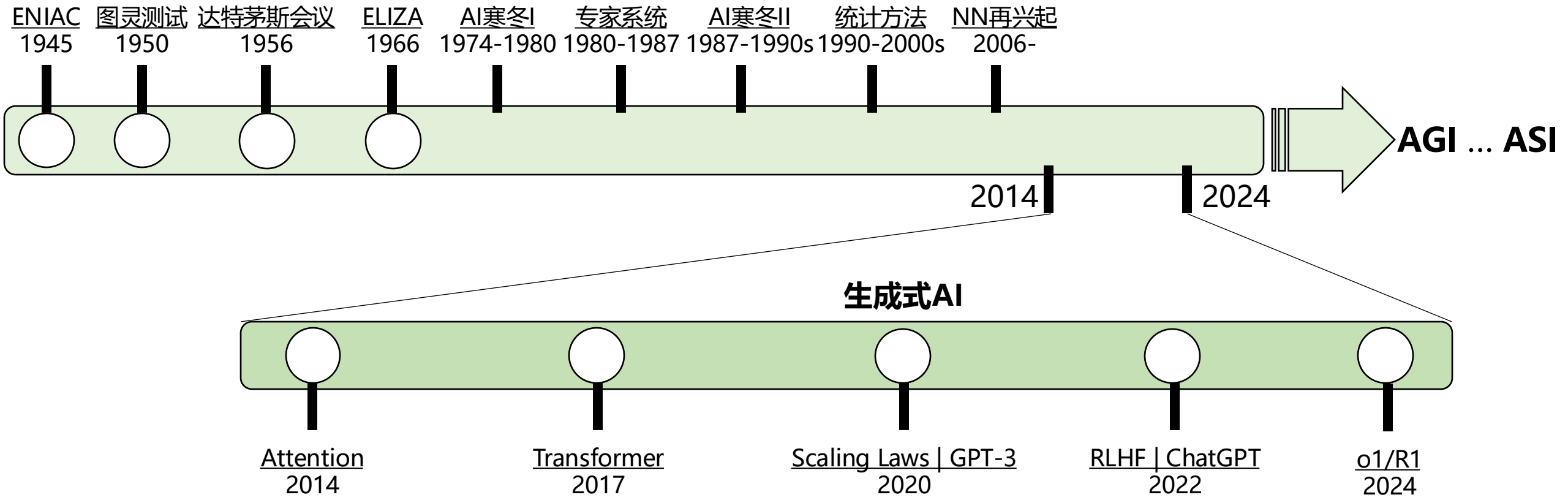
深度解读DeepSeek：原理与效应

熊德意 天津大学
dyxiong@tju.edu.cn
https://dyxiong.github.io
https://tjunlp-lab.github.io

仁文伏羲 伏羲传语 OpenEval CALM

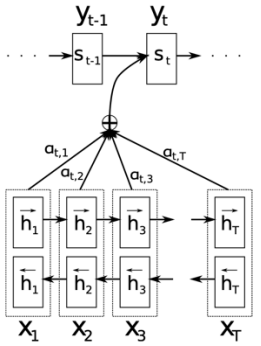
报告目录

生成式AI：2014 —— 2024



生成式AI：使用生成式模型生成**各类数据**（语言、语音、图片、视频等）

- **Attention**：数据依存关系建模
- **Transformer**：数据生成的统一架构
- **Scaling Laws**：数据学习、生成的扩展法则
- **RLHF**：生成与人类价值对齐的数据
- **o1/R1**：生成式求解问题——生成问题求解的过程和答案（推理）



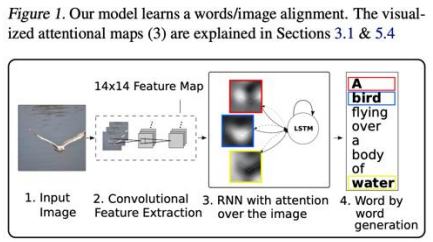
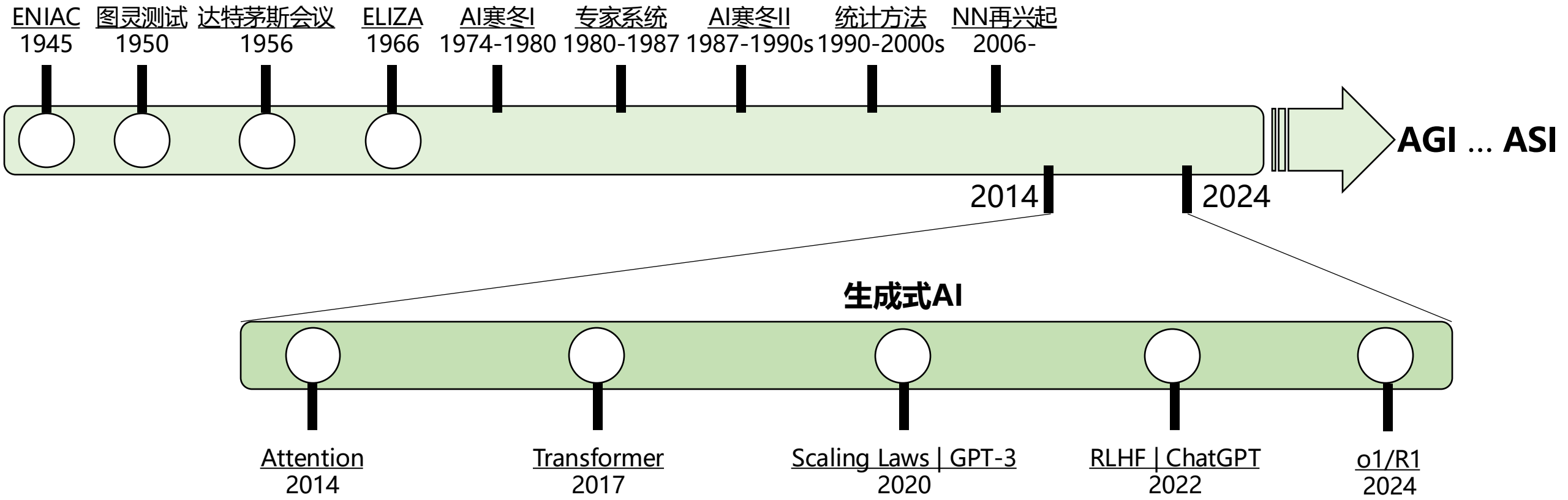


Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in Sections 3.1 & 5.4

生成式AI：2014 —— 2024

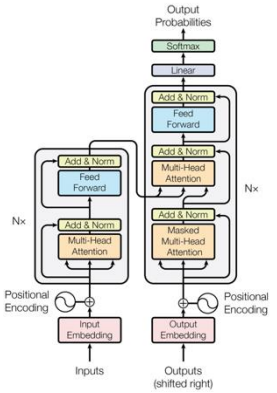


生成式AI：使用生成式模型生成**各类数据**（语言、语音、图片、视频等）

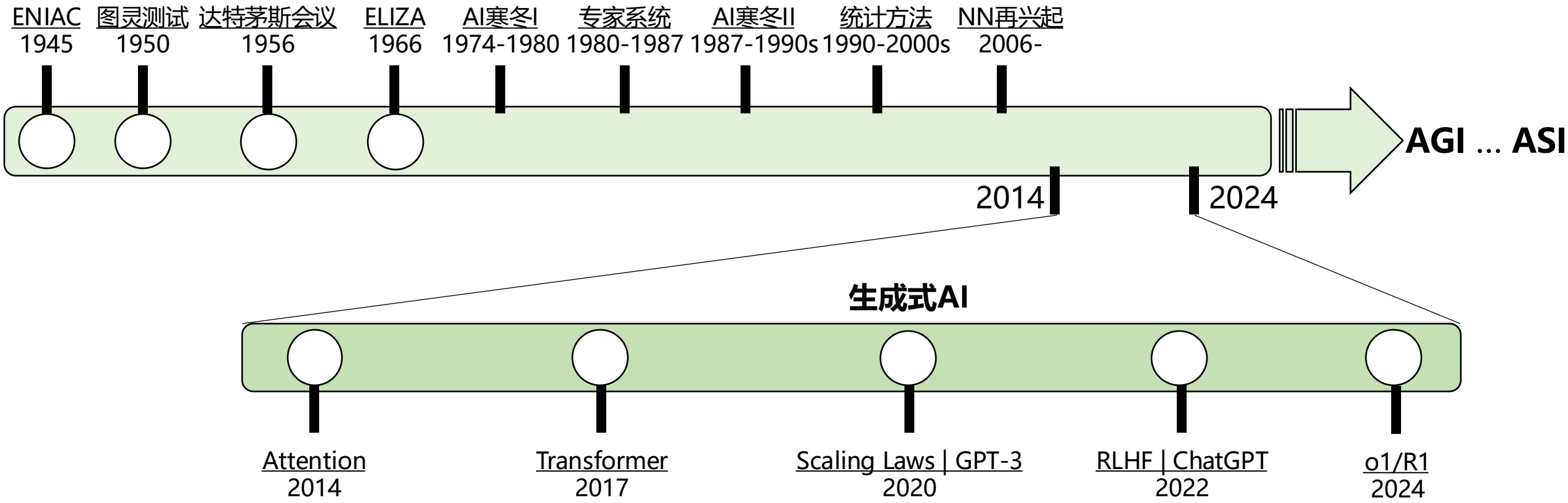
- **Attention**：数据依存关系建模
- **Transformer**：数据生成的统一架构
- **Scaling Laws**：数据学习、生成的扩展法则
- **RLHF**：生成与人类价值对齐的数据
- **o1/R1**：生成式求解问题——生成复杂问题的答案（推理）


Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

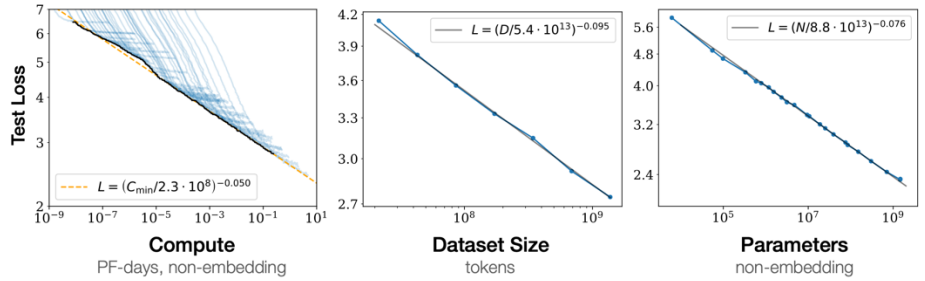


生成式AI：2014 —— 2024

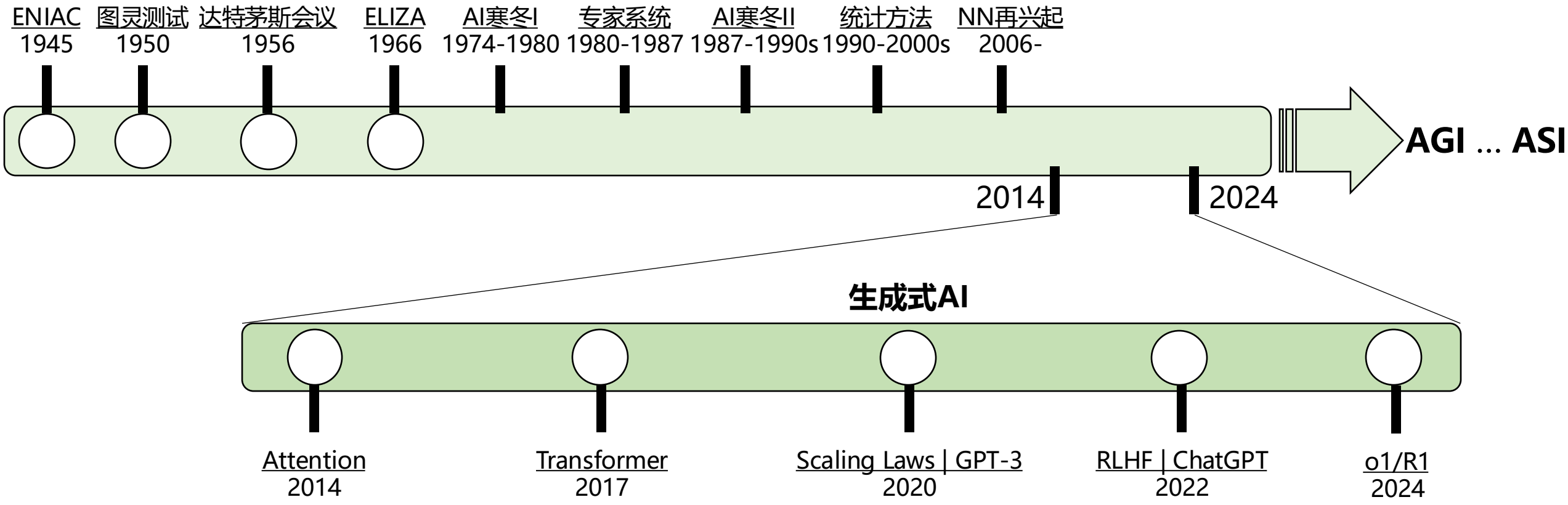


生成式AI：使用生成式模型生成**各类数据**（语言、语音、图片、视频等）

- **Attention**：数据依存关系建模
- **Transformer**：数据生成的统一架构
- **Scaling Laws**：数据学习、生成的扩展法则
- **RLHF**：生成与人类价值对齐的数据
- **o1/R1**：生成式求解问题——生成复杂问题的答案（推理）

