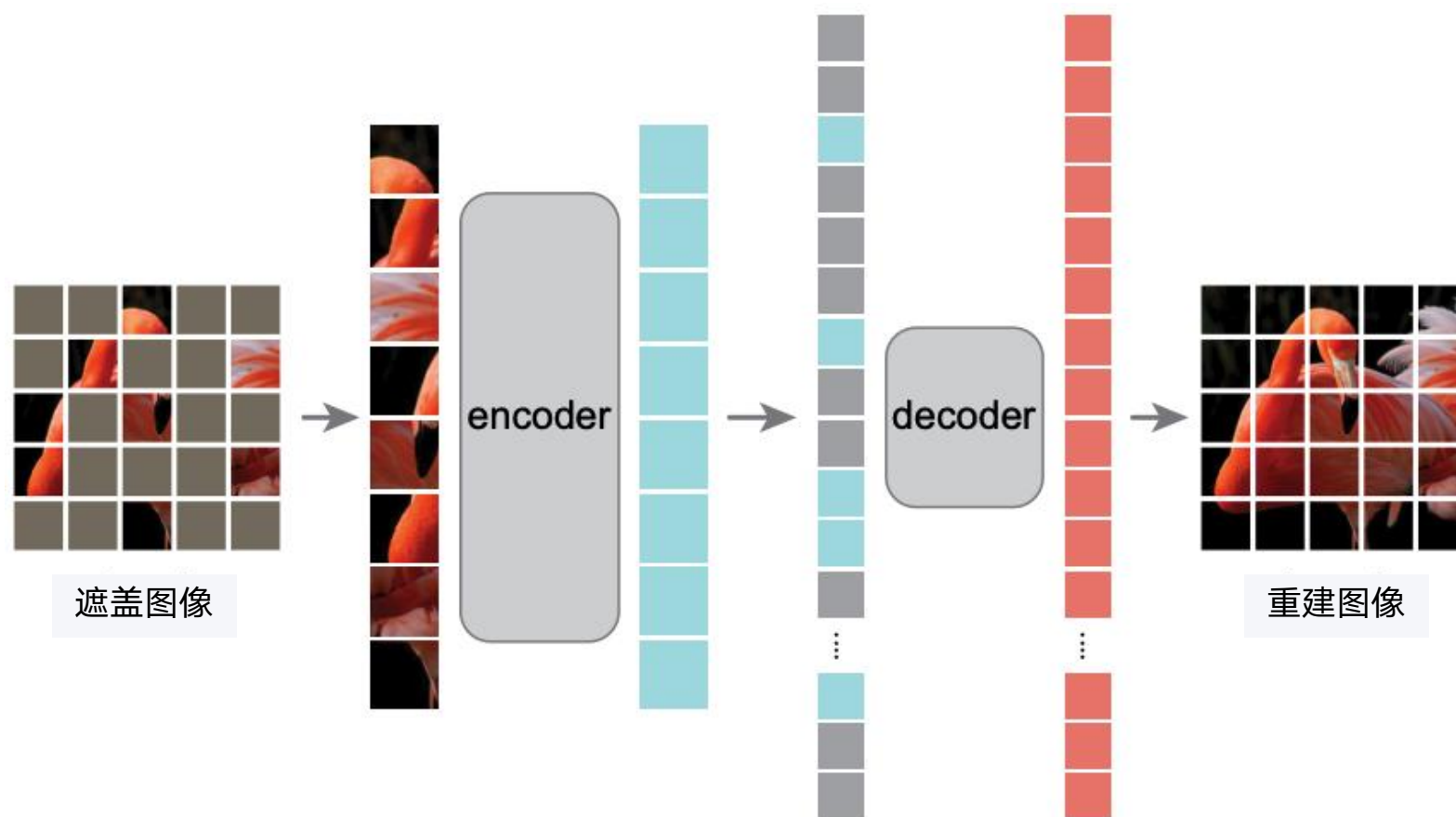
Masked Language Modeling (MLM) 模型会不断地在句子中‘挖去’一个单词，根据剩下单词的上下文来填空，即预测最合适的‘填空词’出现的概率，这一过程为‘自监督学习’

