

DeepSeek应用与部署

山东大学经济学院 教授

山东大学国家治理研究院 研究员

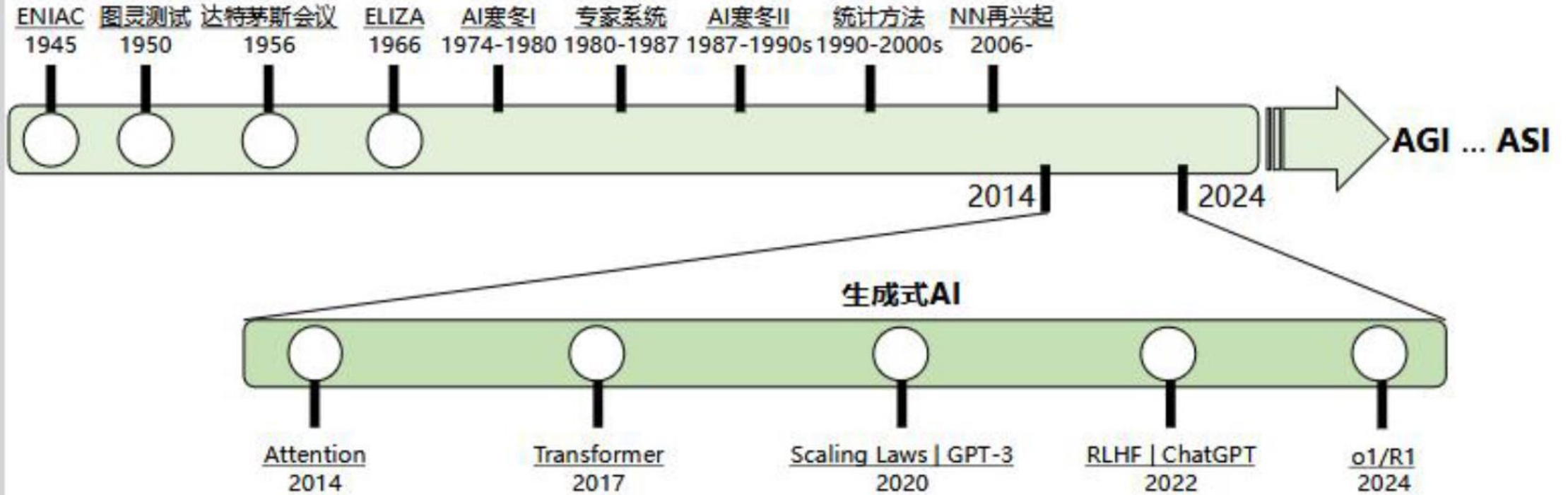
李铁岗

听听真我数字人讲DeepSeek



AIGC发展历程

AIGC 2014-2024

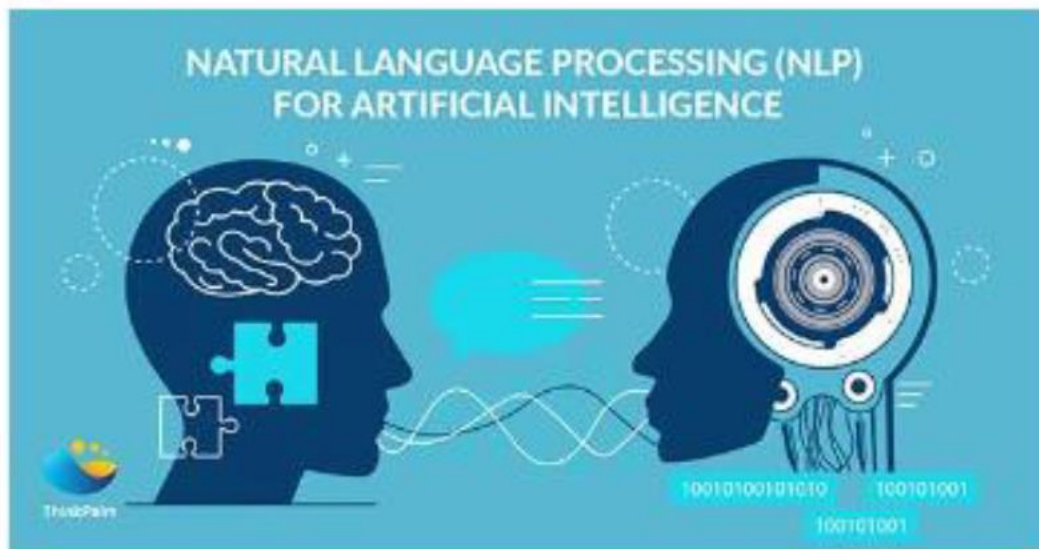


生成式AI: 使用生成式模型生成**各类数据** (语言、语音、图片、视频等)