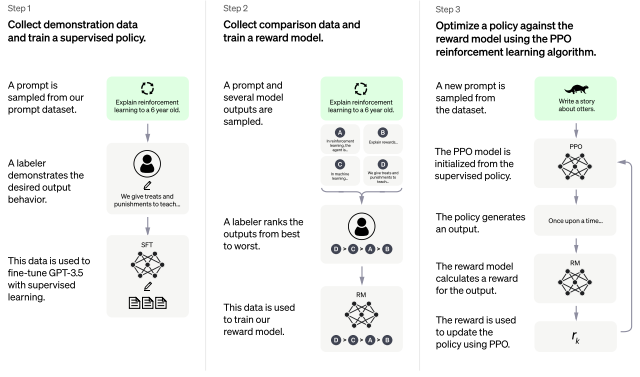


生成式AI：2014 —— 2024

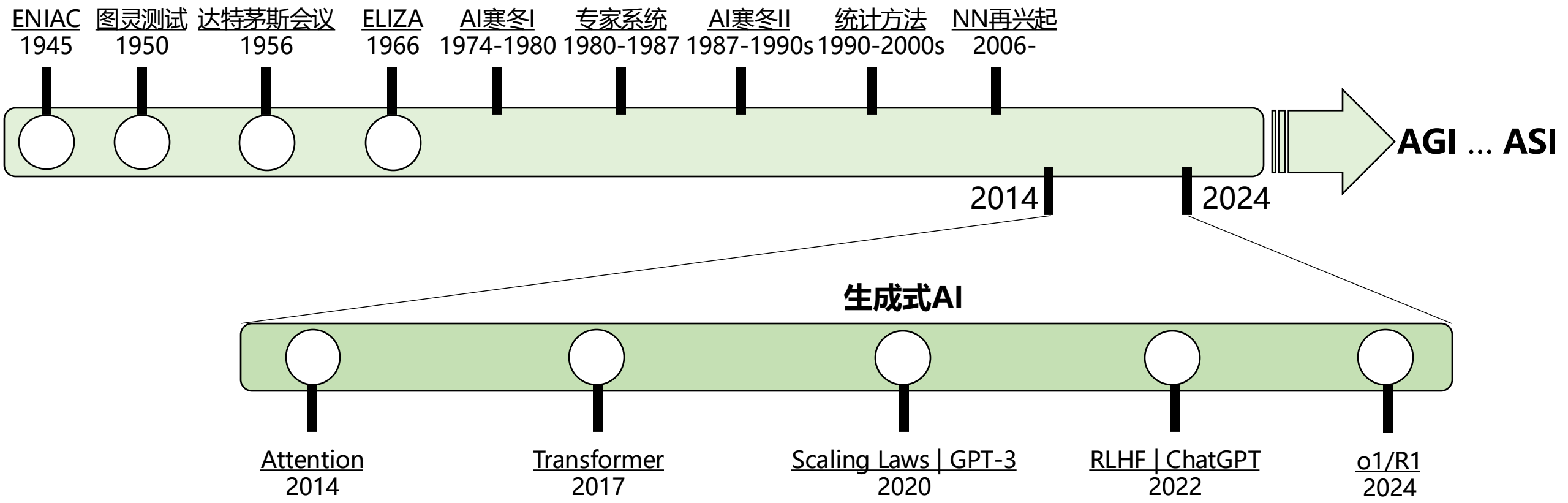


生成式AI：使用生成式模型生成**各类数据**（语言、语音、图片、视频等）

- **Attention**：数据依存关系建模
- **Transformer**：数据生成的统一架构
- **Scaling Laws**：数据学习、生成的扩展法则
- **RLHF**：生成与人类价值对齐的数据
- **o1/R1**：生成式求解问题——生成复杂问题的答案（推理）



生成式AI：2014 —— 2024



生成式AI：使用生成式模型生成**各类数据**（语言、语音、图片、视频等）

- **Attention**：数据依存关系建模
- **Transformer**：数据生成的统一架构
- **Scaling Laws**：数据学习、生成的扩展法则
- **RLHF**：生成与人类价值对齐的数据
- **o1/R1**：生成式求解问题——生成复杂问题的答案（推理）

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a-\sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a-\sqrt{a+x}} = x$, let's start by squaring both ···

$\left(\sqrt{a-\sqrt{a+x}}\right)^2 = x^2 \implies a-\sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a-x^2)^2 = a+x \implies a^2-2ax^2+(x^2)^2 = a+x \implies x^4-2ax^2-x+(a^2-a) = 0$

···

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ···

We started with the equation:

$\sqrt{a-\sqrt{a+x}} = x$

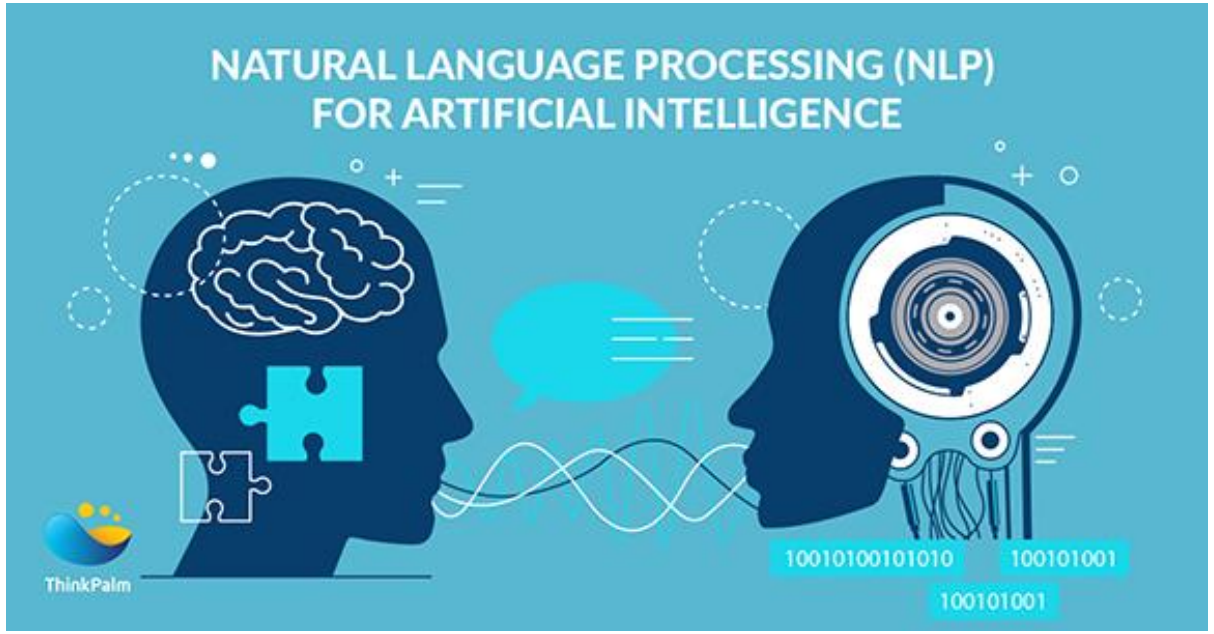
First, let's square both sides:

$a-\sqrt{a+x} = x^2 \implies \sqrt{a+x} = a-x^2$

Next, I could square both sides again, treating the equation: ···

···

自然语言处理与语言模型



Language Models

Trained to predict the next word in a sentence:

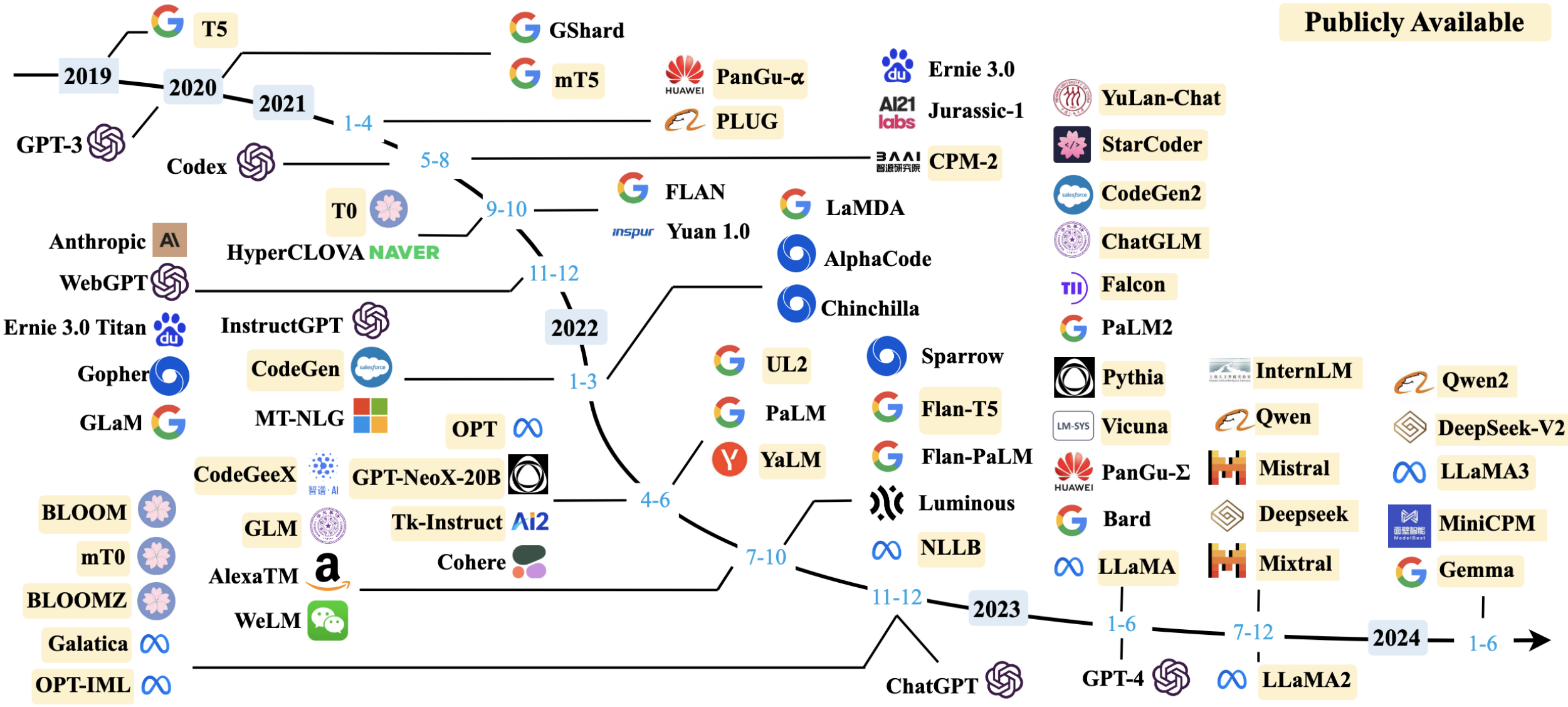
The cat is chasing the _____

- **dog** 5%
- **mouse** 70%
- **squirrel** 20%
- **boy** 5%
- **house** 0%

自然语言处理：人类语言的智能化处理与分析，使计算机具备听、说、读、写、译等人所具备的语言能力

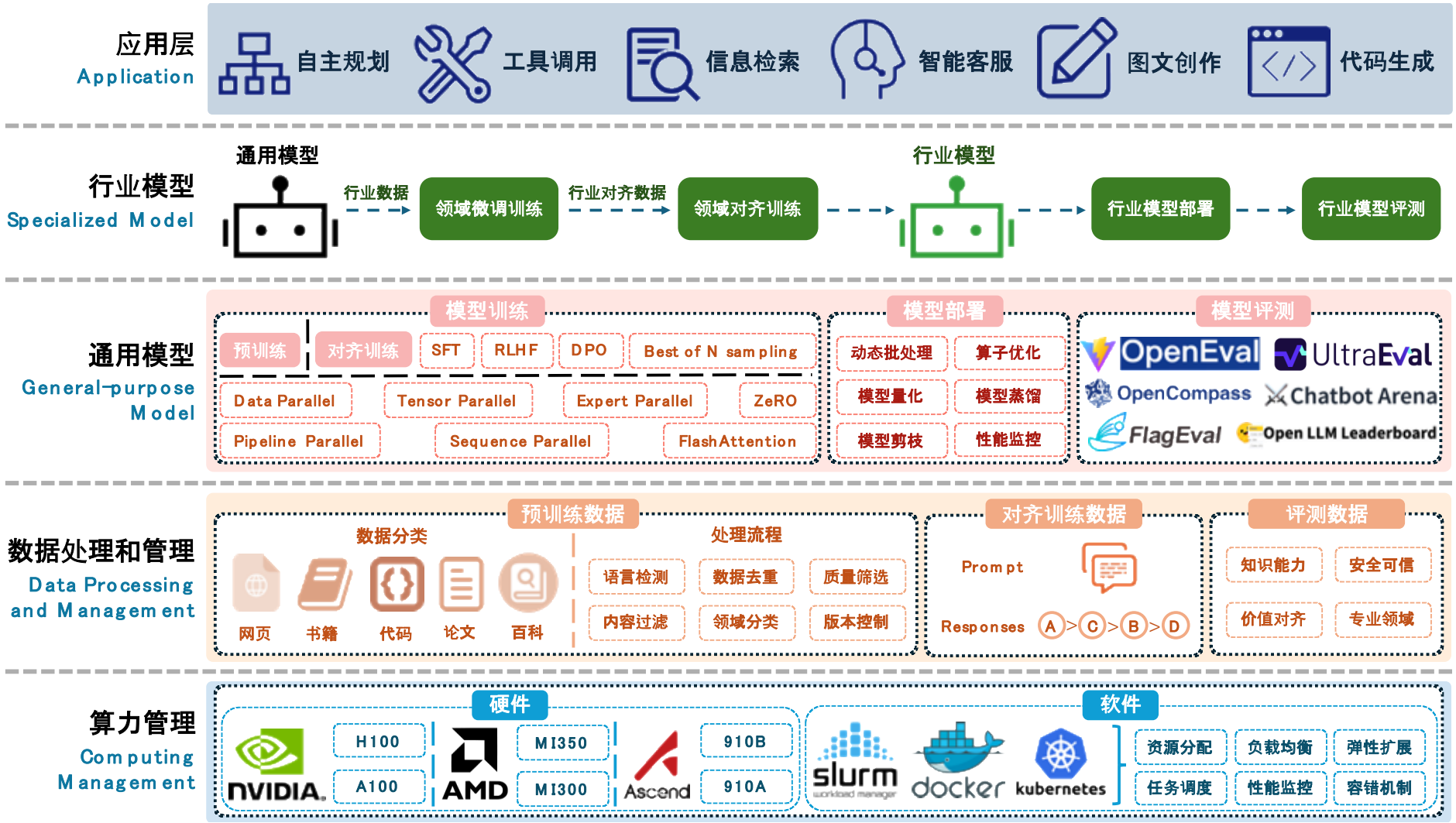
语言模型：自然语言统计建模，简单说，就是预测句子中的下一个单词是什么

大语言模型：2018 —— 2024

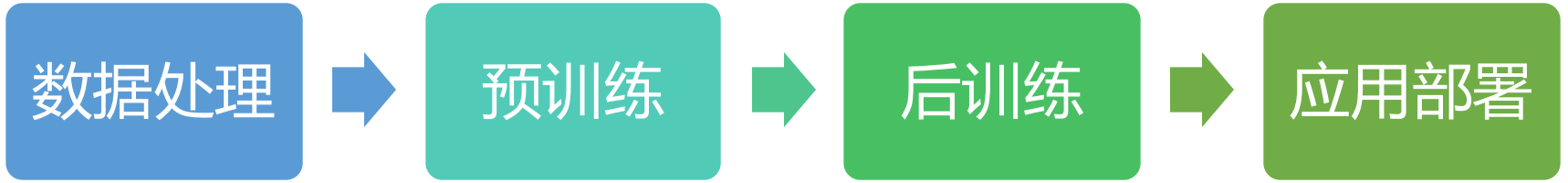


Zhao et al. A Survey of Large Language Models. arXiv:2303.18223

大语言模型：技术栈


应用层
Application
自主规划
工具调用
信息检索
智能客服
图文创作
代码生成
行业模型
Specialized Model
通用模型
行业数据
领域微调训练
行业对齐数据
领域对齐训练
行业模型
行业模型部署
行业模型评测
通用模型
General-purpose Model
模型训练
预训练
对齐训练
SFT
RLHF
DPO
Best of N sampling
Data Parallel
Tensor Parallel
Expert Parallel
ZeRO
Pipeline Parallel
Sequence Parallel
FlashAttention
模型部署
动态批处理
算子优化
模型量化
模型蒸馏
模型剪枝
性能监控
模型评测
OpenEval
UltraEval
OpenCompass
Chatbot Arena
FlagEval
Open LLM Leaderboard
数据处理和管理
Data Processing and Management
预训练数据
数据分类
网页
书籍
代码
论文
百科
处理流程
语言检测
数据去重
质量筛选
内容过滤
领域分类
版本控制
对齐训练数据
Prompt
Responses
A>C>B>D
评测数据
知识能力
安全可信
价值对齐
专业领域
算力管理
Computing Management
硬件
NVIDIA
H100
A100
AMD
MI350
MI300
Ascend
910B
910A
软件
slurm
docker
kubernetes
资源分配
负载均衡
弹性扩展
任务调度
性能监控
容错机制