原话：一辆 列车 缓慢 行驶 在 崎岖 的 山路上

移除单词：一辆 列车 _____ 行驶 在 崎岖 的 山路上

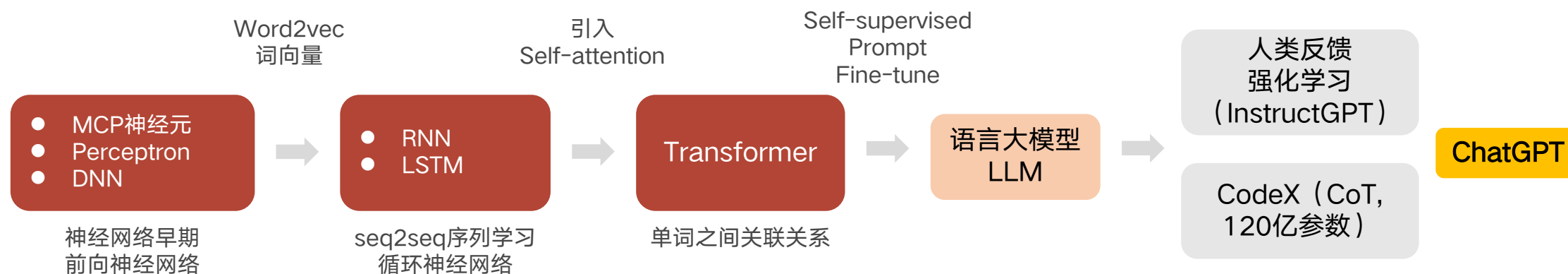
预测填空：一辆 列车 缓慢 行驶 在 崎岖 的 山路上

Masked AutoEncoders (MAE) 通过随机遮盖部分输入数据（如图像）并重建缺失内容，让模型从上下文中学到图像的深层特征，常用于计算机视觉任务。



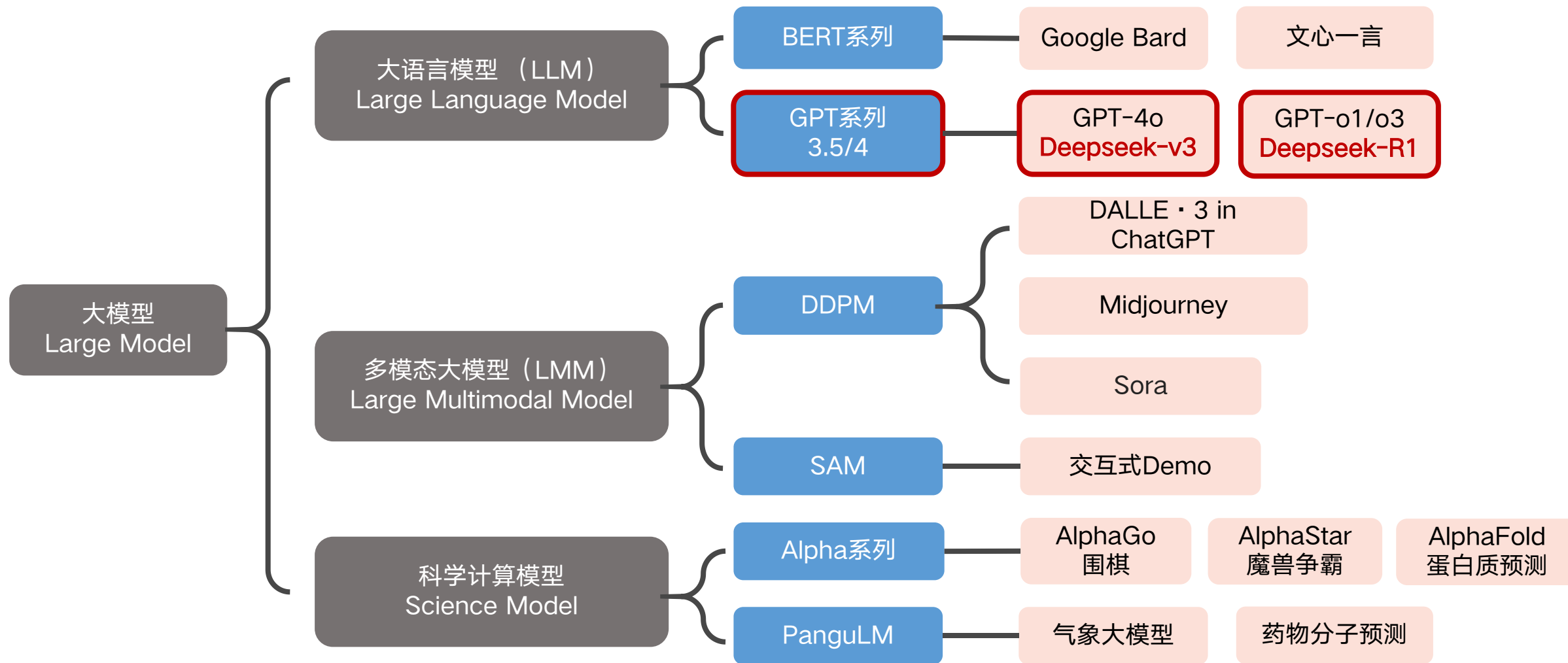
训练 transformer 的通用之力

数据是**燃料**、模型是**引擎**、算力是**加速器**

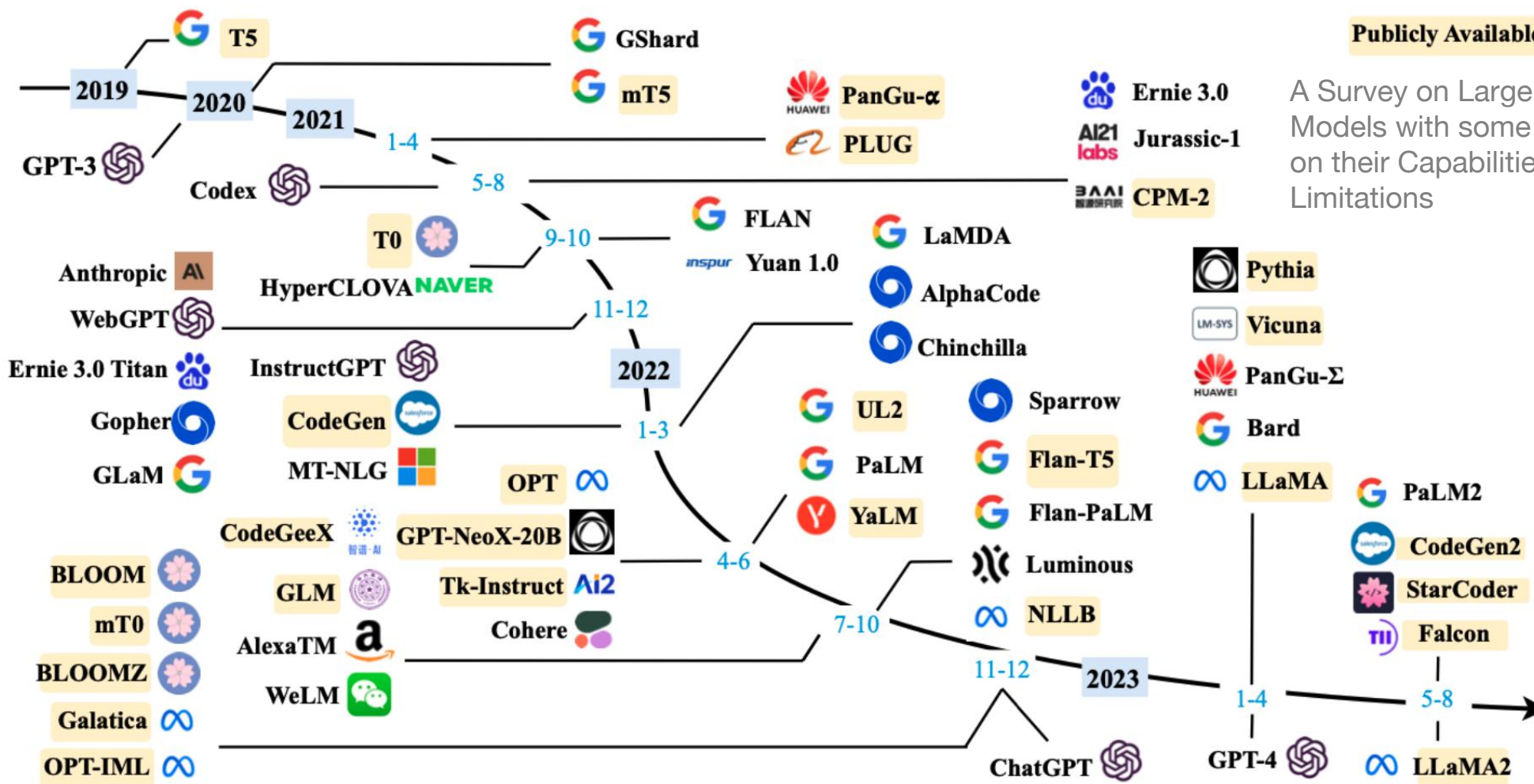


- **数据**: 训练中使用了45TB数据、近 1 万亿个单词（约1351万本牛津词典所包含单词数量）以及数十亿行源代码。
- **模型**: 包含了1750亿参数，将这些参数全部打印在A4纸张上，一张一张叠加后，叠加高度将超过上海中心大厦632米高度。
- **算力**: ChatGPT的训练门槛是1万张英伟达V100芯片、约10亿人民币。
- **大数据、大模型、大算力**下以“共生则关联”原则实现了统计关联关系的挖掘。

大模型脉络



群雄（中美）争霸



OpenAI最新15页报告: DeepSeek缩小中美AI差距

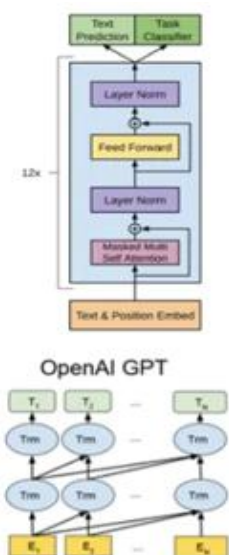
闭源 vs 开源

国际企业	微软	投资OpenAI的GPT-4.0系列	闭源
		自研开源小模型Phi-3 Mini	开源
	亚马逊	自研Titan系列	闭源
		投资Anthropic的Claude 3.5系列	闭源
	谷歌	Gemini系列	闭源
		Gemma系列	开源
	META	Llama3系列	开源
	Mistral AI	Mistral-Large	闭源
		Mistral-Medium	开源
中国企业	阿里	通义千问2.5系列基础模型、行业模型	开源
		Qwen 0.5b-110b系列开源模型	开源
	华为	盘古系列	闭源
	腾讯	混元基础模型、行业模型	闭源
		混元开源模型	开源
	百度	文心一言4.0模型	闭源

： 摩尔定律（大模型时代）

GPT-1 (2018)

12层，每层12个注意力



GPT-2 (2019)

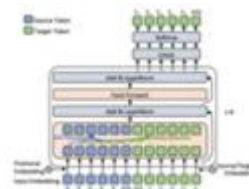
GPT-2做了以下改进：

1. 增加到48层，使用1600维向量进行词嵌入；
2. 将层归一化移动到每个子块的输入，并在最终的自注意块后增加一层归一化；
3. 修改初始化的残差层权重，缩放为原来的 $1/\sqrt{N}$ ，其中， N 是残差层的数量；
4. 特征向量维数从768扩展到1600，词表扩大到50257。

GPT-3 (2020)

GPT-3做了以下优化：

1. 增加到96层，每层有96个注意力；
2. 单词嵌入大小从1600增加到12888；
3. 上下文窗口大小从GPT-2的1024增加到2048，并采用交替密度和局部带状稀疏注意模式。



ChatGPT (2022)

ChatGPT基于GPT-3.5：

1. ChatGPT使用来自人类反馈的强化学习进行训练；
2. 通过近端策略优化算法进行微调，为信任域策略优化算法带来成本效益。



模型	发布时间	参数量	预训练数据量
GPT-1	2018年6月	1.17亿	约5GB
GPT-2	2019年2月	15亿	40G
GPT-3	2020年5月	1750亿	45TB
ChatGPT	2022年11月	千亿级？	百T级？

DeepSeek通过大幅提升模型训练推理效率，缓解AI算力需求？

更多免费行业报告

并购家

www.ipoiipo.cn

一、语言模型

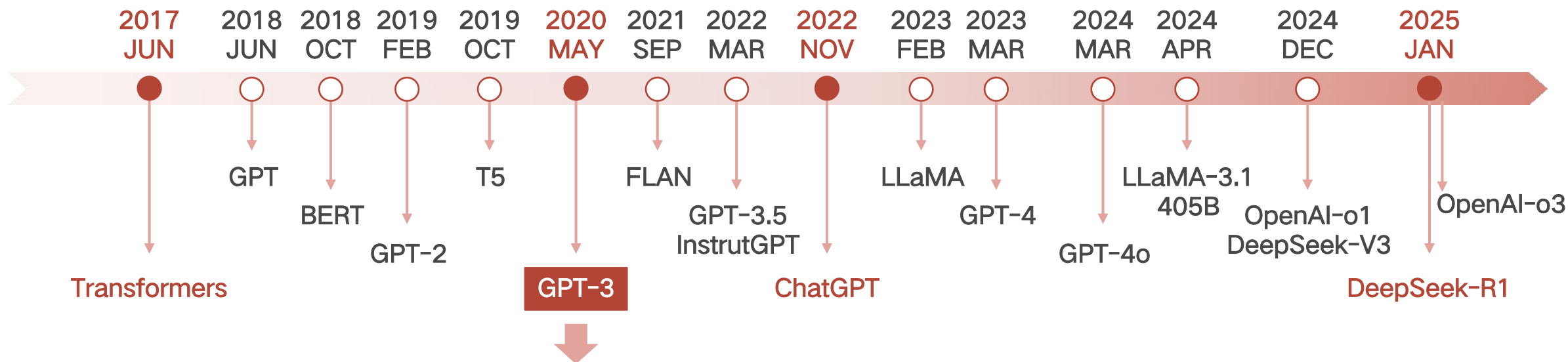
二、Transformer

三、ChatGPT

四、DeepSeek

五、新一代智能体

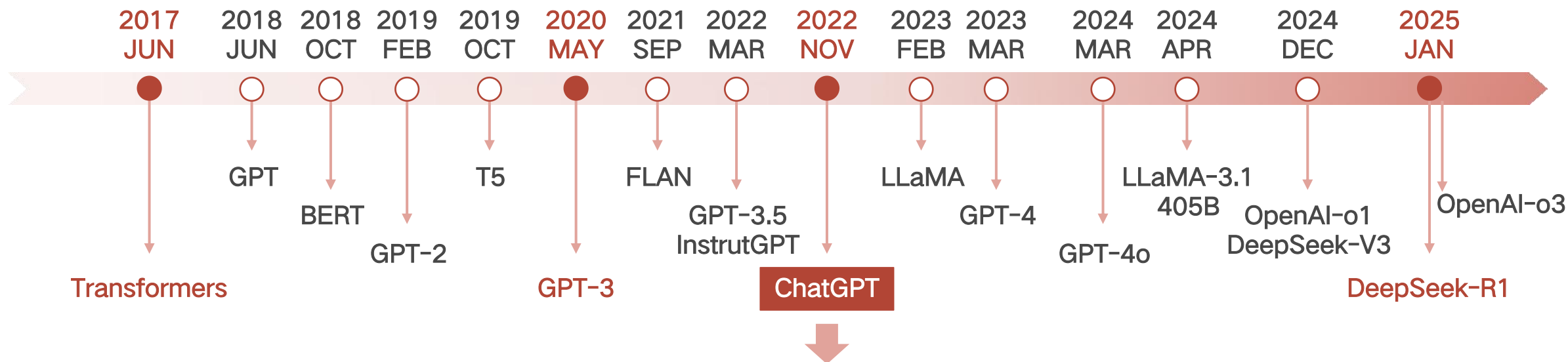
大型语言模型简史



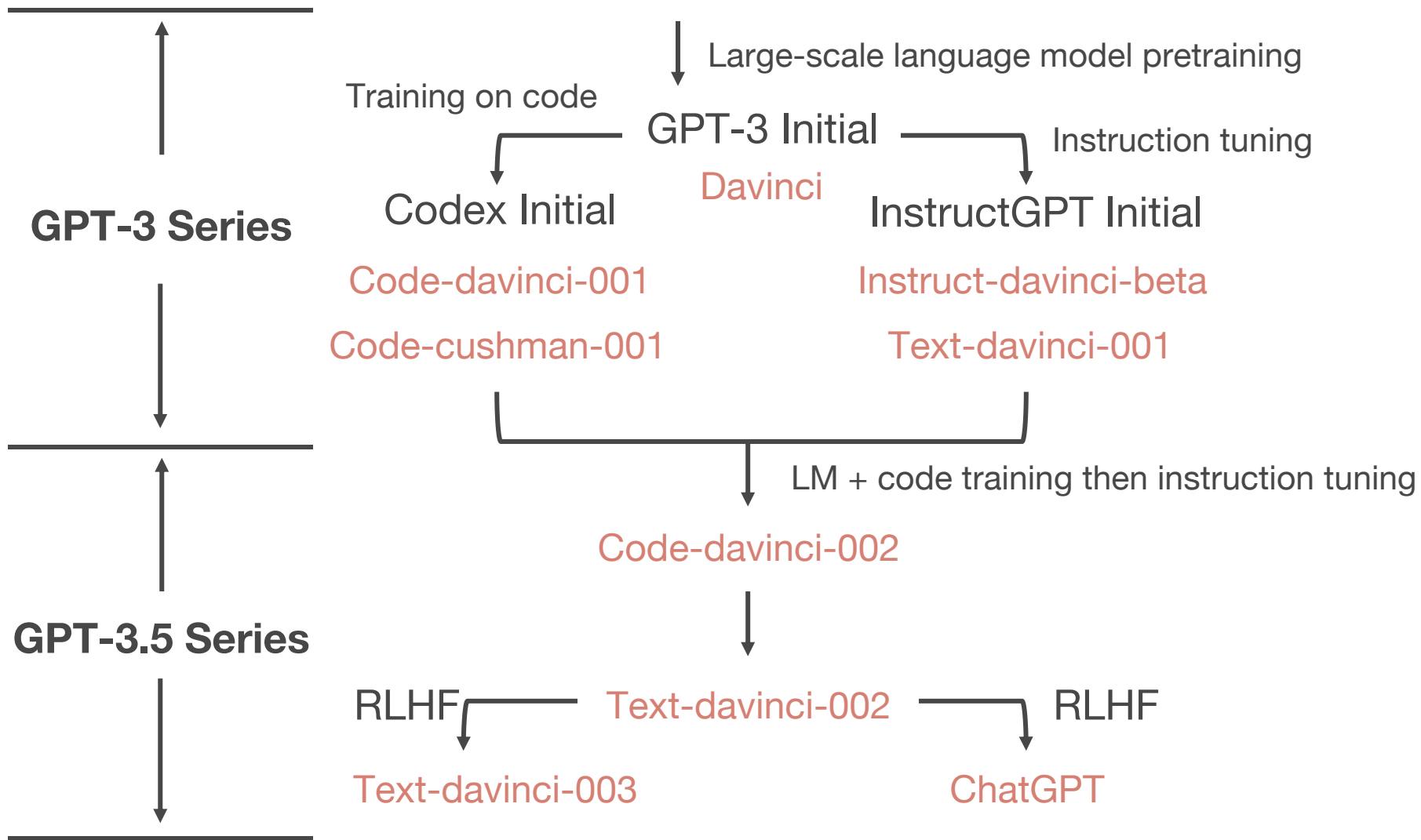
GPT-3：语言模型的转折点

- 大语言模型：1750亿参数
- **涌现**能力：随着模型规模增大而出现的新能力
- 生成/创造：**Art**ificial Intelligence（人工 => 艺术）

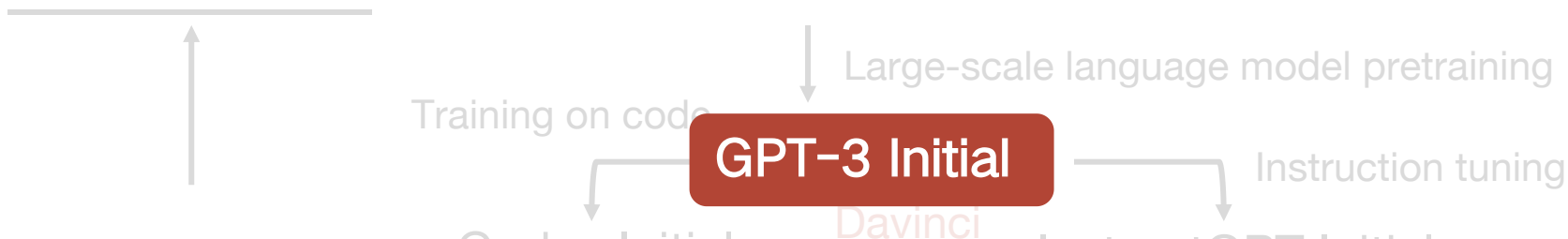
大型语言模型简史



ChatGPT: 人工智能的IPHONE时刻



■ GPT3 Initial



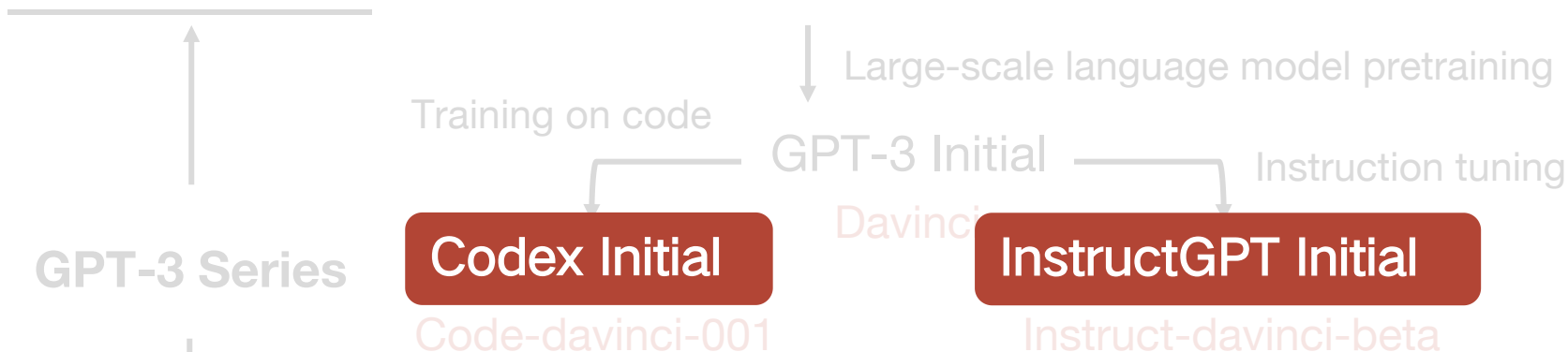
初代 GPT-3 展示了三个重要能力（来自于大规模的预训练）

- 语言生成：来自语言建模的训练目标（**说人话**）
- 世界知识：来自 3000 亿单词的训练语料库（**百晓生**）
- 上下文学习：上下文学习可以泛化，仍然难以溯源（**触类旁通**）