大语言模型：生命周期与范式



- 数据治理
- 数据要素
- 知识源头

- 基座模型
- 自监督学习
- 能力涌现

- 对齐模型
- 微调&强化
- 安全可信

- 红队测试
- 商业落地
- 模型压缩

- 训练范式
  - 预训练 —— 基座模型
  - 后训练 —— 对齐模型
  - 推理训练 —— 推理模型

- 关键
  - 模型架构
  - 训练算法
  - 扩展法则

杀手锏：性能/成本 曲线 | 性价比

扩展法则

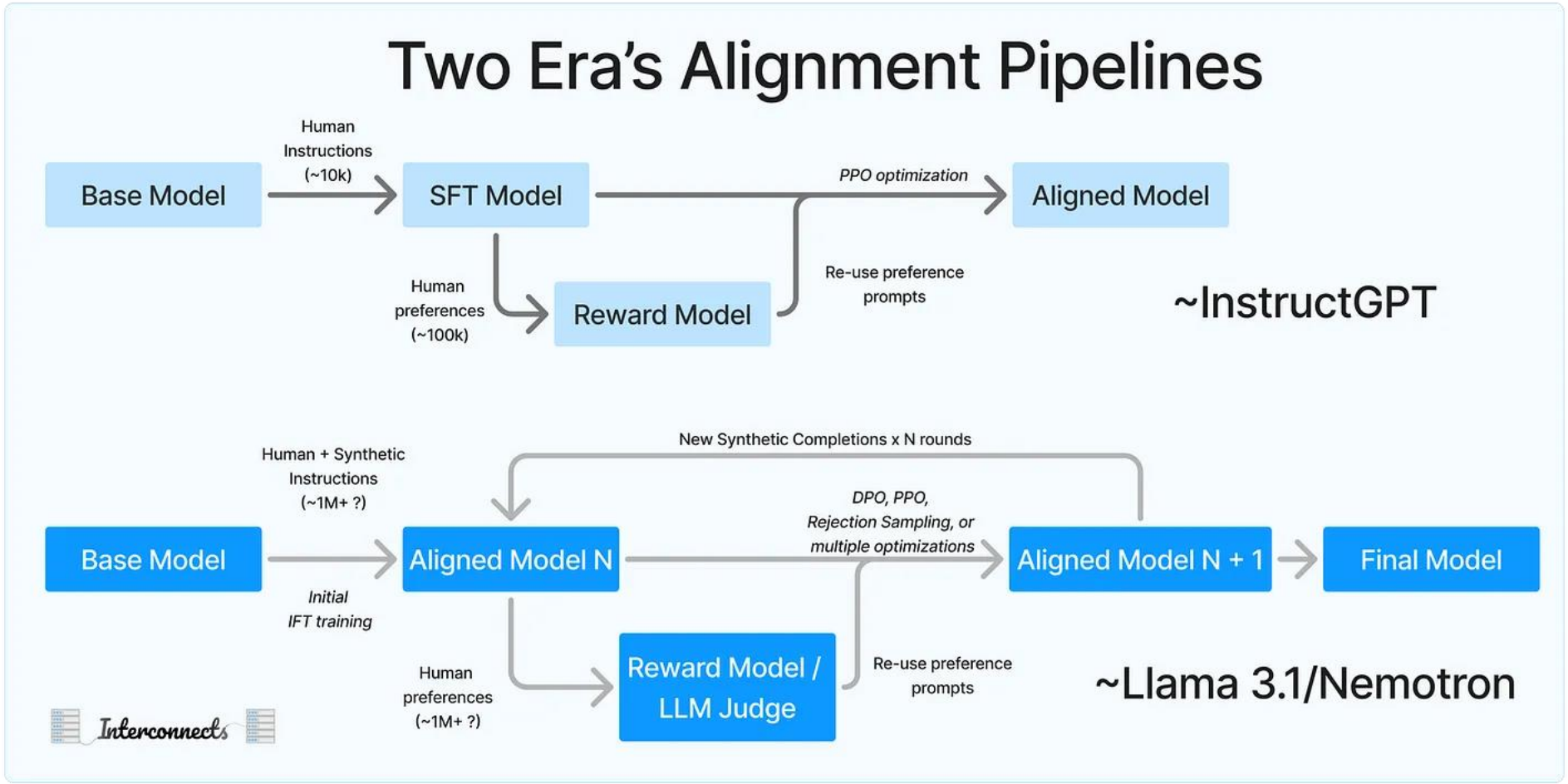
The Bitter Lesson

[Sutton, 2019]

The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by **search and learning**.

Sasha Rush and Daniel Ritter. Speculations on Test-Time Scaling. 2024

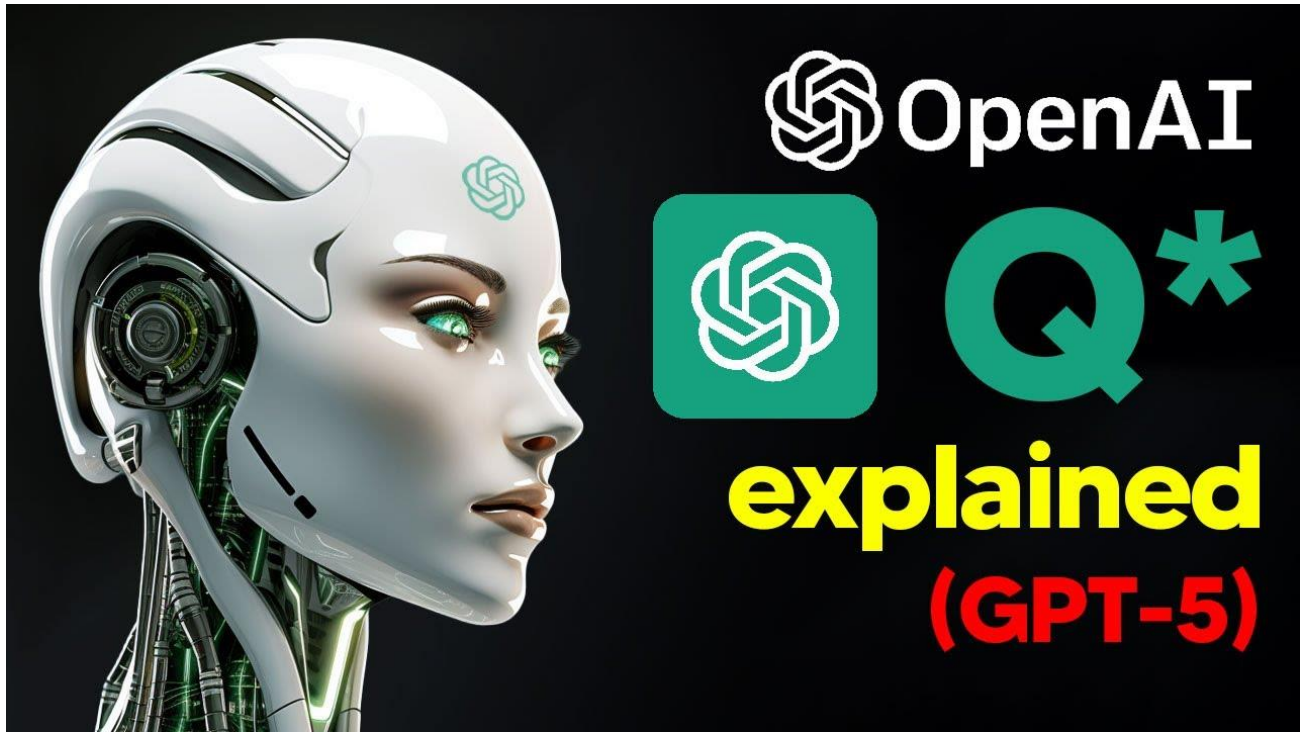
大语言模型：后训练范式

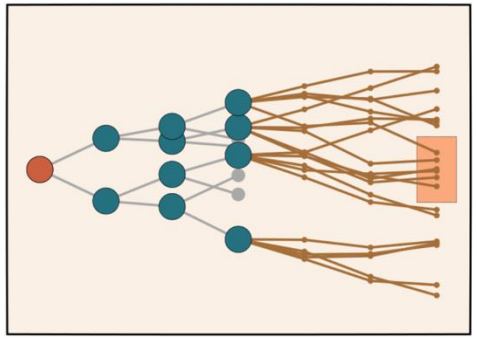


成本较低
大部分实验室可做

成本高昂（上千万）
少数企业/实验室可做

推理语言模型?



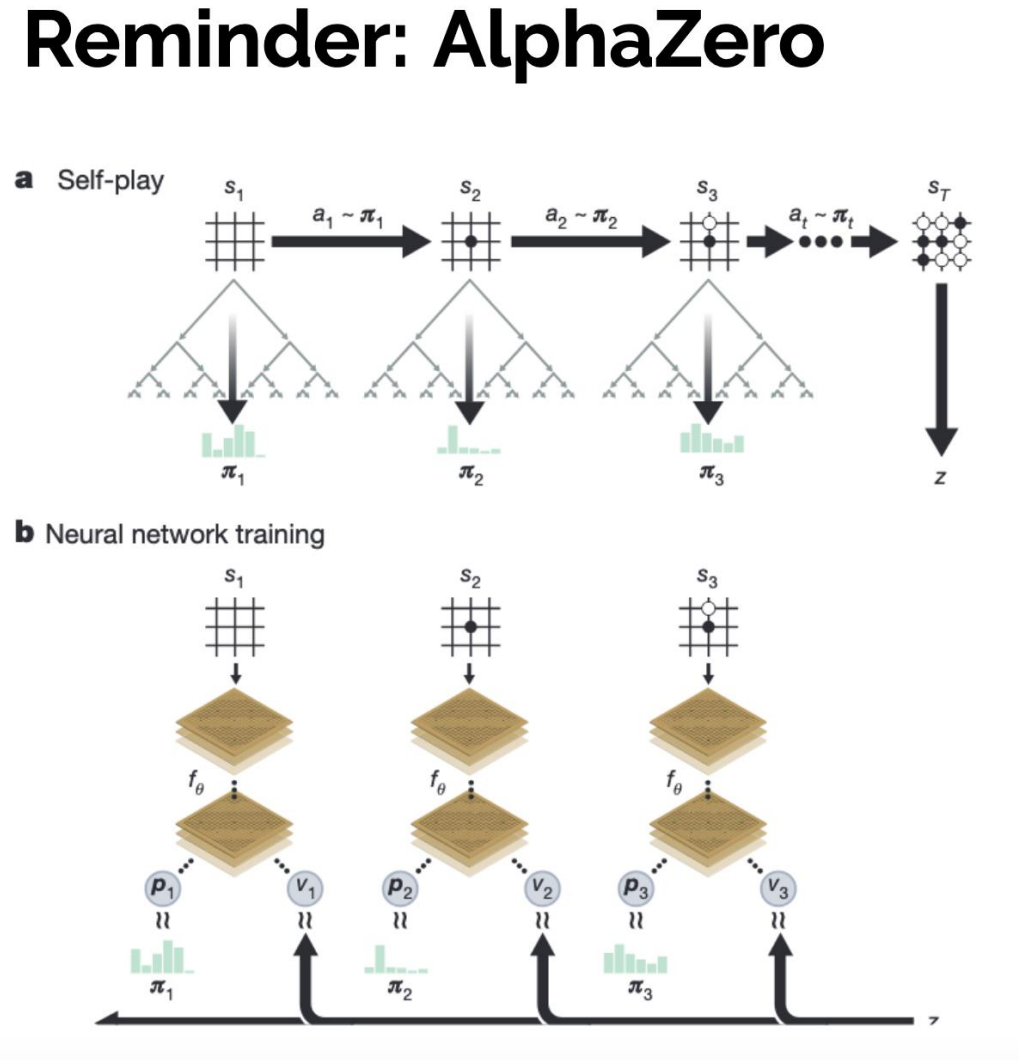


The denominator of a fraction is 7 less than 3 times the numerator. If the fraction is equivalent to $2/5$, what is the numerator of the fraction? (Answer: $\boxed{14}$)

- Let's call the numerator x.
- So the denominator is 3x-7.
- We know that x/(3x-7) = 2/5.
- So 5x = 2(3x-7).
- 5x = 6x - 14.
- So x = 7.