初代 GPT-3 表面看起来很弱，但有非常强的潜力，展示出极为强大的“涌现”能力



Codex + Instruct



2020 - 2021 年，OpenAI 投入了大量的精力通过**代码训练**和**指令微调**来增强 GPT-3。

使用**思维链**进行复杂推理的能力很可能是代码训练的一个神奇副产物
使用**指令微调**将 GPT-3.5 的分化到不同的技能树（数学家/程序员/...）



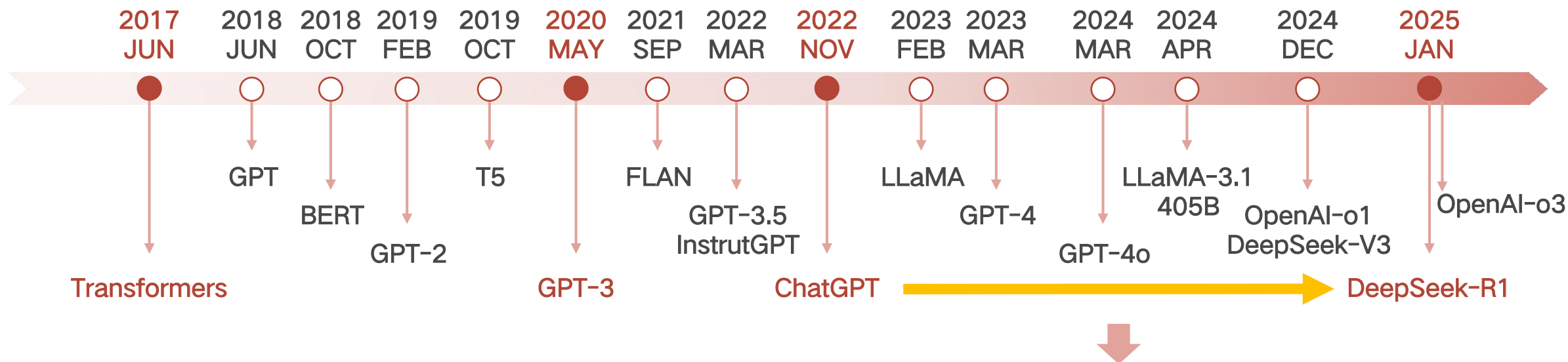
GPT3.5



ChatGPT（技术到产品）



大型语言模型简史

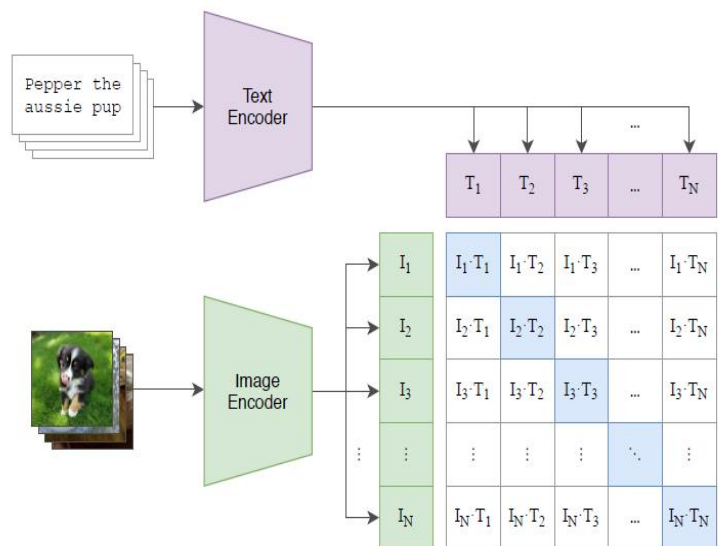


多模态模型：连接文本、图像及其他

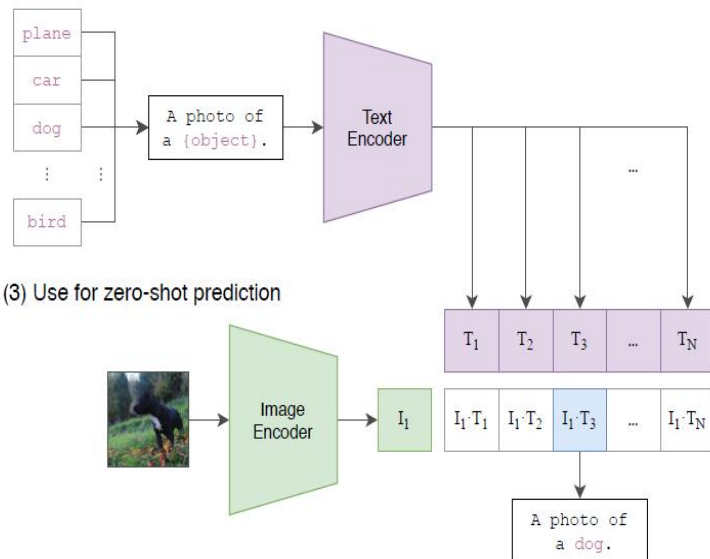
- 开源：Meta的LLaMA系列（普惠学术领域）
- GPT-4v: 视觉遇见语言（跨模态）
- GPT-4o: 全模态前沿（交互能力）

2023.06

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

- ✓ GPT-4可提供多模态能力
- ✓ zero-shot及few-shot的能力
- ✓ GPT-4逻辑推理能力的飞跃
- ✓ GPT-4的安全性已经大幅提升
- ✓ 更强的专属能力（如编程）
- ✓ 处理其它语言的能力
- ✓ 处理更长序列的能力

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

☞ GPT-4o (文科博士生)

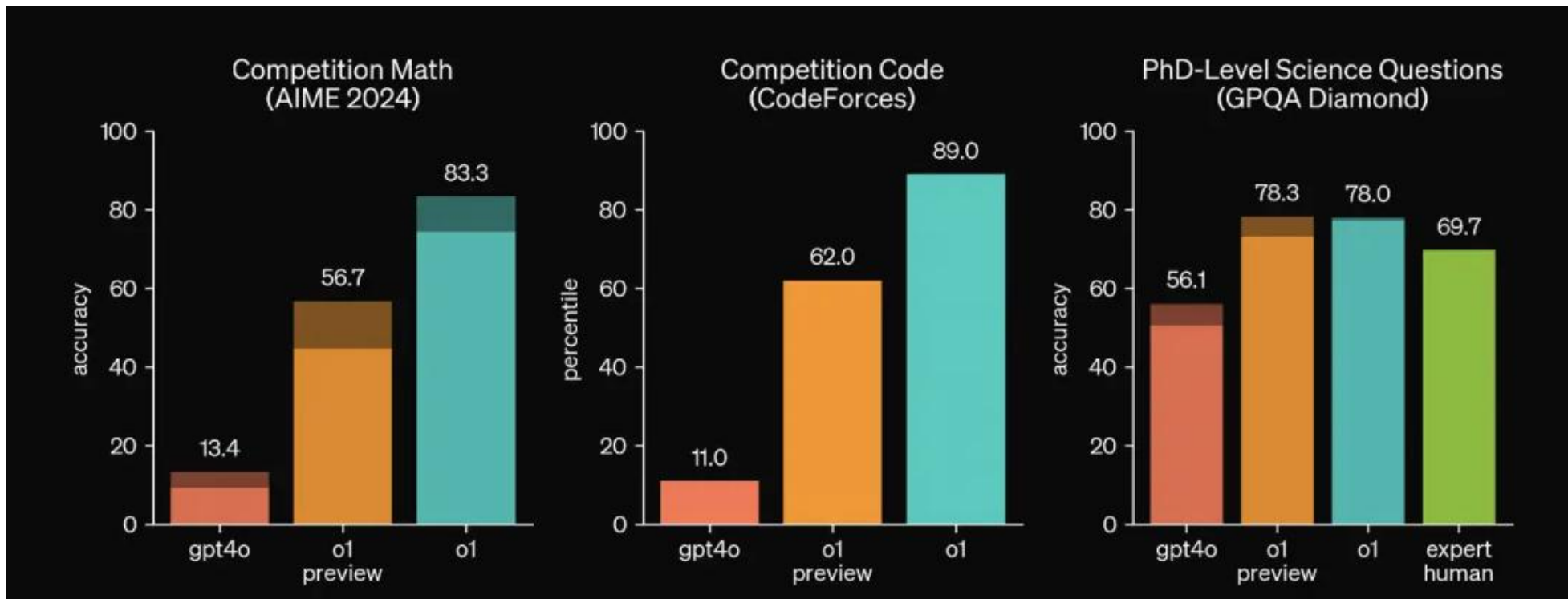
2024.06



- ✓ 多模态输入输出 (交互能力)
- ✓ 响应速度 (接近人类响应)
- ✓ 数学推理、编程等能力提升
- ✓ 非英文文本性能大幅提升
- ✓ 视觉和音频理解能力
- ✓ 成本优势

📊 GPT-o1 (理科博士生)

2024.09



- ✓ 推理能力大幅提升：数学和编程能力爆表
- ✓ 更像人类一样思考：全新安全训练方法 & 更强的“越狱”抵抗力

一、语言模型

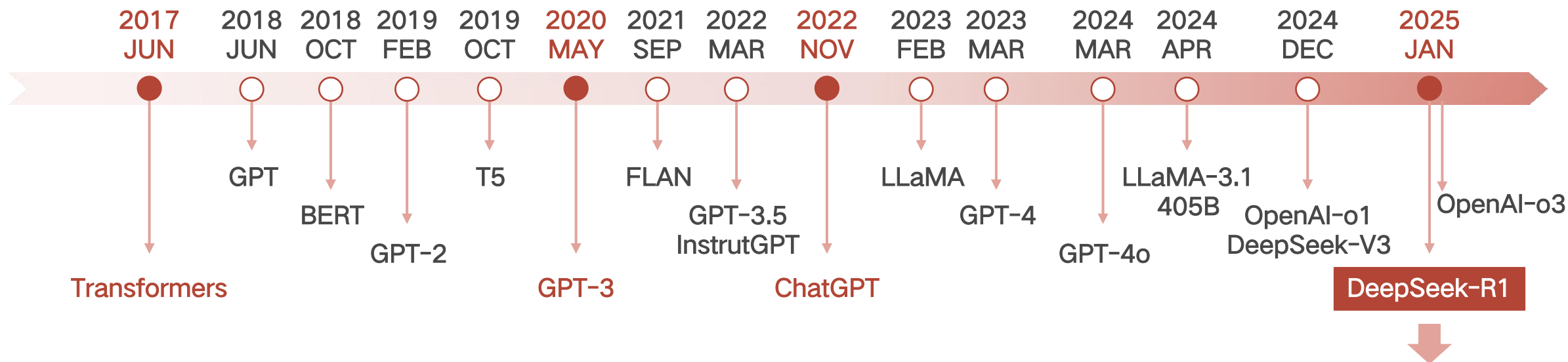
二、Transformer

三、ChatGPT

四、DeepSeek

五、新一代智能体

大型语言模型简史



推理模型：从「生成」到「推理」的重心转变

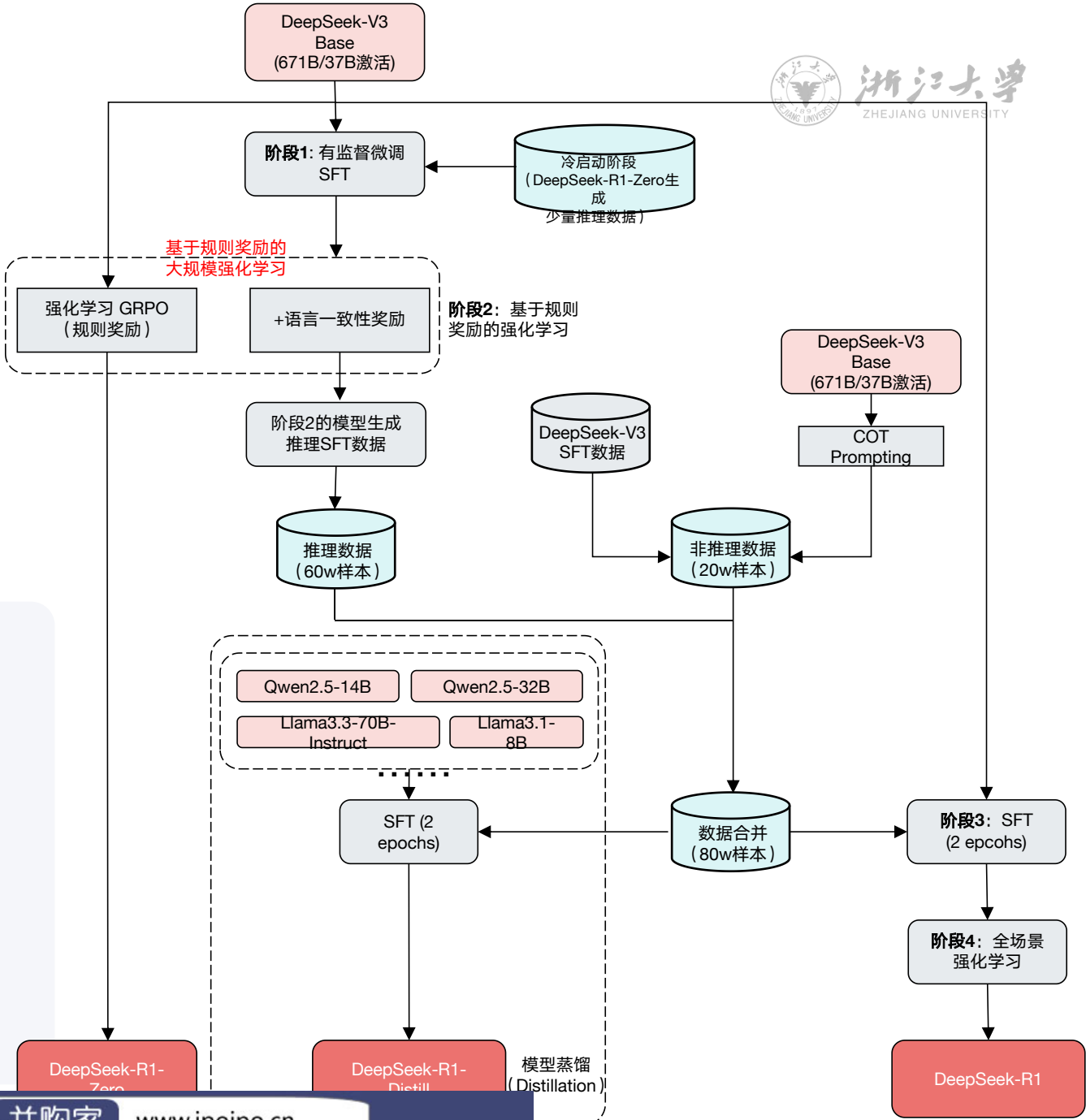
- OpenAI-o1/o3：推理能力的一大飞跃
- DeepSeek-V3/R1：专家模型、强化学习，开源，效率

DeepSeek 技术全景图

DeepSeek

- Step 1: DeepSeek-V3 Base 基础生成模型
- Step 2: DeepSeek-R1-Zero 推理模型初试
- Step 3: DeepSeek-R1 推理模型大成
- Step 4: DeepSeek-R1-Distill R1 蒸馏小模型

DeepSeek模型并非是颠覆性基础理论创新（Transformer-based），其对算法、模型和系统等进行的**系统级协同工程创新**，打破了大语言模型以大算力为核心的预期天花板，为**受限资源下探索通用人工智能**开辟了新的道路。



DeepSeek 技术揭秘

DeepSeek

Step 1: DeepSeek-V3 Base

基础生成模型

Step 2: DeepSeek-R1-Zero

推理模型初试

Step 3: DeepSeek-R1

推理模型大成

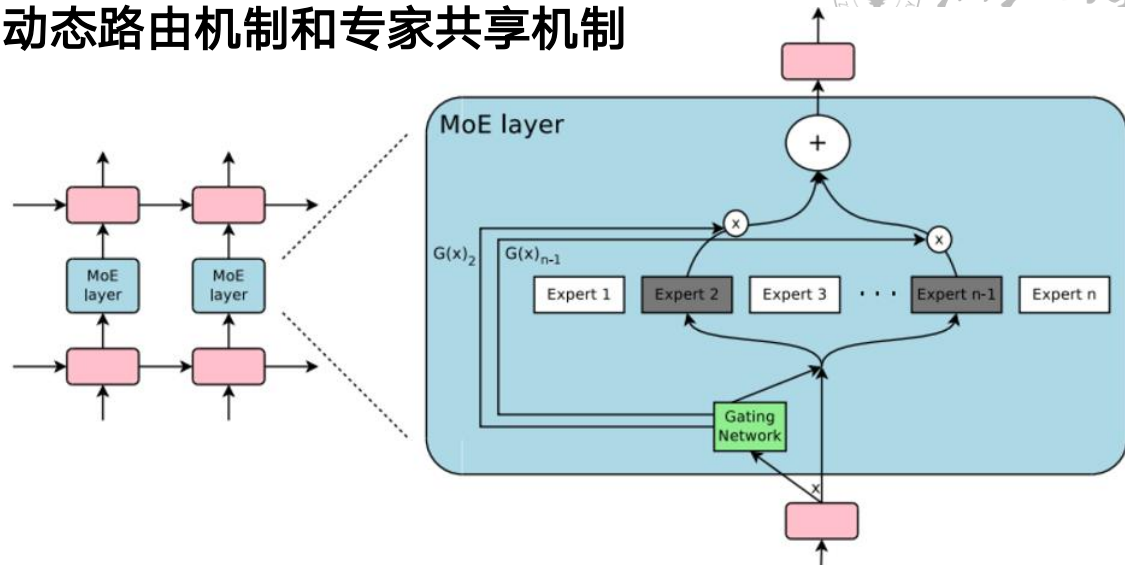
Step 4: DeepSeek-R1-Distill

R1 蒸馏小模型

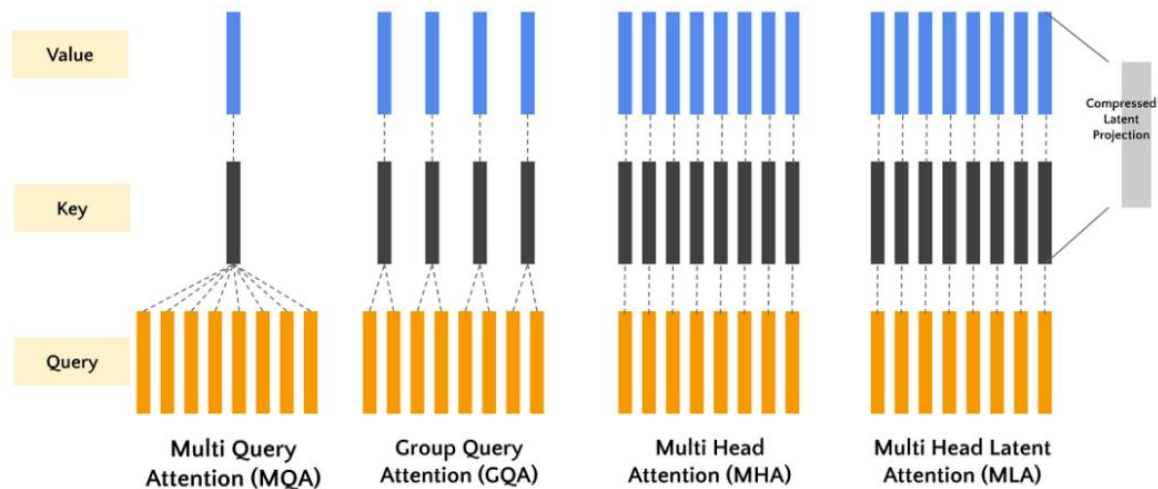
DS-V3对标GPT-4o（文科博士生）：

- **混合专家模型**：V3基座模型总共有6710亿参数，但是每次token仅激活8个专家、370亿参数（~5.5%）。
- **极致的工程优化**：多头潜在注意力机制(**MLA**)，使用FP8混合精度，DualPipe算法提升训练效率，将训练效率优化到极致，显存占用为其他模型的**5%-13%**。