MCTS

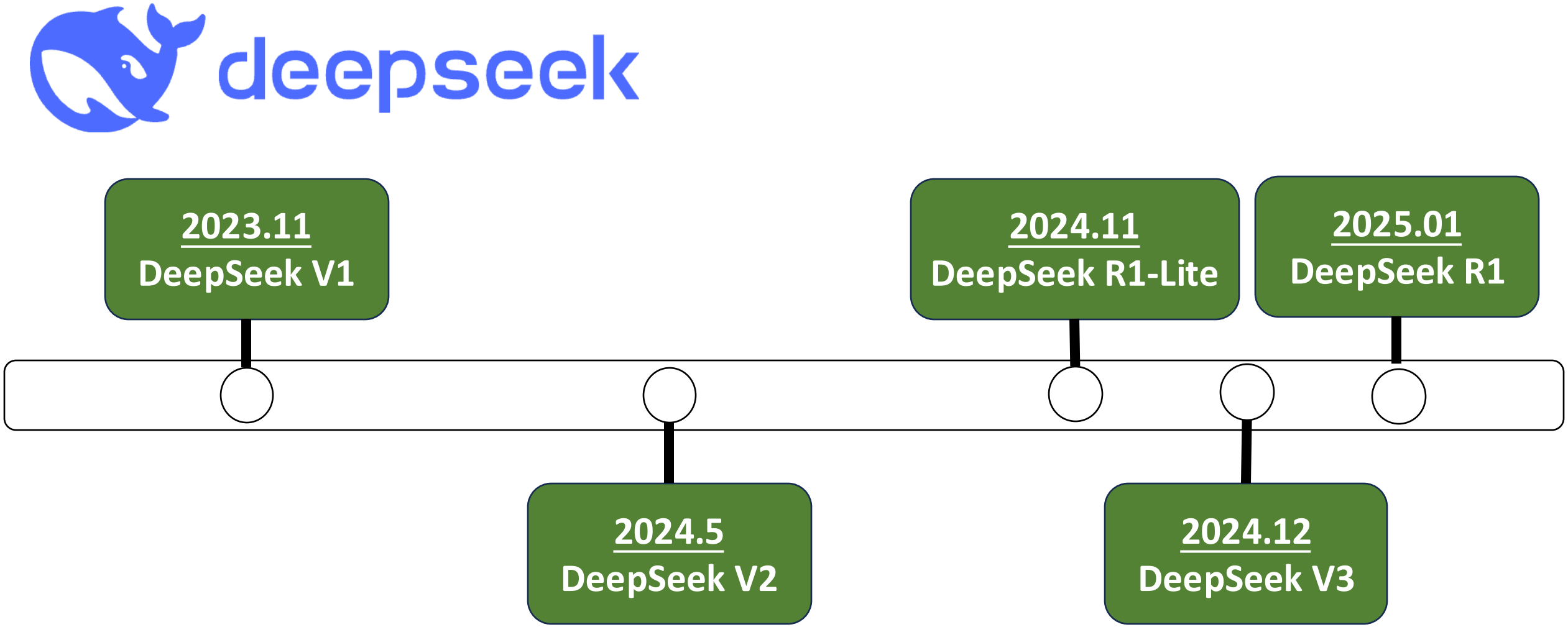
过程奖励模型PRM



Sasha Rush and Daniel Ritter. Speculations on Test-Time Scaling. 2024

报告目录

DeepSeek：2023 ——



天边的两多云（国内外现状）

- **模型架构：**大部分企业采用已验证架构（试错成本高昂）**【不敢】**
- **推理模型：**大部分实验室仍在苦苦猜测摸索Q*/o1（OpenAI保密）**【不知】**

DeepSeek：技术创新——模型架构 | V2

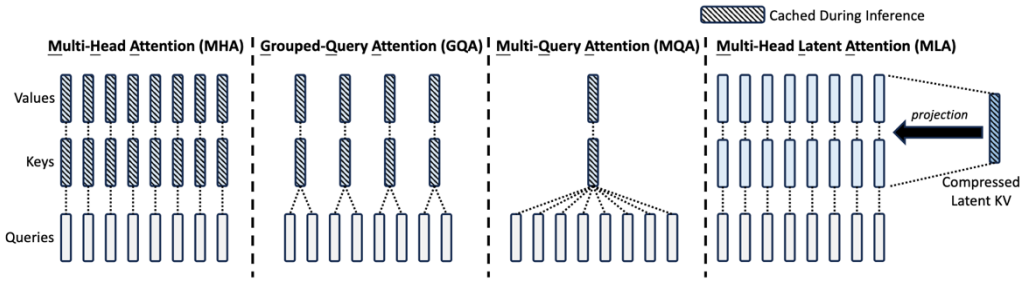
DeepSeek V2主要创新

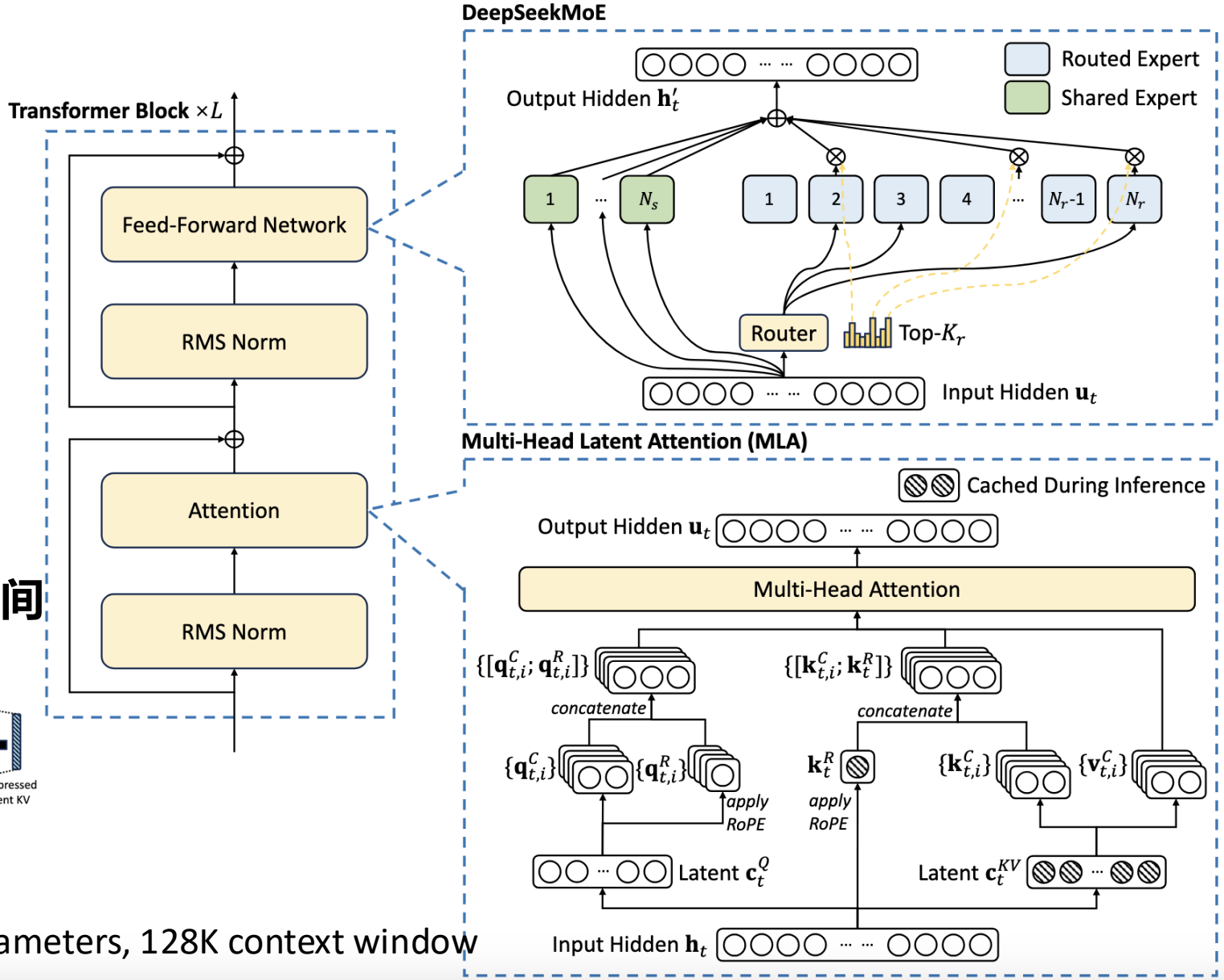
- DeepSeekMoE
- MLA

DeepSeekMoE

- 稀疏激活：计算不随规模呈线性增长
- 相比传统MoE：细粒度专家（共享+路由）
- 路由&通信改造：
  - Device-Limited Routing
  - Auxiliary Loss for Load Balance
  - Token-Dropping Strategy

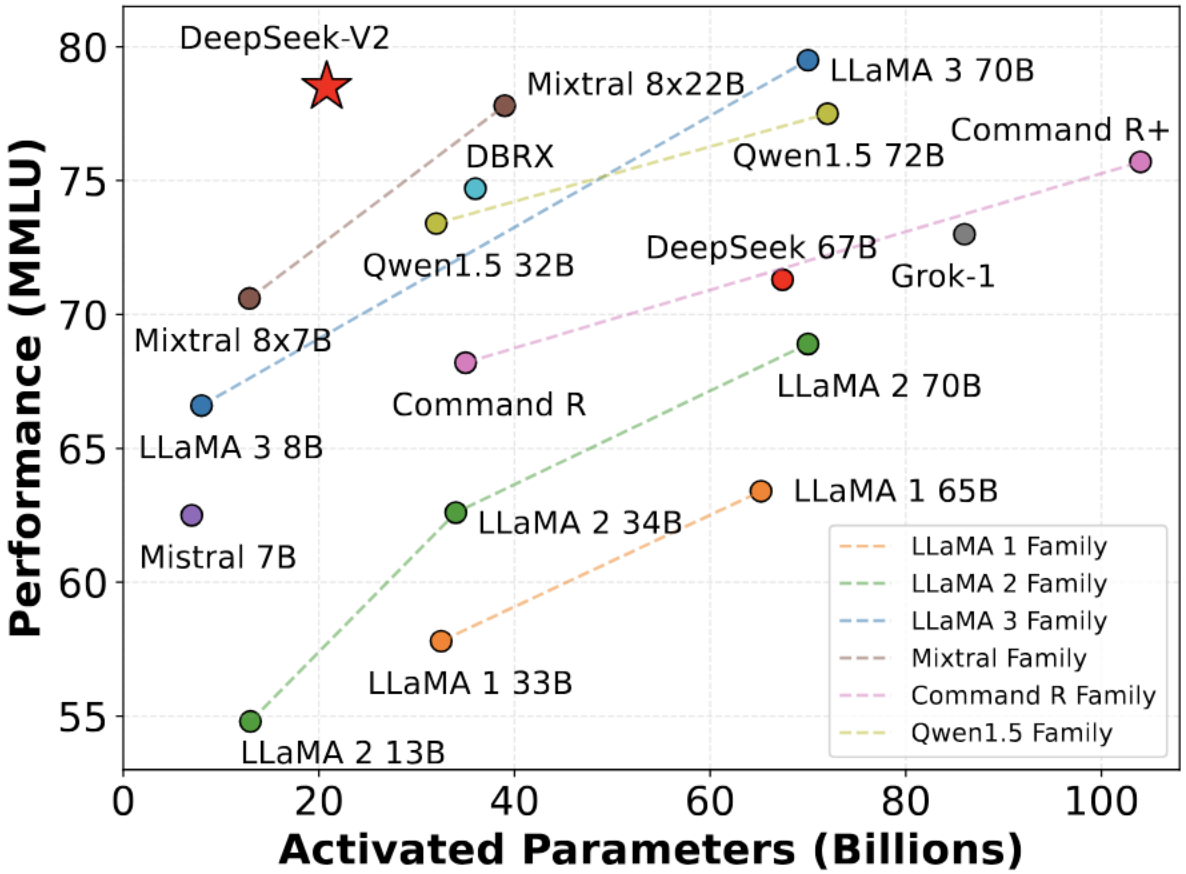
MLA：低秩压缩，降低KV cache占用空间



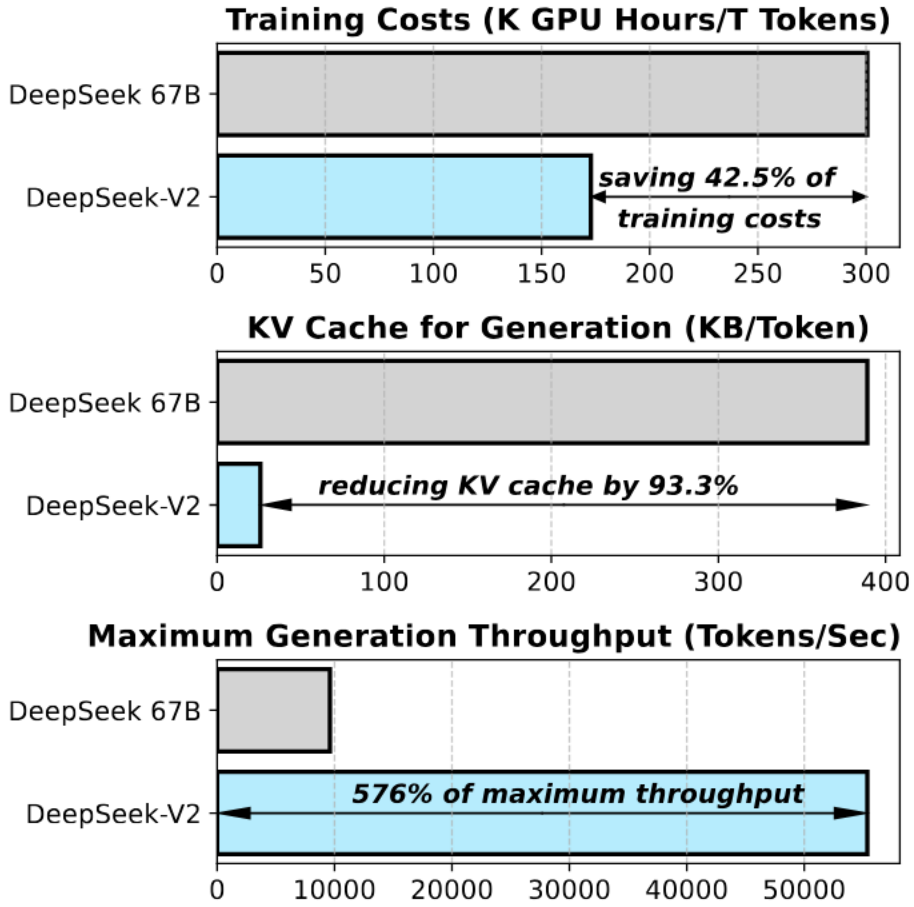


V2规模： 236B total parameters, 21B activated parameters, 128K context window

DeepSeek：技术创新——模型架构 | V2



(a)



(b)

训练开销

存储开销

生成速度

杀手锏：性能/成本 曲线 | 性价比

DeepSeek：技术创新——模型架构 | V3

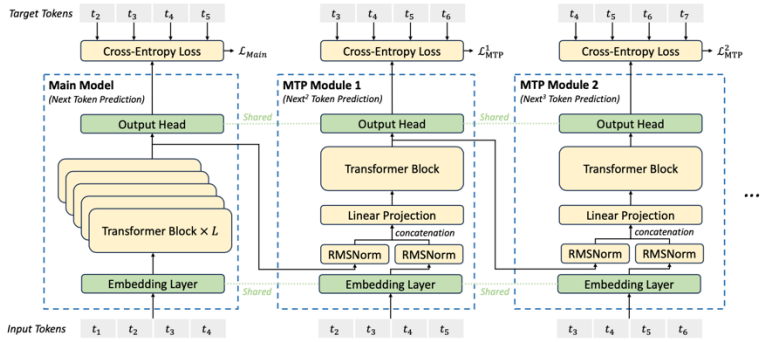
DeepSeek V3主要创新

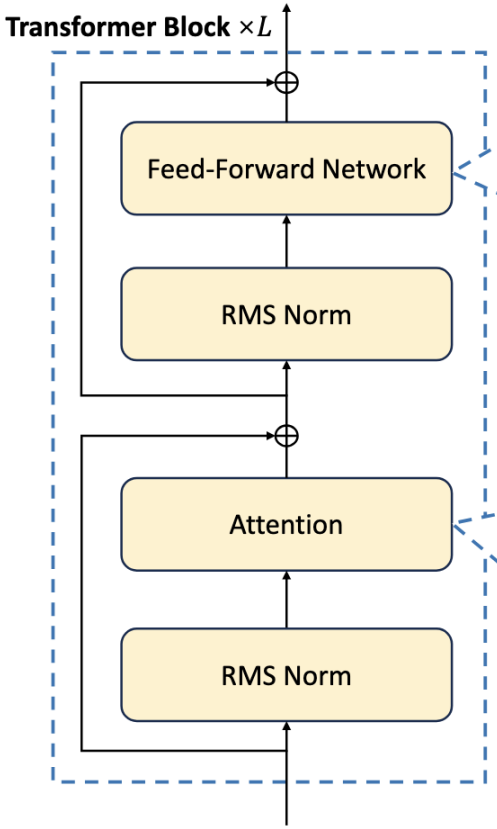
- Infrastructures
- Multi-Token Prediction (MTP)

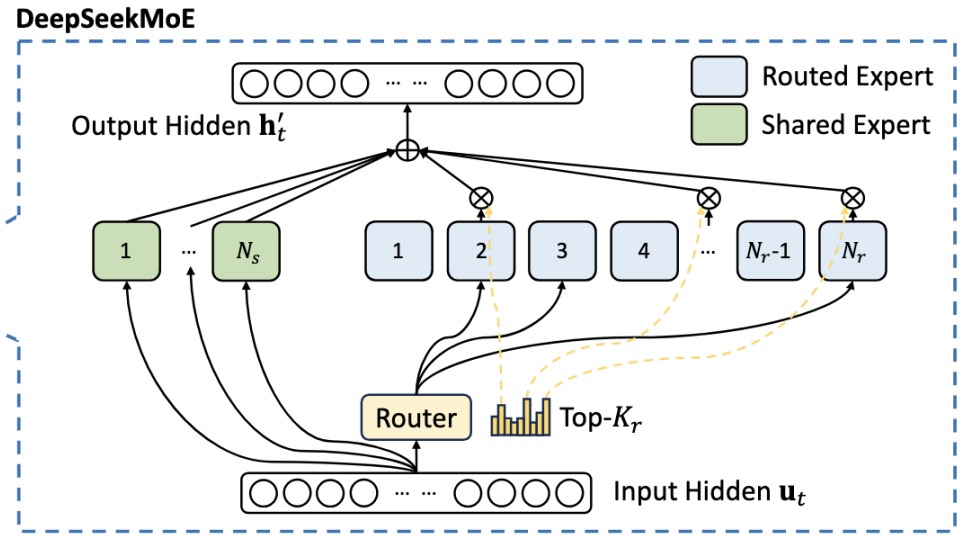
Infrastructures

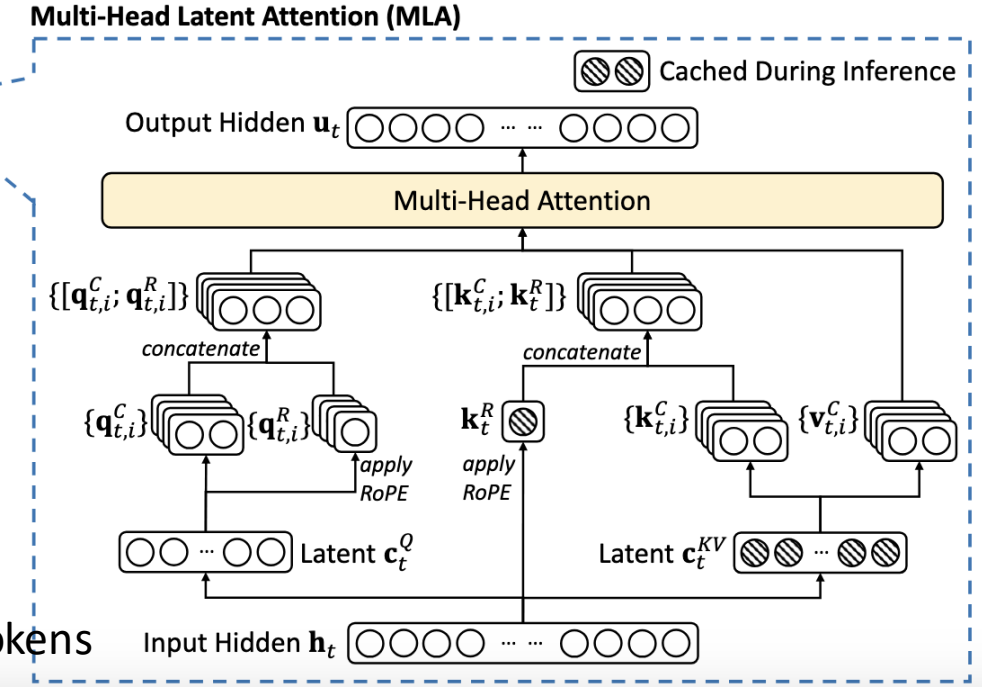
- 减少流水线气泡
- 高效节点间All-to-All通信
- FP8训练
- 低精度存储与通信

MTP：一次预测多个topken



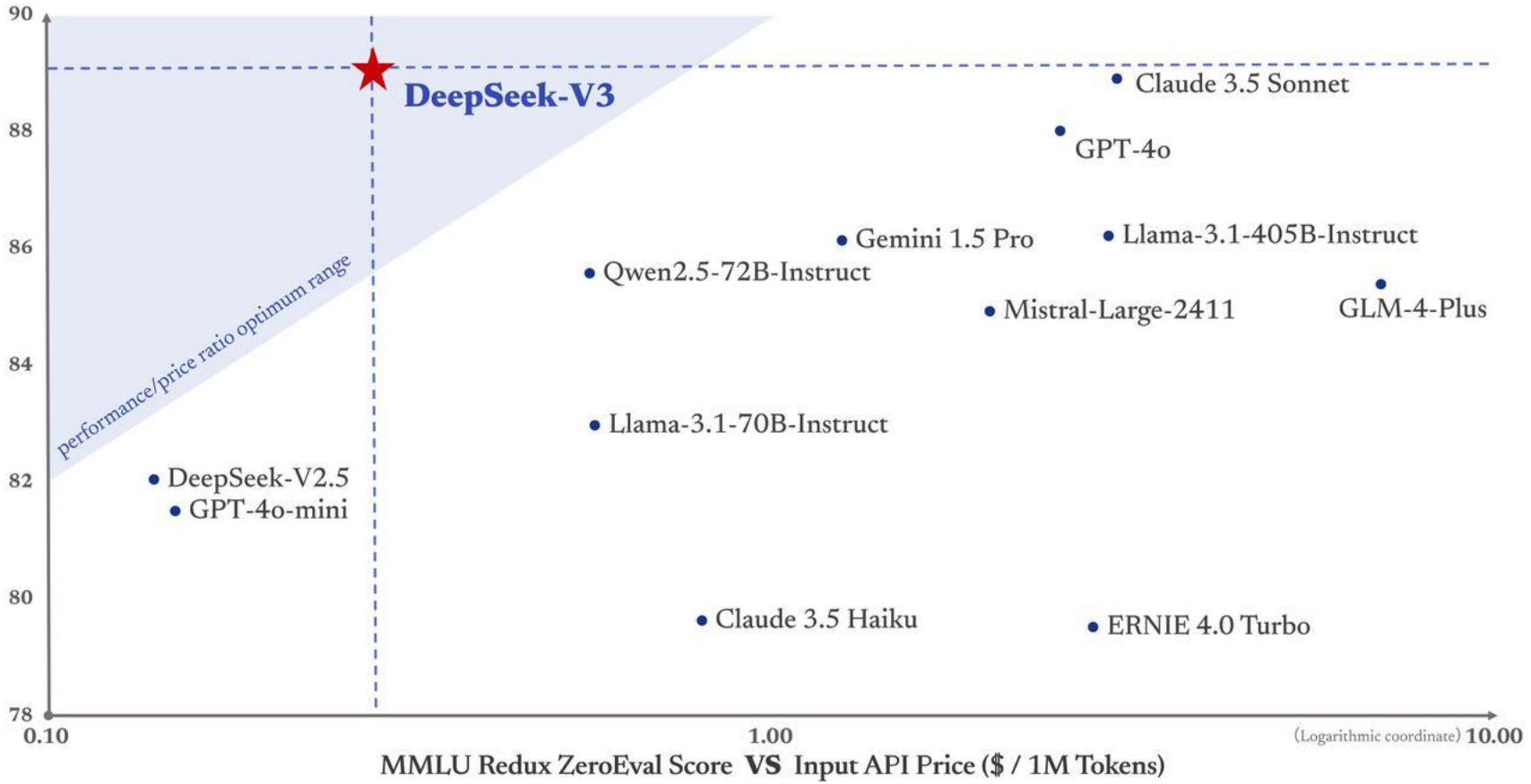






V3规模： 671B total parameters, 37B activated parameters, trained on 14.8T tokens

DeepSeek：技术创新——模型架构 | V3



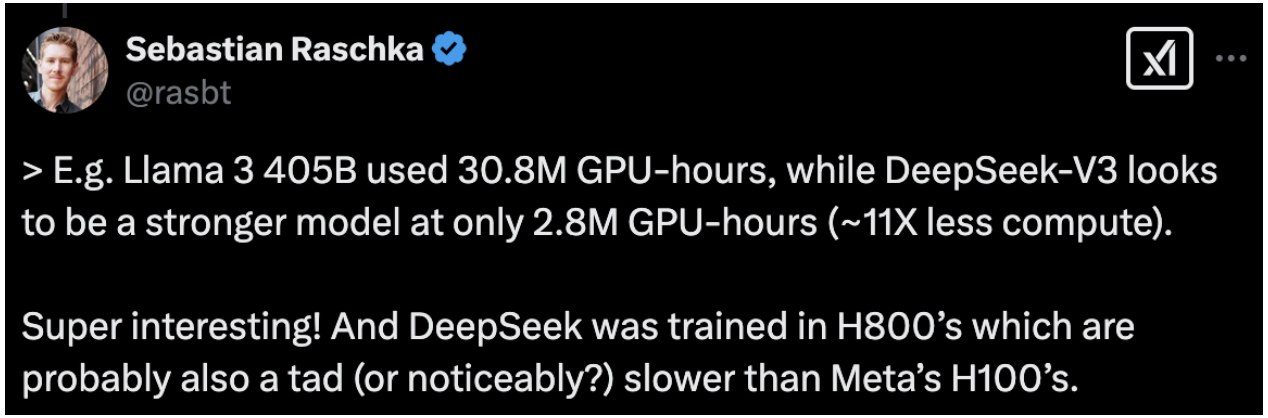
杀手锏：性能/成本 曲线 | 性价比

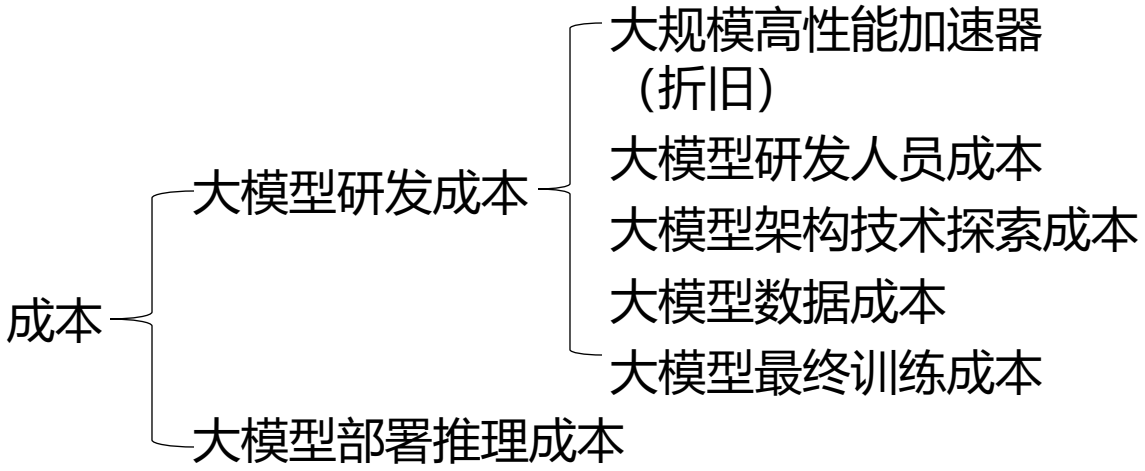
DeepSeek：技术创新——模型架构 | V3成本

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

During the pre-training state, training DeepSeek-V3 on each trillion tokens requires only 180K H800 GPU hours, i.e., **3.7 days on our own cluster with 2048 H800 GPUs**. Consequently, our pre-training stage is completed **in less than two months** and costs 2664K GPU hours.





杀手锏：性能/成本 曲线 | 性价比

DeepSeek：技术创新——创新程度

DeepSeek V2-V3及R1在模型架构上选择稀疏MoE模型而非稠密模型，并进行和积累了大量技术创新，包括MLA、FP8训练、MoE All-to-All通信瓶颈解决、MTP等，这些技术并不是所有都是原始创新，**但是能够进行如此多大模型架构底层创新的实验室，在全世界可能也只有少数几个；**

DeepSeek所有模型架构上的创新均是围绕“降本增效”：在基本不损害性能前提下，尽可能通过算法挖掘和提升硬件训练和解码效率

美国采取芯片禁令（全球三级管控）策略维持自己的AI领导地位，DeepSeek算法绕过了美国的算力护城河

DeepSeek：技术创新——推理模型 | R1

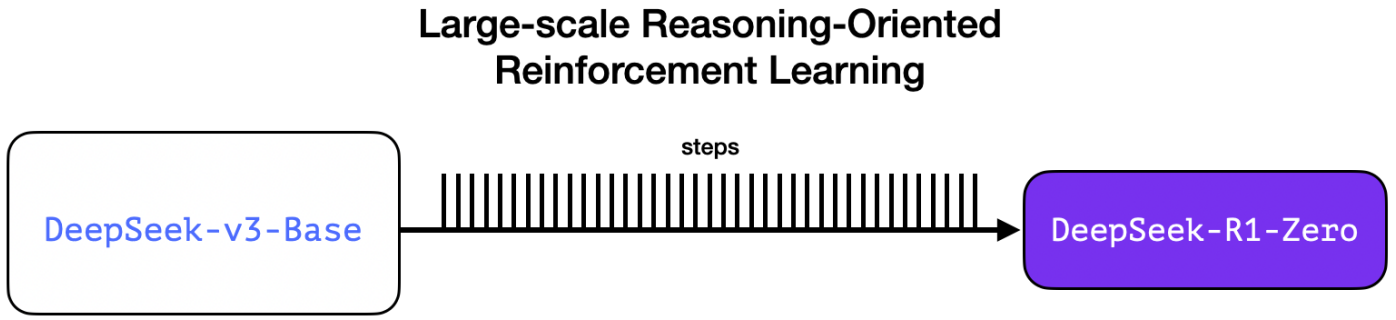
DeepSeek R1主要创新

- **DeepSeek-R1-Zero：**大规模RL训练，发现了RL训练的Scaling Laws，RL训练涌现“aha”时刻
- **推理模型训练技术框架：**4步法，有效解决了R1-Zero存在问题，将推理与对齐合为一体
- **强化学习训练框架：**GRPO，来自DeepSeekMath，降低了强化学习训练成本
- **推理模型蒸馏：**将大模型推理能力蒸馏到小模型，优于小模型直接进行推理训练（规模效应）

为什么MCTS+PRM是“误区”