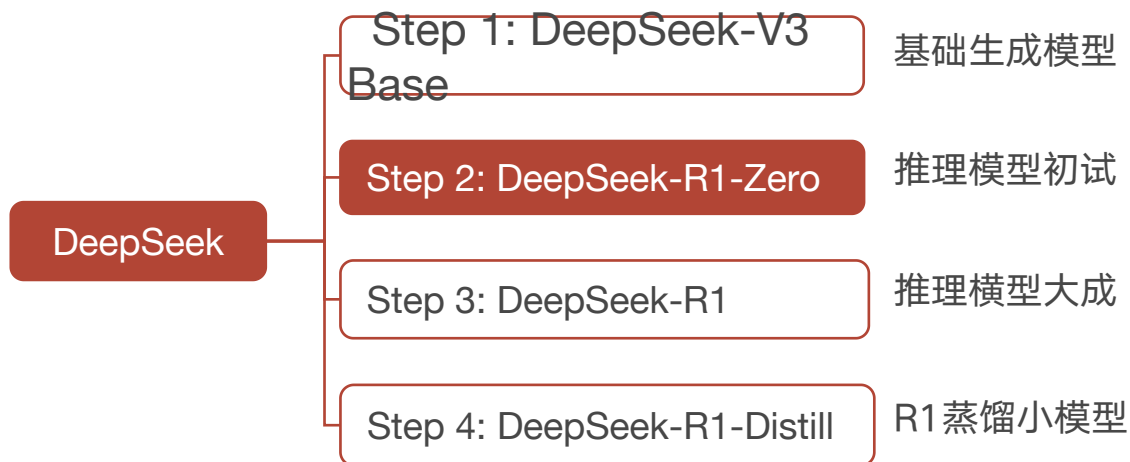
动态路由机制和专家共享机制



Attention - MQA | GQA | MHA | MLA



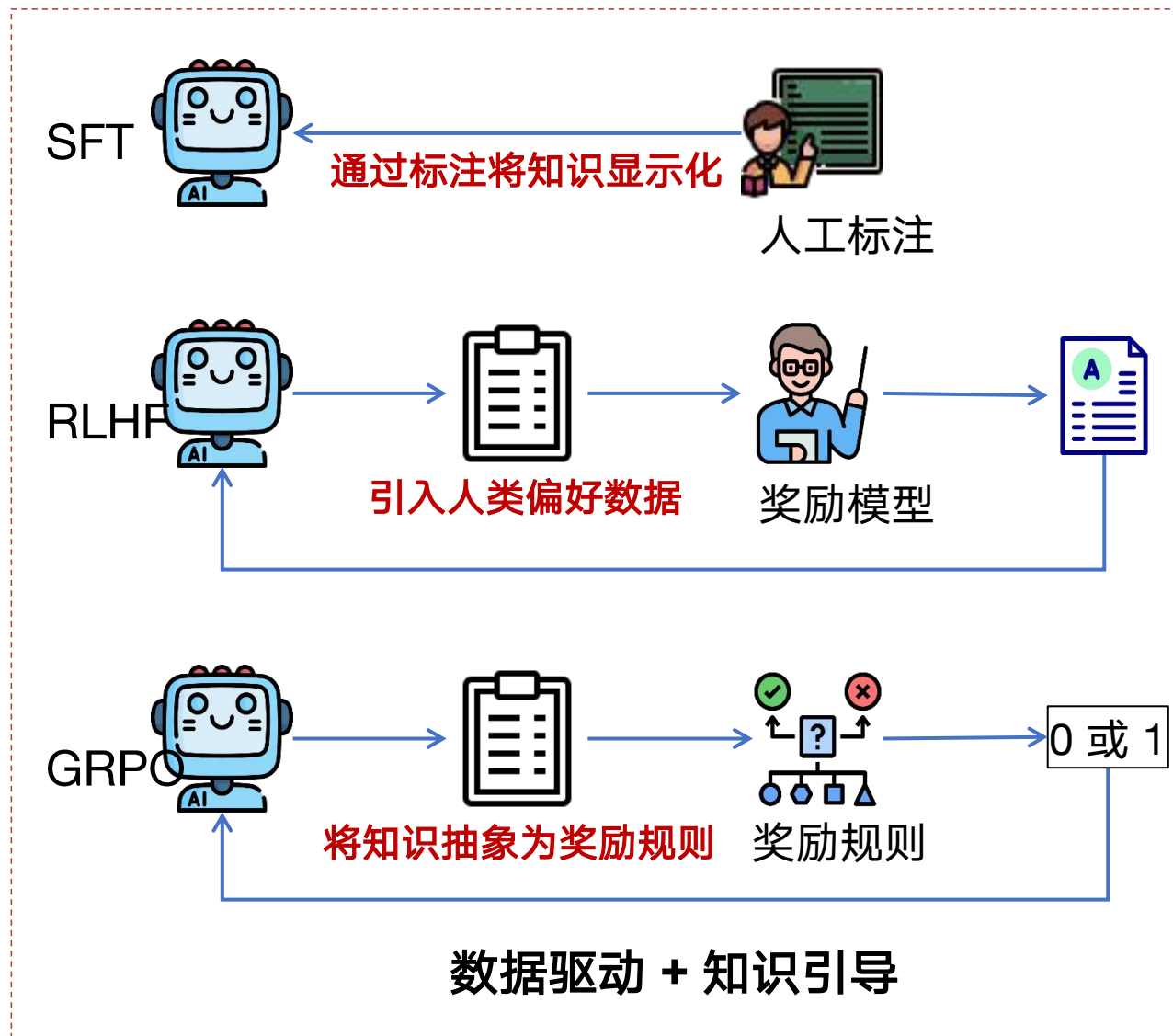
DeepSeek 技术揭秘



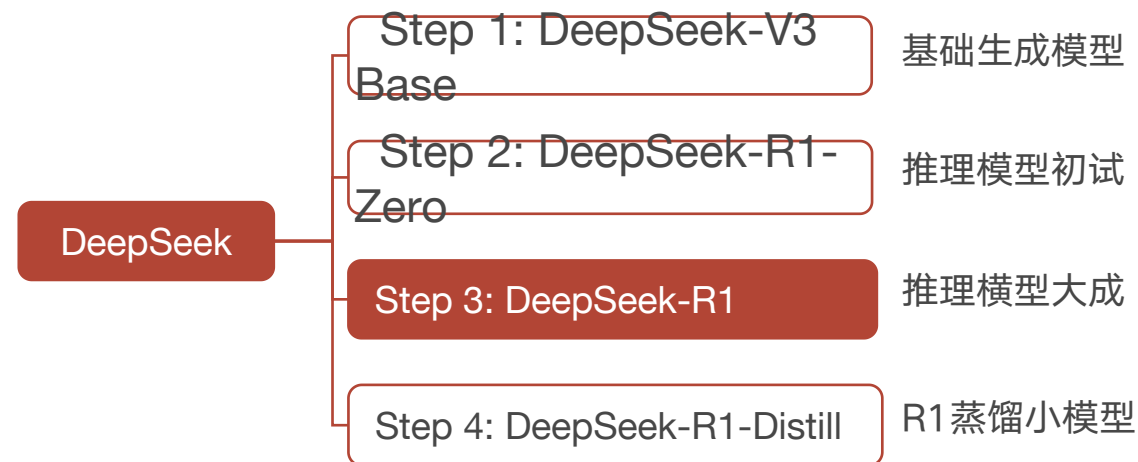
赋予DeepSeek-V3最基础的推理能力：

R1-Zero使用DeepSeek-V3-Base作为基础模型，直接使用 **GRPO** 进行强化学习来提升模型的推理性能：

- 准确度奖励 (Accuracy rewards)
- 格式奖励 (Format rewards)

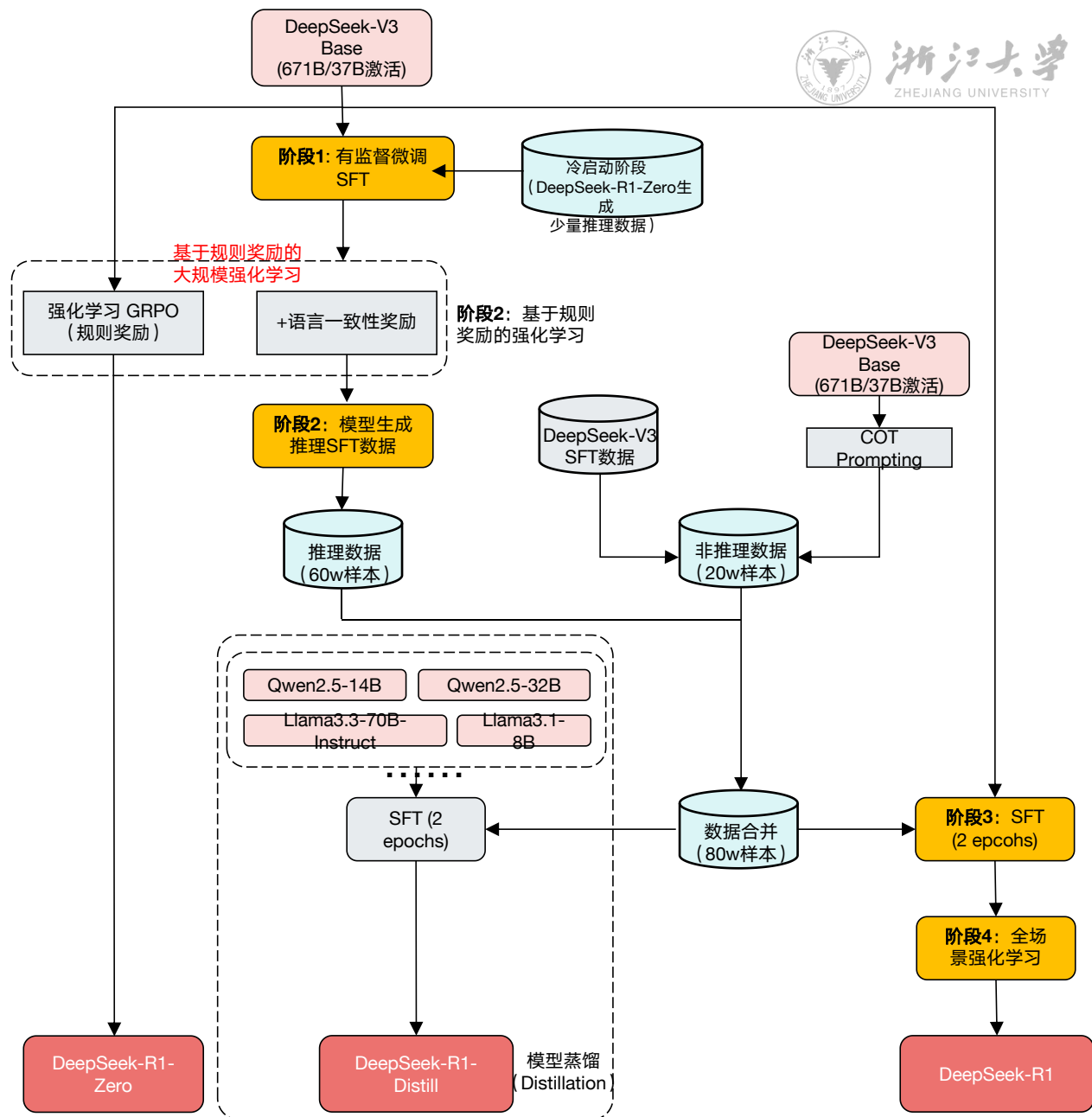


DeepSeek 技术揭秘

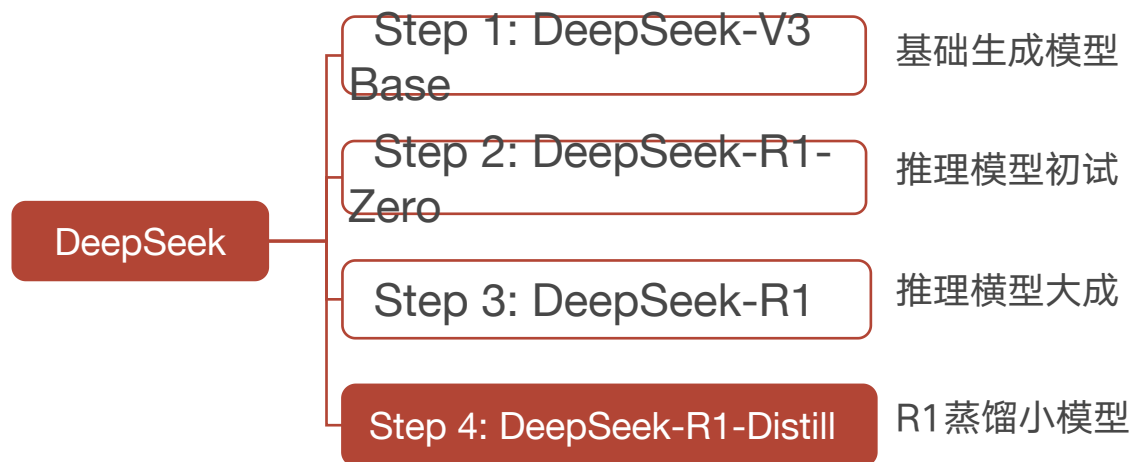


DS-R1对标OpenAI-o1（理科博士生）：

- 阶段1：DeepSeek-R1-Zero生成少量推理数据 + SFT => 为V3植入初步推理能力（冷启动）
- 阶段2：根据规则奖励直接进行强化学习（GRPO）训练=> 提升推理能力（多轮迭代，获取大量推理数据）
- 阶段3：迭代生成推理/非推理样本微调 => 增强全场景能力
- 阶段4：全场景强化学习 => 人类偏好对齐（RLHF）