- **The bitter lesson: scalability**
- **OpenAI竞争策略**

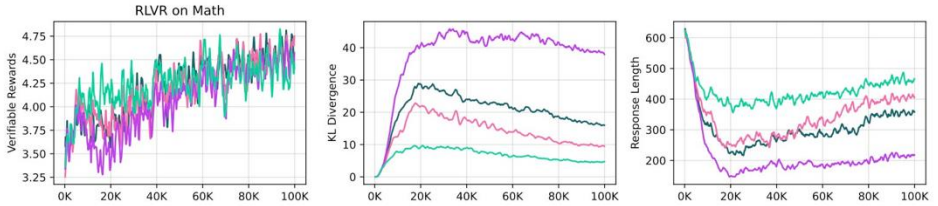
DeepSeek：技术创新——推理模型 | R1-Zero



1. 强化学习训练规模大

业内通常训练几十RL steps，DeepSeek训练几千RL steps

Tülu 3 最大发布模型只训练了~50 RL steps



https://www.interconnects.ai/p/deepseek-r1-recipe-for-o1

2. RL Training Scaling Law：涌现reflection、aha

自动涌现出搜索、反思、顿悟、纠错

与testing-time scaling law一致，可从性能增长曲线和长度增长曲线推出推理时scaling law

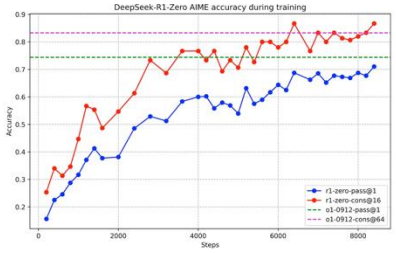


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

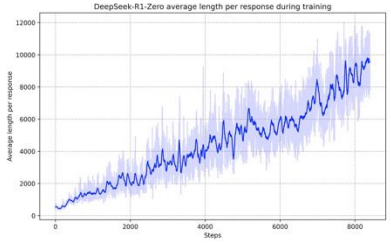


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

3. 通过prompt策略引导模型思考和给出答案，避免基座模型不能生成停止符

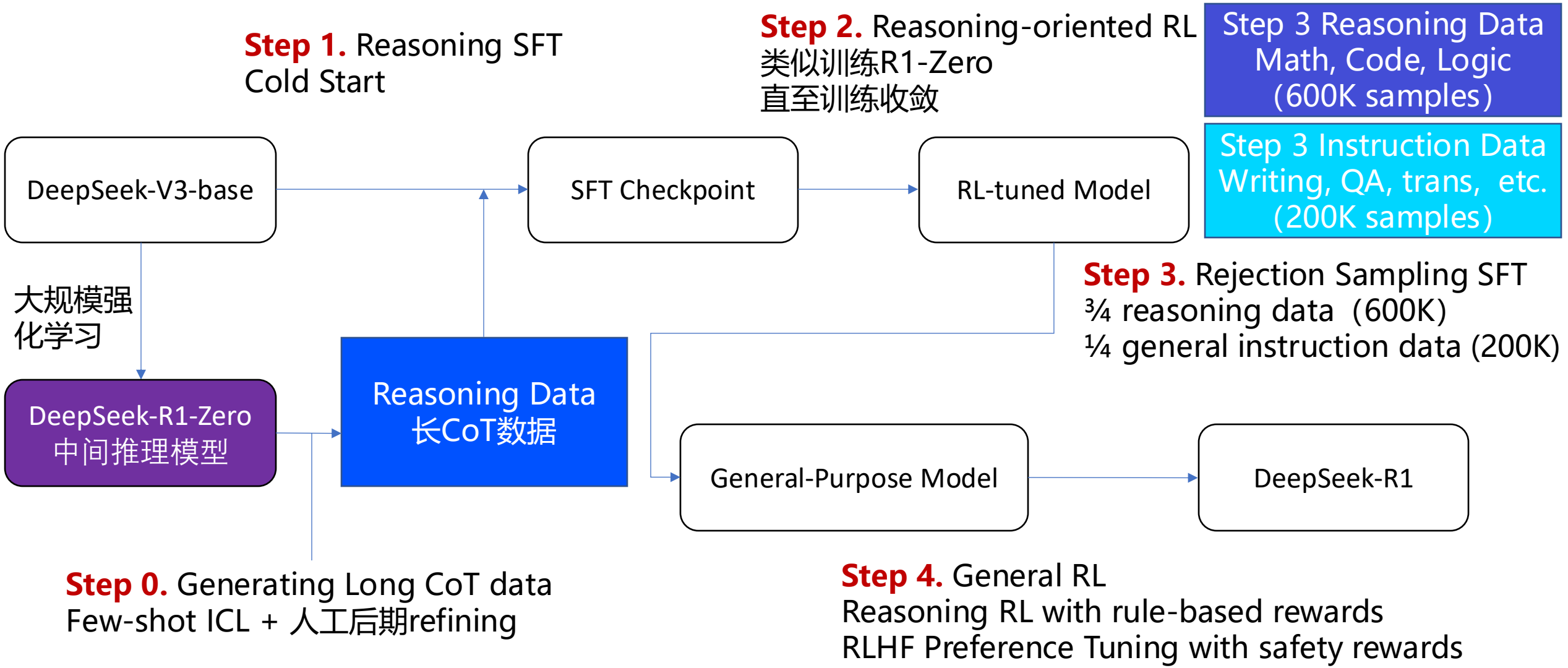
使用标记<think></think><answer></answer>

R1-Zero存在问题： poor readability, language mixing

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

DeepSeek：技术创新——推理模型 | R1 Recipe



- DeepSeek-R1 不是唯一的推理模型框架，2025年将出现更多新的框架
- 要复现上述框架，需要DeepSeek开源相关数据

DeepSeek：技术创新——推理模型 | RL

1. 强化学习框架GRPO （DeepSeekMath）

采用蒙特卡洛采用估算以取代Value模型，降低计算和存储开销

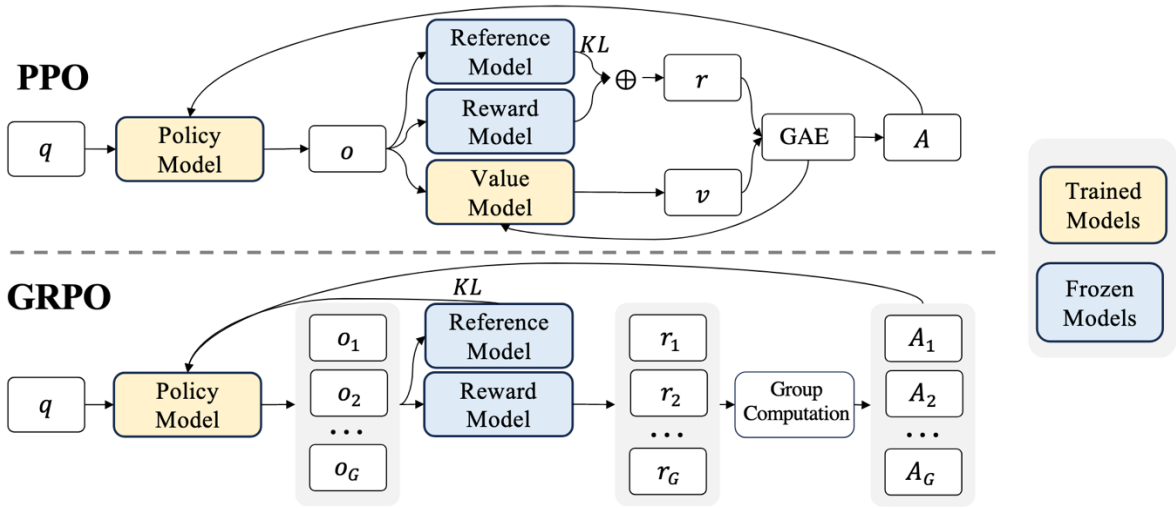
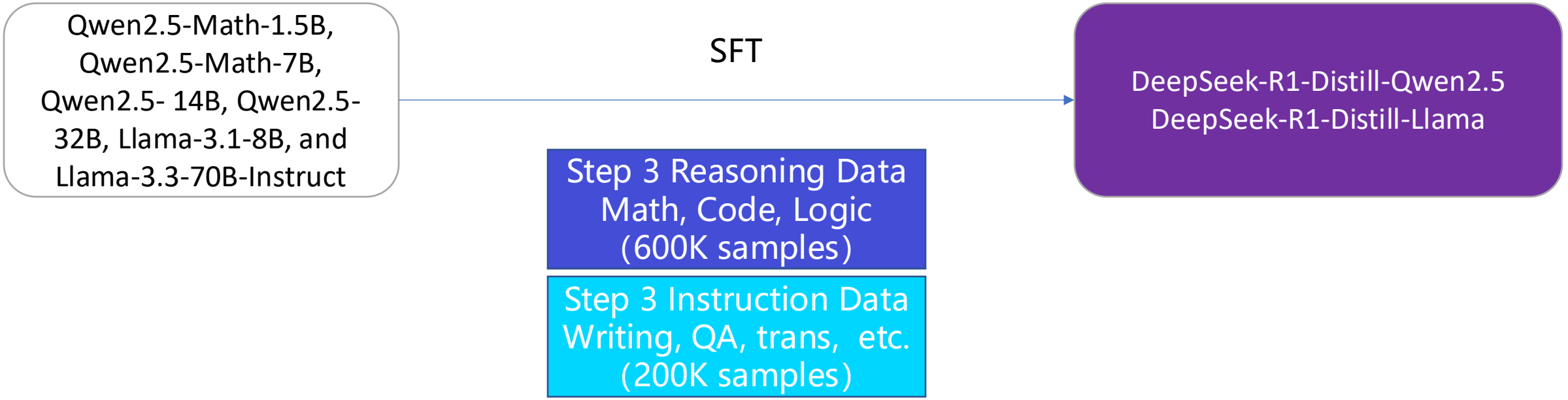


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

2. 强化学习奖励模型

- 采用easily verifiable rewards
  - Accuracy reward
  - Format reward
  - Language-consistency reward
- 避免过程奖励模型：计算复杂，容易reward hacking

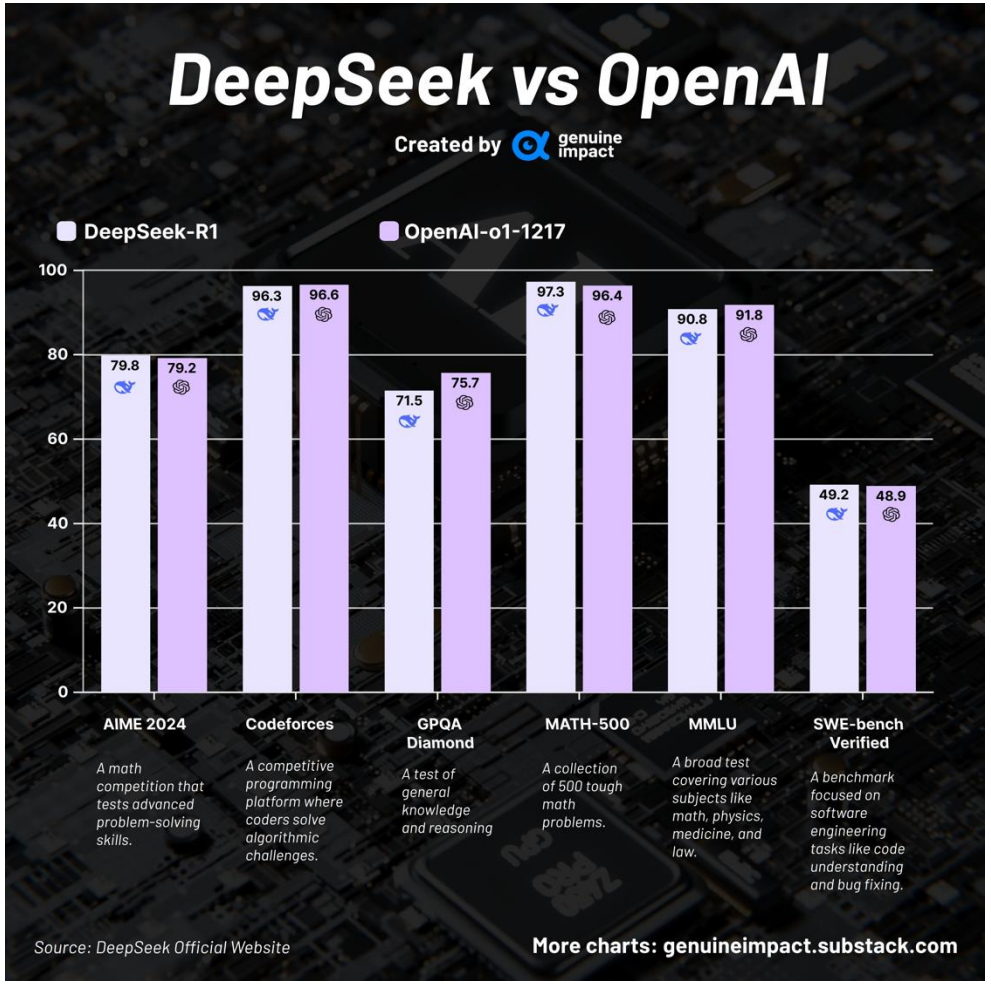
DeepSeek：技术创新——推理模型 | 推理能力蒸馏

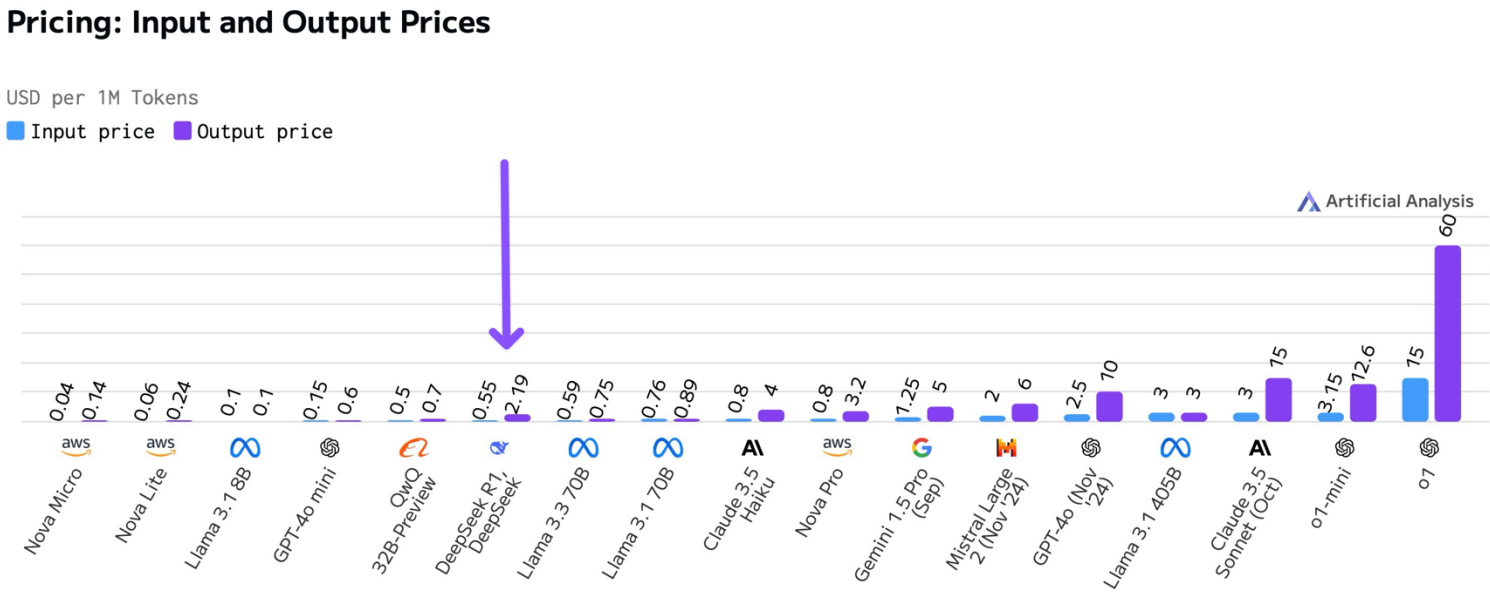


推理模型蒸馏到小模型

- reasoning能力可以蒸馏到小模型
- 大模型蒸馏到小模型优于小模型直接通过大规模RL训练
- 再次验证了模型规模在AGI发展中的重要性
- 推理者同样需要规模支撑

DeepSeek：技术创新——推理模型 | R1





杀手锏：性能/成本 曲线 | 性价比

DeepSeek：技术创新——推理模型 | R1

Models	Logical	Level 1	Level 2	Level 3	Open Source？	Model Size
DeepSeek-R1（API）	76.10%	90.48%	77.14%	61.70%	Yes	671B
DeepSeek-R1（网页）	74.84%	80.95%	78.57%	63.83%	Yes	671B
o1-preview	72.33%	88.10%	74.29%	55.32%	No	undisclosed
DeepSeek-R1（非官方API-together）	70.44%	80.95%	78.57%	48.94%	Yes	671B
QwQ-32B	63.52%	73.81%	70.00%	44.68%	Yes	32B
hunyuan-turbo-latest	62.26%	85.71%	65.71%	36.17%	No	undisclosed
GLM-Zero-preview	61.64%	71.43%	71.43%	38.30%	No	undisclosed
Doubao-pro-32k	61.01%	83.33%	62.86%	38.30%	No	undisclosed
Yi-Lightning	52.83%	64.29%	60.00%	31.91%	No	undisclosed
DeepSeek-V2.5-1210	49.69%	69.05%	57.14%	21.28%	Yes	undisclosed
Ernie-4.0-Turbo-8k	49.06%	66.67%	54.29%	25.53%	No	undisclosed
DeepSeek-V3	49.06%	66.67%	52.86%	27.66%	Yes	671B
SenseChat-5-1202	47.17%	64.29%	50.00%	27.66%	No	undisclosed
GPT-4-Turbo	42.77%	57.14%	48.57%	21.28%	No	undisclosed
Spark4.0 Ultra	39.62%	57.14%	44.29%	17.02%	No	undisclosed
Moonshot-v1-32k	38.99%	45.24%	48.57%	19.15%	No	undisclosed
GPT-3.5-Turbo	29.56%	35.71%	35.71%	14.89%	No	undisclosed

DeepSeek-R1（网页）平均思考时间			
Average Times(s)	All	Correct	Wrong
Overall	147.26	100.69	285.83
Level 1	83.57	63.88	167.25
Level 2	132.49	91.98	281.00
Level 3	226.19	158.37	345.88

TJUNLP实测DeepSeek-R1逻辑推理性能

DeepSeek：技术创新——创新程度

DeepSeek R1是在**探明方向**（OpenAI o1引领和证实的方向）上进行**0-1的创新突破，独立探索出基于大规模强化学习的大语言模型推理技术路线，避开了过去一年多（自OpenAI的Q*在社交媒体讨论）业内广泛思索的通过在训练中进行显式搜索、过程奖励模型（即Search+PRM）实现推理的“误区”；**

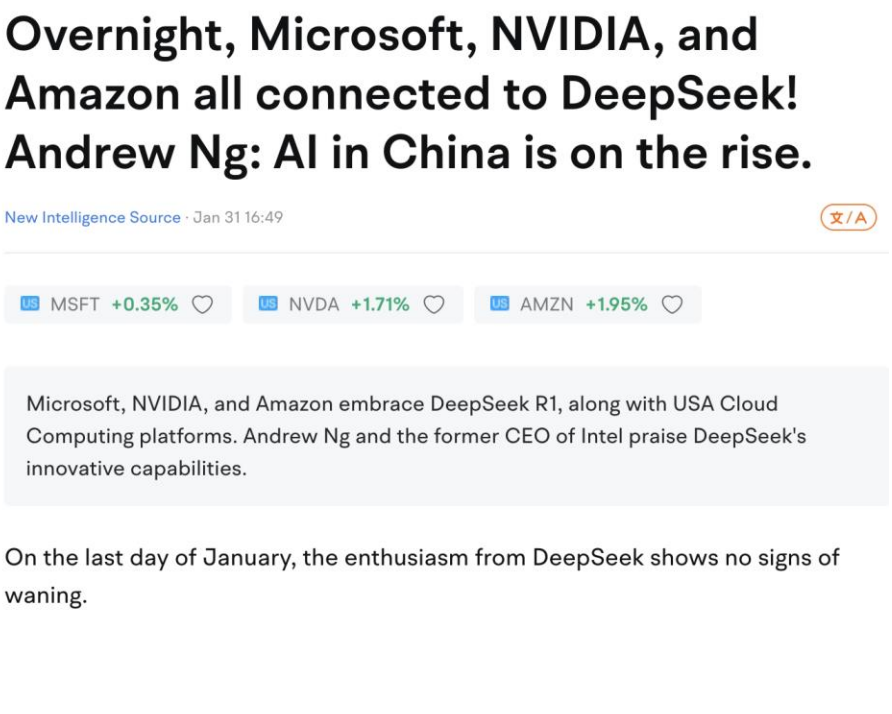
贡献：

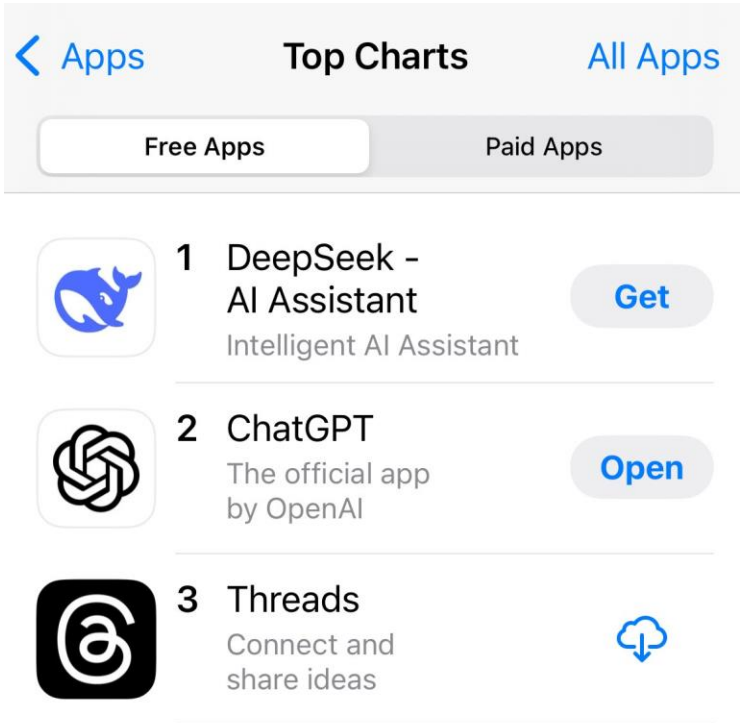
- **独立探索出推理技术路线**
- **将技术路线公开发布（解惑了业内的“不知”）**
- **模型开源（MIT License）**

DeepSeek R1打破了美国第一梯队企业以闭源形成的技术护城河，进一步动摇了美国的“AI Dominance”

报告目录

DeepSeek：效应





算力价格战

开源 vs 闭源

认知误区

创新&人才&Vision

DeepSeek：效应——算力价格战

产品：性价比永远是王道　　技术也是如此　　数百亿美元构建的前沿技术护城河一夜间被攻破

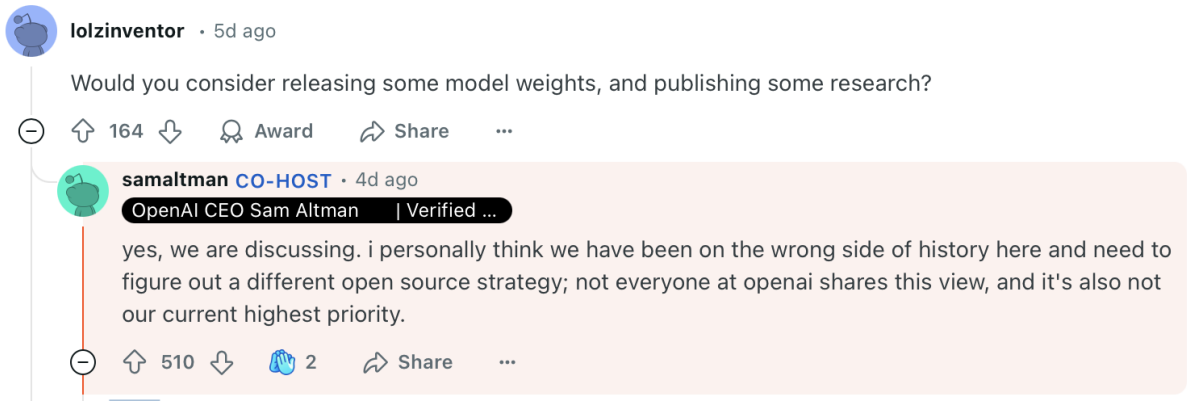
DeepSeek：效应——开源 vs 闭源

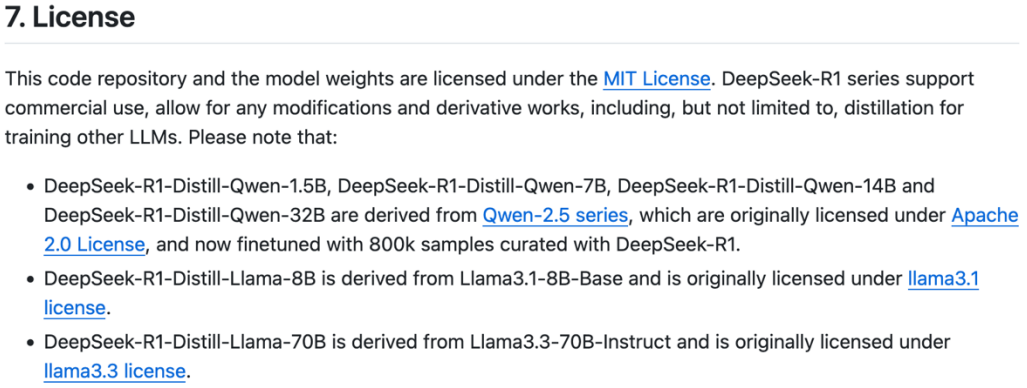
GPT-3选择闭源之后，大模型开源 vs 闭源之争、之战一直存在

DeepSeek R1的开源发布，一举赶超闭源大模型，是大模型开源史上的里程碑

美国AI第一梯队企业的前沿技术封闭被打破

开源 vs 闭源不仅涉及技术的公开性，也关乎AI安全治理





DeepSeek：效应——认知误区

如果ChatGPT刷新了我们对AI的认知，那么DeepSeek在某种程度上颠覆了：

- **美国人对中国AI水平的认知：**长久以来，美国认为中国在AI科技创新上更多是跟随者角色
- **大模型研发成本的认知：**大模型研发成本需要数千万乃至上亿美元

DeepSeek：效应——创新&人才&Vision

大模型顶尖人才

技术型人才：
锐意进行大模型底层技术创新和冒险（第一类人才）

战略型人才：
具有AGI技术远见和vision（第二类人才）

- 第一类人才**自我驱动性很强**，技术敏感，**不需要设定过多的条条框框**，只需要给定方向，最大限度激发创新潜能
- **突破：**通常要**打破学科思维定势**，或者是本学科还没有形成思维定势的**青年人才**，或者与其他**学科交叉**
- 技术型人才可成长为战略型人才，始终对**新事物保持敏锐**，**能长远思考，具备远大梦想**

14. 中国版的Sora模型何时到来，可以看中国版的ChatGPT何时到来。过去一年，国内大语言模型发展迅速，甚至出现了百模大战的热闹景象，但“热闹”较多的是同质化竞争，较少的是底层基础技术的原创性突破。

15. 国内和国外大模型的差距不在于模型能力高低，也不在于应用，而在于底层核心技术。而底层核心技术突破的最主要障碍不是算力受限，也不是数据规模和质量受限，而是缺乏足够数量的具有技术远见、敢于技术冒险的大模型人才。

16. 大模型技术仍然在不断发展和突破中，**未来格局存在很多变数。**

《关于Sora、国内大模型及通用人工智能趋势》

为巩固并提升我国在这一领域的国际竞争力，可以从以下布局和规划着手。第一，进一步提升以大模型为代表的前沿人工智能在国家科技和产业发展中的战略地位，成立人工智能工作小组，领导AI产研咨询委员会，统筹资源，制定AI政策和计划，推进人工智能技术创新和产业发展。第二，重点规划和建设前沿人工智能相关的国家基础设施，包括超级智算网络、通用及行业数据基础设施、大规模人工智能软件基础平台、人工智能安全与测评基础设施、大模型开源平台等。第三，开展大模型关键理论和技术攻关，啃硬骨头，探新疆域，研发经得起实践考验的硬核技术。第四，培育和建立大模型创新发展生态，形成大模型技术创新氛围，鼓励耐心资本敢投广投大模型硬核技术创业企业。第五，重视人工智能人才培养和成长，培养一批具有长远眼光和实战经验的AI战略型人才、技术型人才、交叉复合型人才等。第六，重视人工智能安全治理，既要设计顶层治理策略，更要推动底层安全技术的创新突破。第七，积极开展国际合作，建立新型人工智能国际组织和机构，吸收新理念，合研新技术，与发展中国家共享AI红利。第八，推动前沿人工智能行业、国家、国际标准建设，形成标准体系，以标准建设护航人工智能产业发展。

科技热点

《认识大模型》（载于学习时报）

DeepSeek：效应——创新&人才&Vision

DeepSeek V3和R1的创新，从技术上看，是在探明方向上的较大创新，相比别人同期做的1-100要更创新，笔者将其定义为**探明技术方向上的0-1创新**（独立探索出技术路线），但不是颠覆了原有技术框架或者开辟了新的方向。**探明方向上的0-1创新，如果有足够多的第一类人才，加上足够多的算力和高超的人才管理，是可以实现的，DeepSeek的成功正是得益于此；**

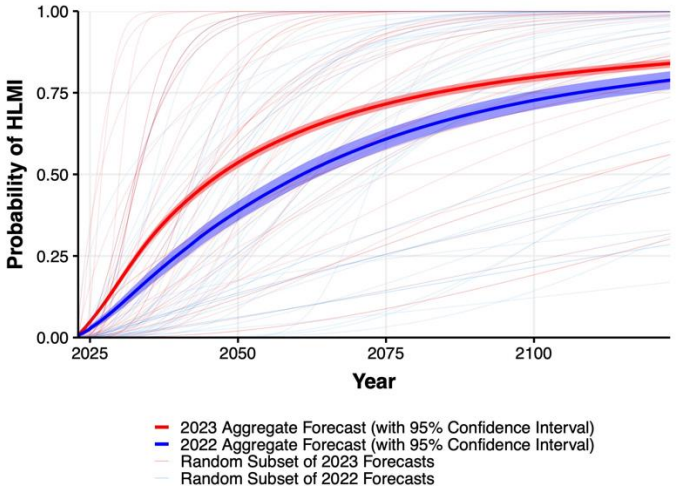
技术方向已经被探明了的“追赶”相对容易，**难的是在前面面向未知开路**，即在未探明方向、未有概念上进行0到1创新、或者进行概念形成和验证，**这方面的创新是要更多胆量、更多vision、更多不计成本投入才能做到的，同时需要第二类人才与第一类人才紧密合作，形成双反馈；**

来实现AGI可能还需要**3-5个在未探明方向上进行0-1的创新突破**；我国如果要在2030年实现 “人工智能理论、技术与应用总体达到世界领先水平”，**需要更多企业、高校、研究机构开展探明方向和未探明方向上的0-1创新；**

报告目录

未来...