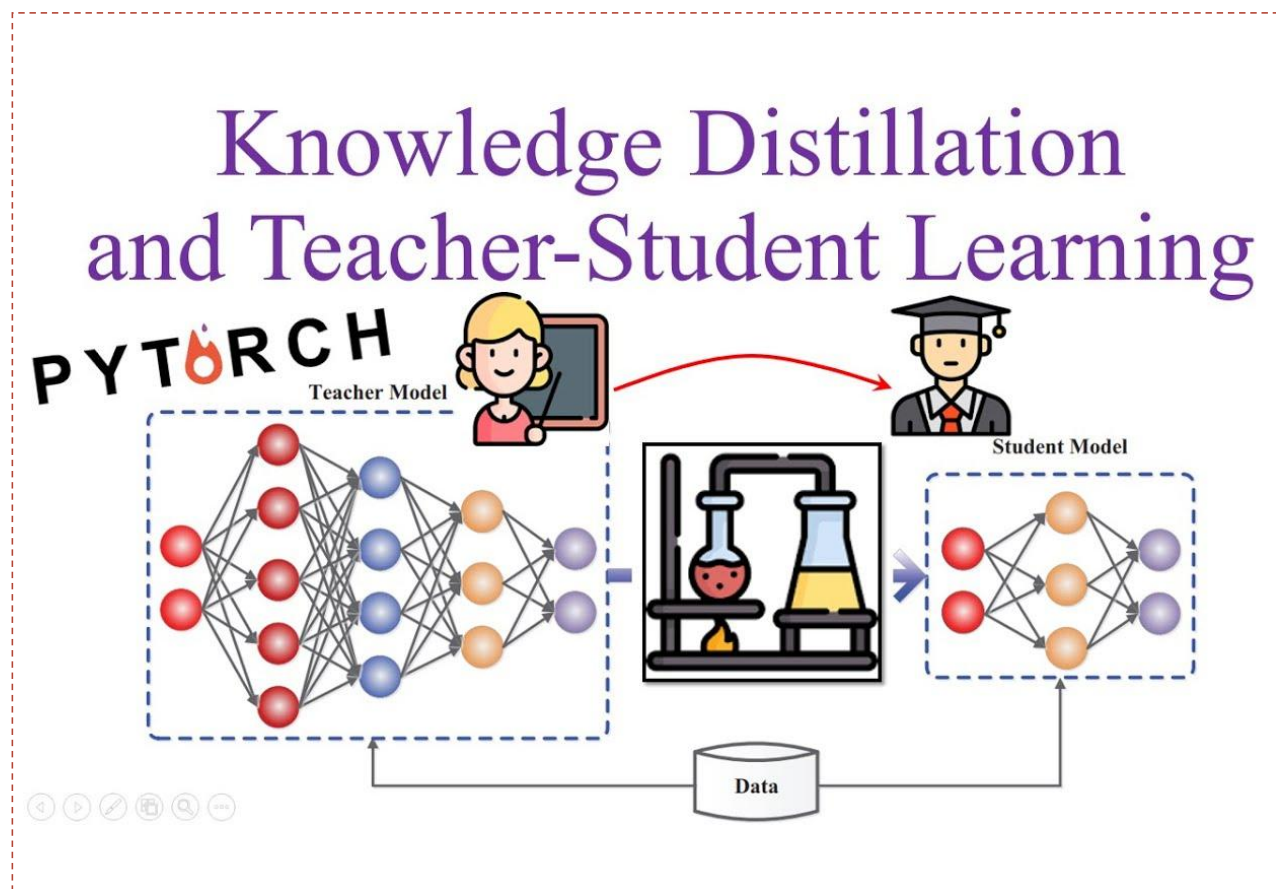


DeepSeek 技术揭秘



DeepSeek-R1-Distill模型:

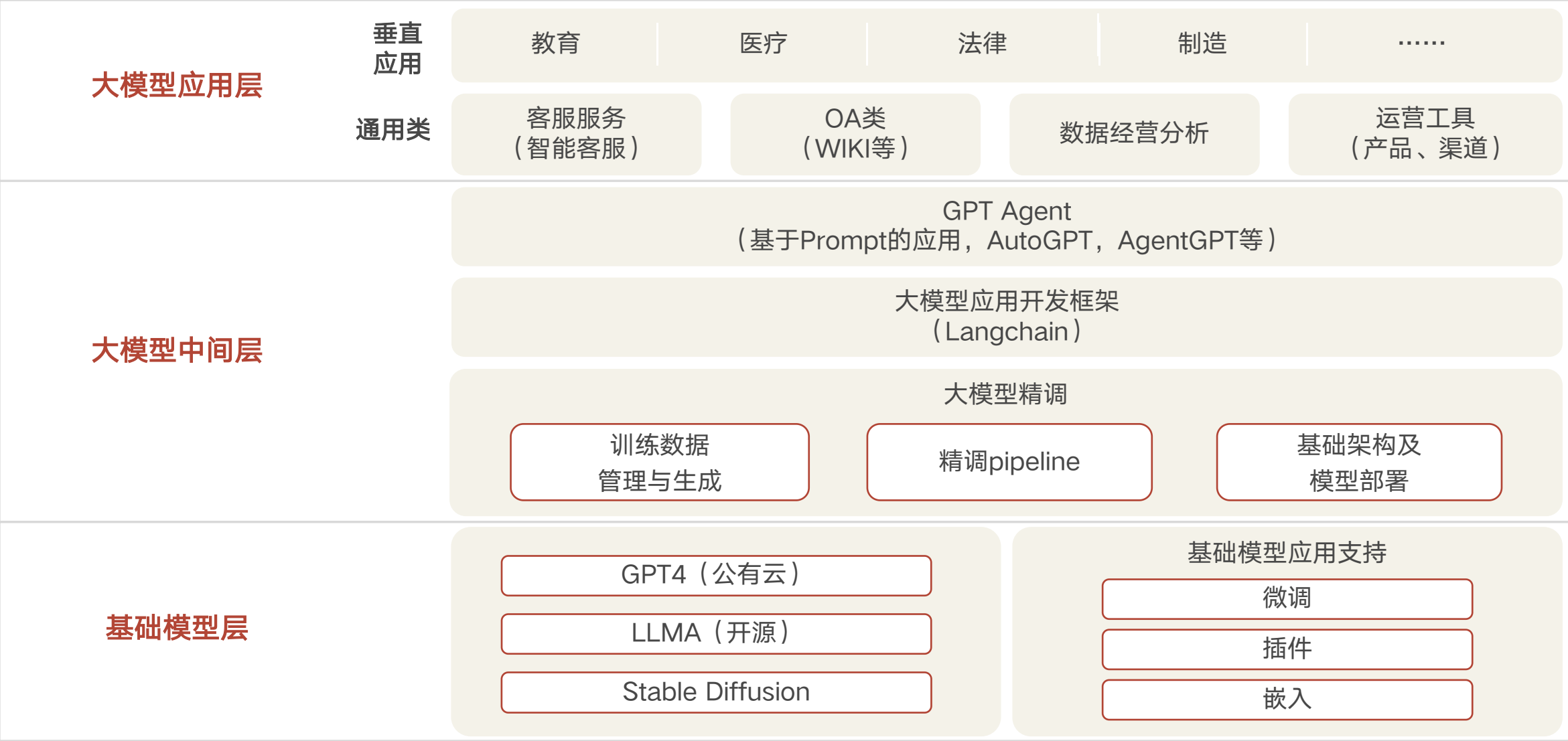
- (1) 基于各个低参数量通用模型（千问、Llama等）
- (2) 使用DeepSeek-R1同款数据微调
- (3) 大幅提升低参数量模型性能



知识蒸馏:

- 老师教学生：“解题思路”，不仅给答案（硬标签），还教“为什么”（软标签）
- 模型瘦身：大幅压缩参数（如671亿→7亿参数），手机也能跑AI

DeepSeek 带来的全栈影响



一、语言模型

二、Transformer

三、ChatGPT

四、DeepSeek

五、新一代智能体

从 LLM 到 Agent

通用LLM

ChatGPT (2022)

LLaMA (2023)

Vicuna (2023)

垂类LLM

Code Llama (2023)

MathGLM (2023)

LawBench (2023)

基于LLM的Agent

HuggingGPT (2023)

AutoGPT (2023)

JARVIS (2024)

01

02

03

04

05

06

07

技术架构

Transformer (2017)

Bert/GPT (2018)

大模型开发工具

LangChain (2022)

LlamaIndex (2023)

垂类应用

LLM VSCode (2023)

DB GPT-Hub (2023)

Kore.ai (2023)

Uchat (2024)

Agent开发平台

GPTs (2023)

Coze (2024)

Agent Builder (2024)