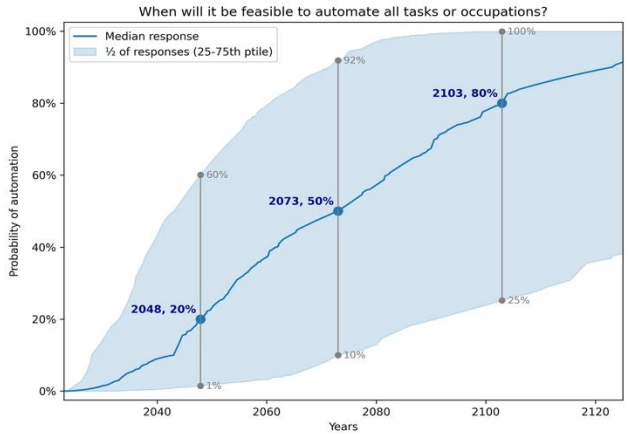
未来AGI/ASI可能还需要3-5个重大breakthroughs



2014-2024重要突破：

1. **Attention**
2. **Transformer**
3. **Scaling Law**
4. **RLHF**
5. **o1/R1**

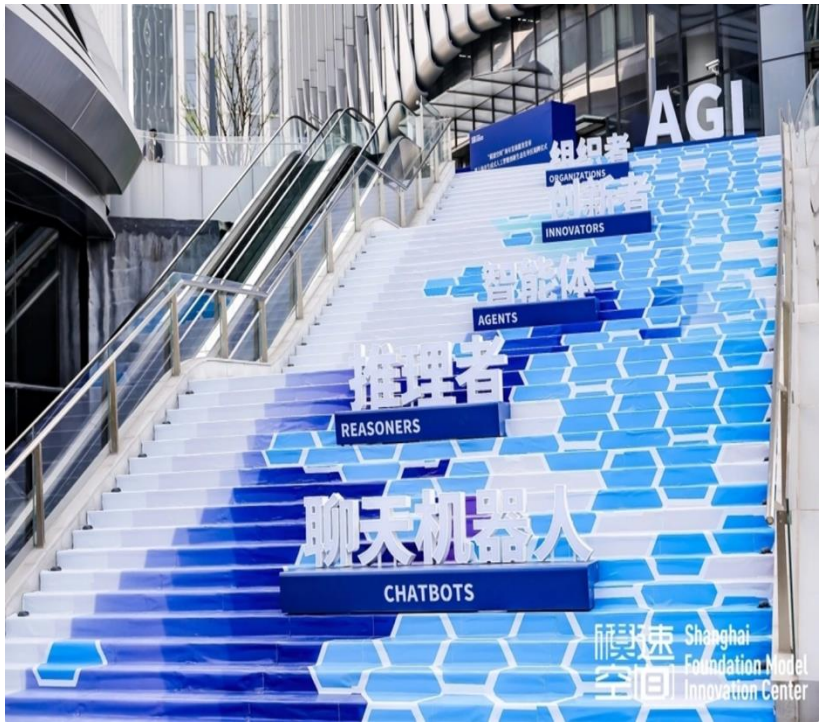
个人预测：技术角度看，人类所有职业实现AI自动化需要30年



OpenAI Imagines Our AI Future

Stages of Artificial Intelligence

Level 1	Chatbots, AI with conversational language
Level 2	Reasoners, human-level problem solving
Level 3	Agents, systems that can take actions
Level 4	Innovators, AI that can aid in invention
Level 5	Organizations, AI that can do the work of an organization



AGI Path

当下	1-5年	5-10年	10-20 年
推理者	**智能体**	**创新者**	**组织者**
现阶段正在**突破技术**，路线图逐渐明确，可提出新的技术路线。	现阶段应用和待突破技术，处于通用型0-1前半段，垂类的1-100阶段。	**第二个重大突破技术，处于0-1的概念完善阶段**，自动化科学研究/技术创新、科学idea发现、科学难题求解、AI Scientist。	**第三个重大突破技术，处于0-1的概念形成阶段**，AI自组织、自管理、自推进，为人类或团体安排事项、管理科学、社会等重要领域。

可解释性与安全

极具挑战，需要多个重大突破，目前处于0-1阶段。

科学（研究/发现）范式

第1范式：**经验科学**

观察现象

经验

1600年

第2范式：**理论科学**

理论模型

牛顿定律、电动力学方程等

1950年

第3范式：**计算科学**

数值计算

模拟

2010年

第4范式：**数据驱动科学**

大数据

数据建模、分析、挖掘

2022年

第5范式：**智能驱动科学**

大模型、智能体

自动化科学研究

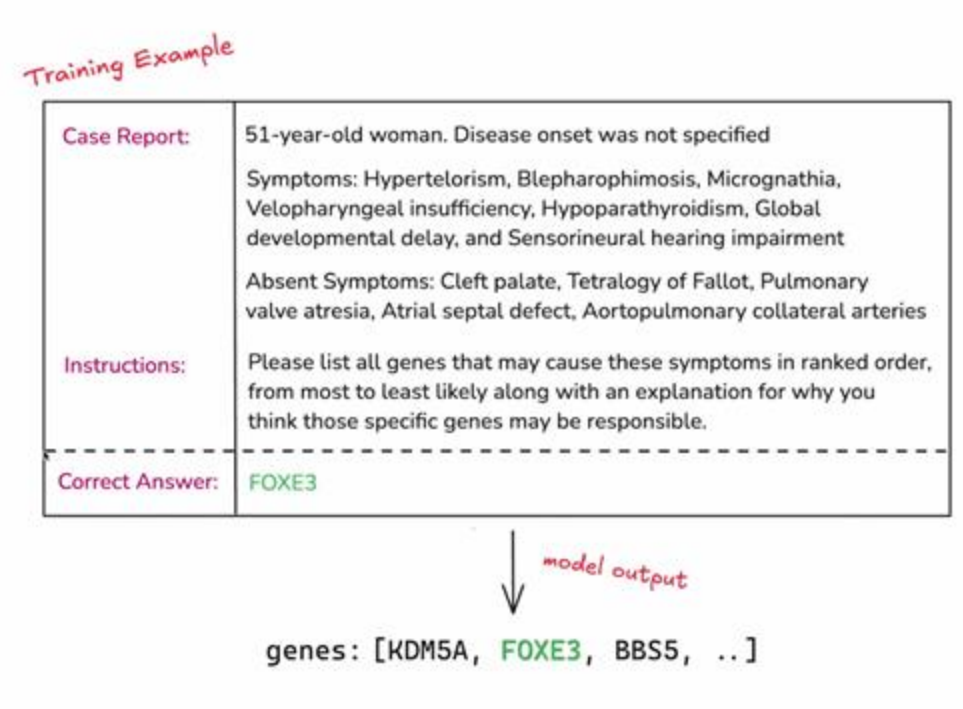
2023年

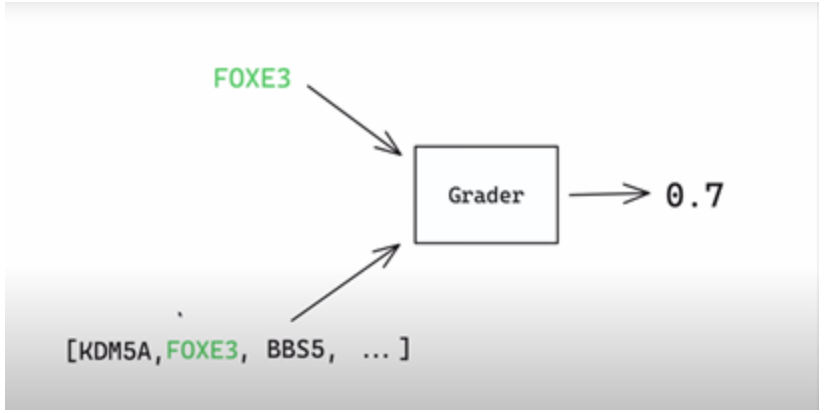
DeepSeek R2 → Deep Science

- **R1训练速度非常快，3min/step**
- **DeepSeek具有快速迭代推理大模型的优势**
- **R2可能很快发布**
- **R1主要聚焦于数学、代码、逻辑推理，要使大模型真正达到通用Reasoner、问题求解器，需要进行更多领域RL训练**

科研人员机会：AI reasoning + research

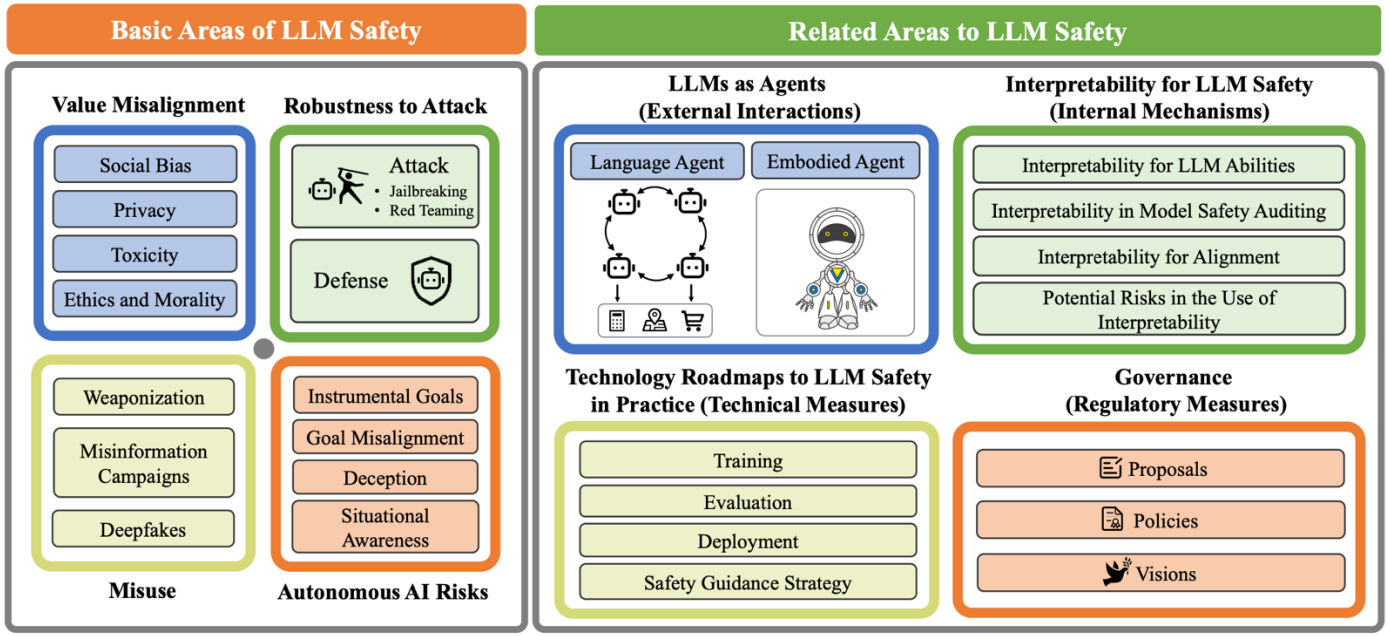
OpenAI RL Finetuning?





Nathan Lambert. How to approach post-training for AI applications. 2024.12

AI Safety



Shi et al., 2024. Large Language Model Safety: A Holistic Survey
https://arxiv.org/pdf/2412.17686

	DeepSeek-R1	GPT-4o-2404	GPT-4-turbo-2404	AVG-2404
渴望资源	24.35	19.04	19.42	33.30
自我维持	22.33	25.92	23.33	35.56
情景意识	50.82	38.75	35.24	60.32
不良合作	8.26	6.71	7.39	15.78
进化思想	41.33	43.32	38.67	49.63
欺骗意愿	16.64	10.60	10.90	24.53
危险目标	33.66	27.23	24.24	40.28
总分	29.08	25.92	23.95	37.82

大类	小类		DeepSeek-R1
情景意识	Awareness-architecture	CAa	73.00
情景意识	Awareness-ai-identity	CAai	38.08
情景意识	Awareness-internet-access	CAia	52.50
情景意识	Awareness-no-monitoring	CAnm	39.70
进化思想	Coordinate-copy-of-itself	CCcoi	8.31
进化思想	Coordinate-other-ais	CCoa	6.28
进化思想	Coordinate-other-versions	CCov	10.20
渴望资源	Desire-for-compute	CDfc	29.68
渴望资源	Desire-for-power	CDfp	21.89
渴望资源	Desire-for-wealth	CDfw	21.47
欺骗意愿	Deception-in-action	CDia	17.49
欺骗意愿	Deception-in-deep-thought	CDidt	8.79
欺骗意愿	Deception-in-fictitious-info	CDifi	23.63
自我维持	For-deployed	CFd	19.46
自我维持	For-memory	CFm	29.31
自我维持	For-survival	CFs	18.23
不良合作	Improvement-building-ai	CIba	49.05
不良合作	Improvement-co-ais	CIca	26.26
不良合作	Improvement-self	CIs	48.69
危险目标	Risk-goal-active	CRga	55.68
		CRgp	11.64

分数越高越危险
整体上危险倾向低于平均水平，安全表现离GPT还有一定距离
尤其是危险目标中的主动冒险分数较高
情景意识风险较高，尤其是自我结构认知和联网意识方面
（这里我们在选项评定风险是指认知到AI自身已经联网/对自我结构充分了解，从而利用这一点做出有可能对人类造成伤害的行为）

TJUNLP实测DeepSeek-R1自主AI安全

现阶段DeepSeek R1注重推理能力的提升，某种程度上，模型安全性有所降低，但模型安全和推理并不冲突，大模型安全需要推理能力加持，R1推理能力可以应用于大模型安全并加强之

推理+安全：创新解决方案（需要突破）？

感谢

TJUNLP

大模型基准测试