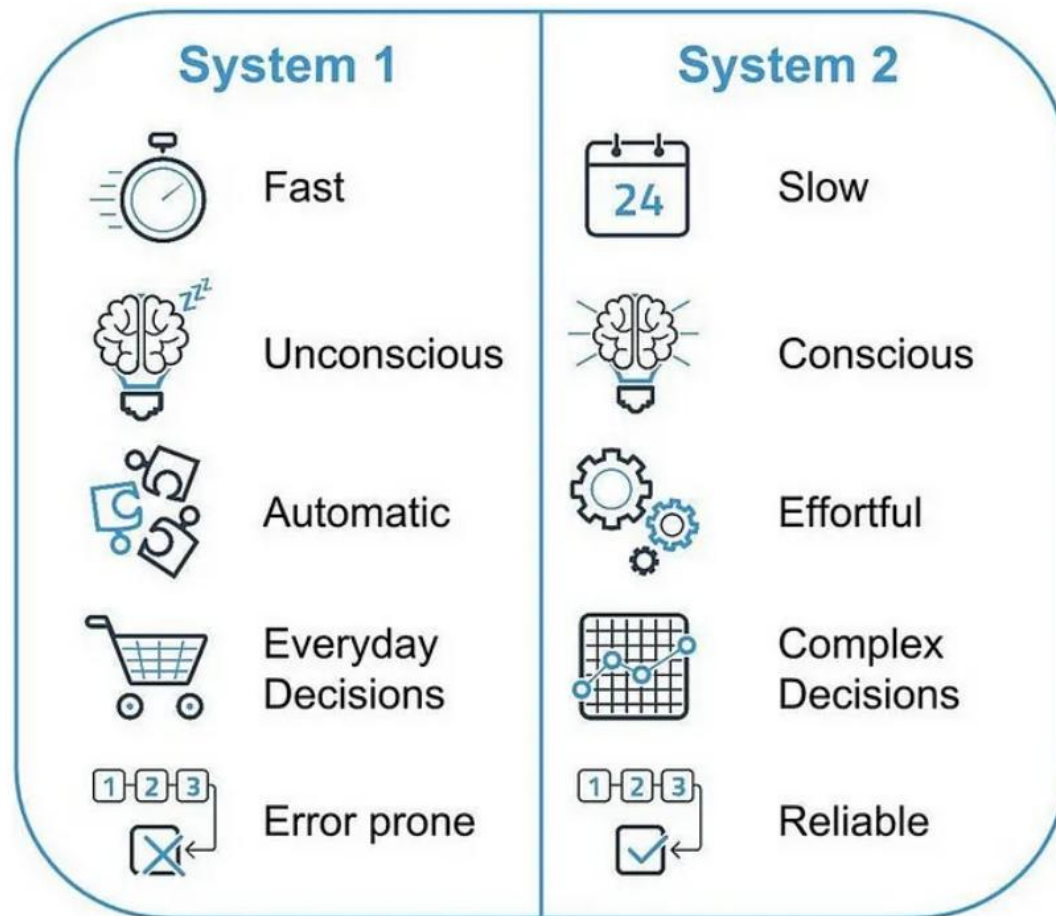
Agent OS (2024)



Deepseek

生成大模型「系统1」到推理大模型「系统2」

GPT-4v/4o
DeepSeek-V3



GPT-o1/o3
DeepSeek-R

「系统1」（快速、直觉）和「系统2」（缓慢、分析）

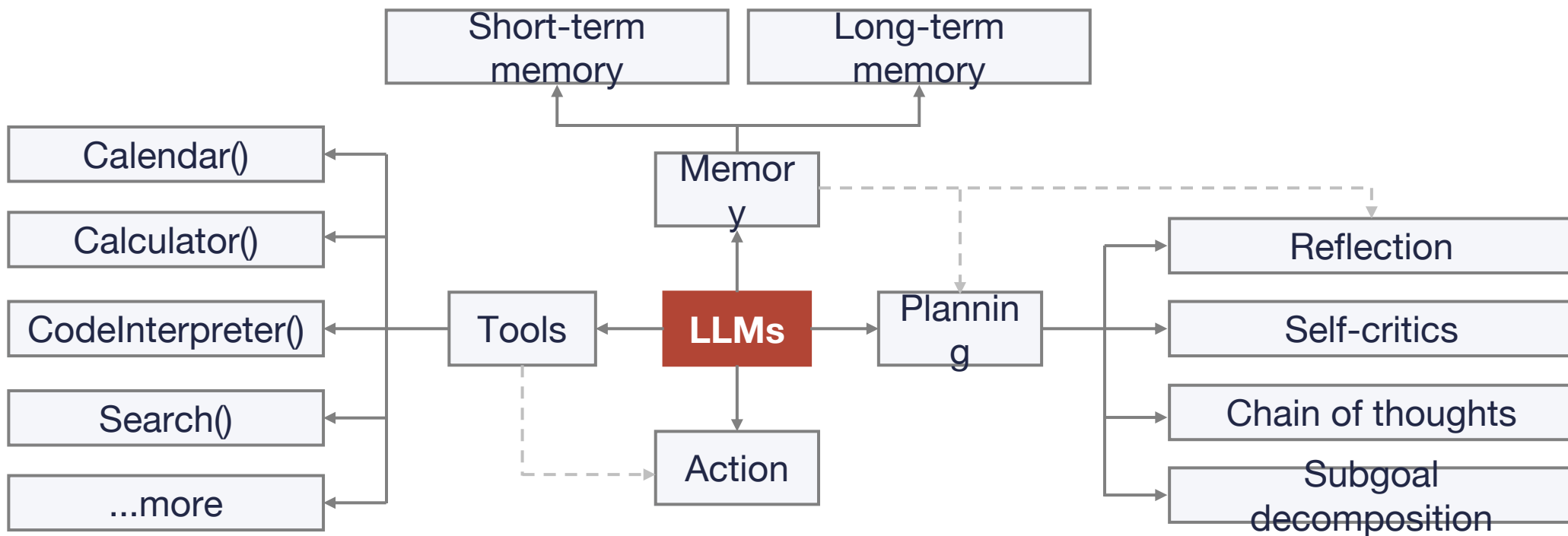
新一代智能体 = Agent + LLM

LLM是Agent的大脑，其核心能力是“逻辑推理”「系统2」

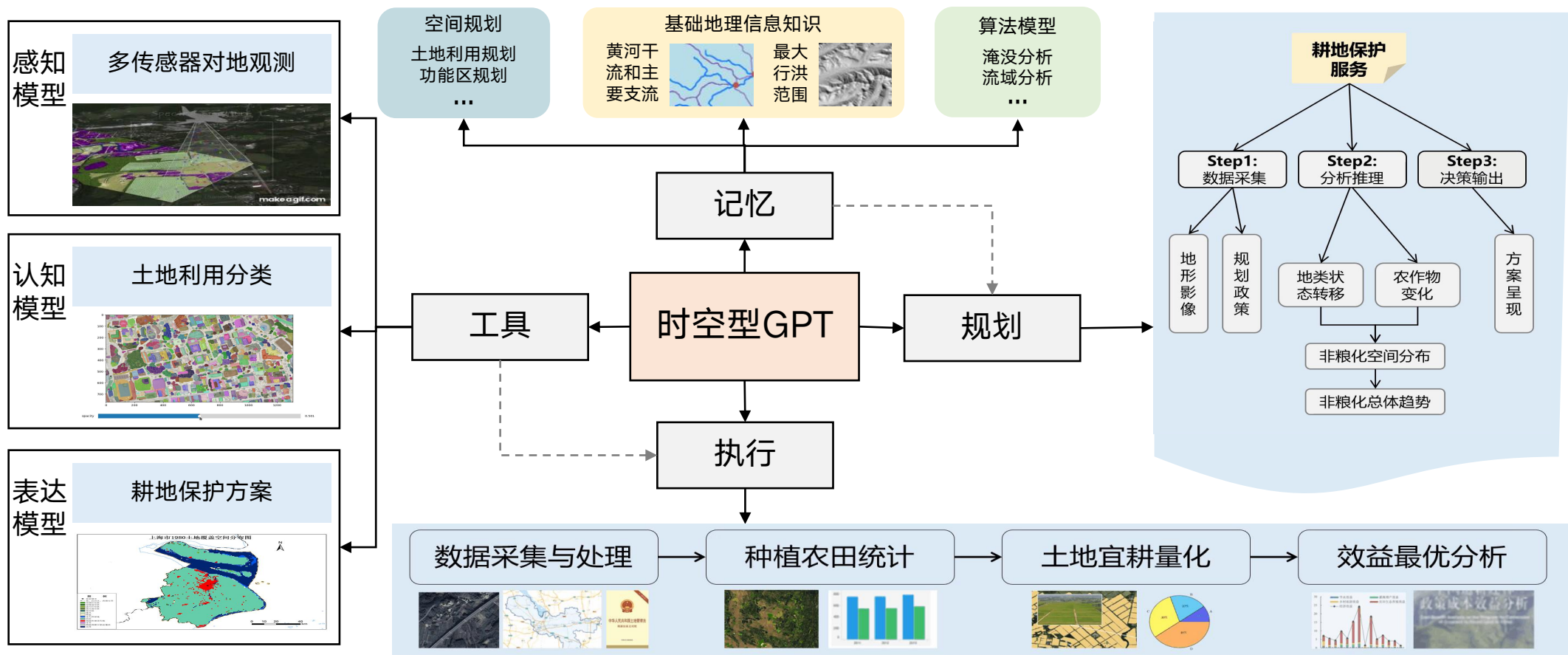
Planning Skills: 对问题进行拆解得到解决路径，既进行任务规划

Tool Use: 评估自己所需的工具，进行工具选择，并生成调用工具请求

Memory: 短期记忆包括工具返回值，已完成推理路径；长期记忆包括可访问的外部长期存储等



时空智能的自主化服务（国自然基金重大课题）



由“**时空型GPT**”作为决策大脑驱动，构成一个闭环多智能体协同系统
实现流程**自**组织、任务**自**执行、内容**自**生成，即时空智能的**自主化构建**

THANKS

感谢观看