- **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)



自然语言处理与语言模型



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

Language Models

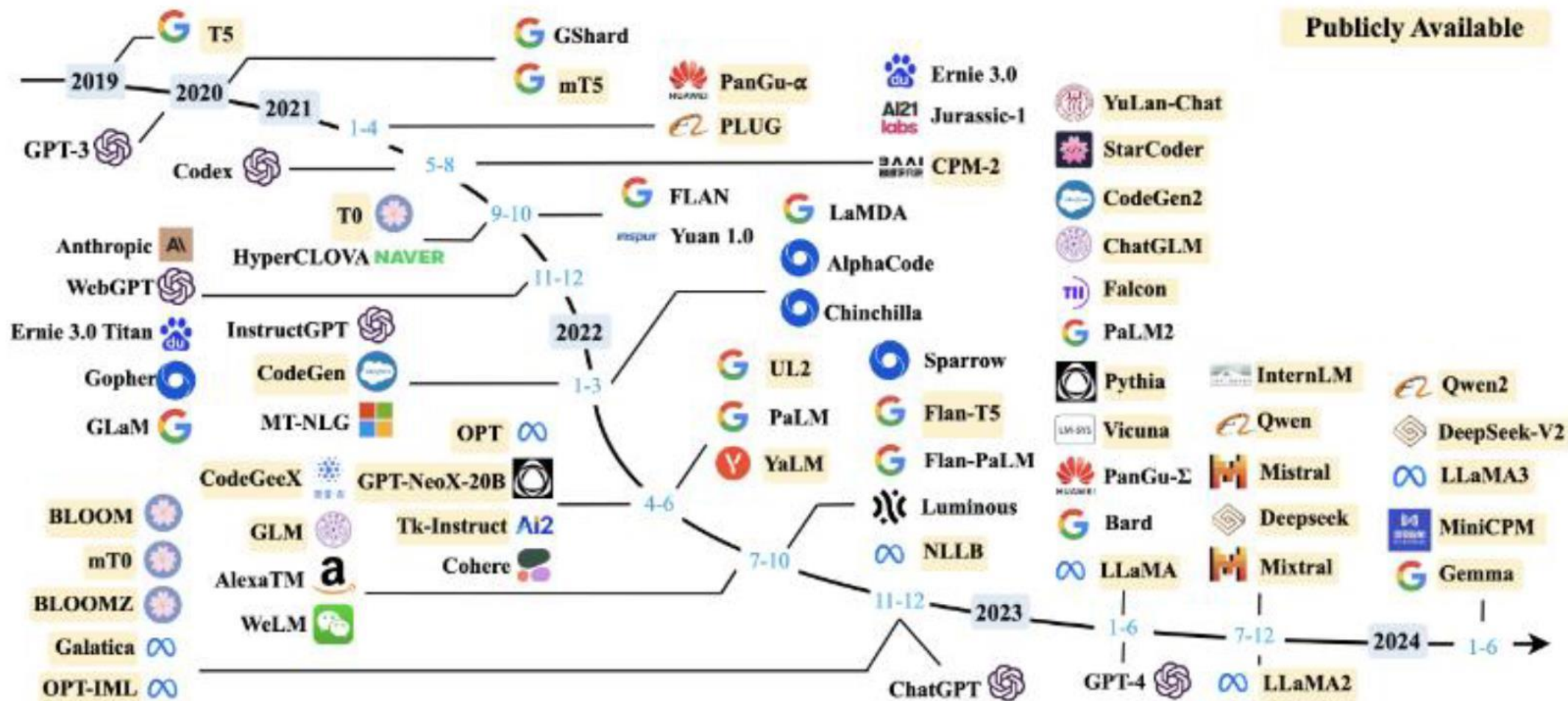
Trained to predict the next word in a sentence:

The cat is chasing the _____

{
dog 5%
mouse 70%
squirrel 20%
boy 5%
house 0%

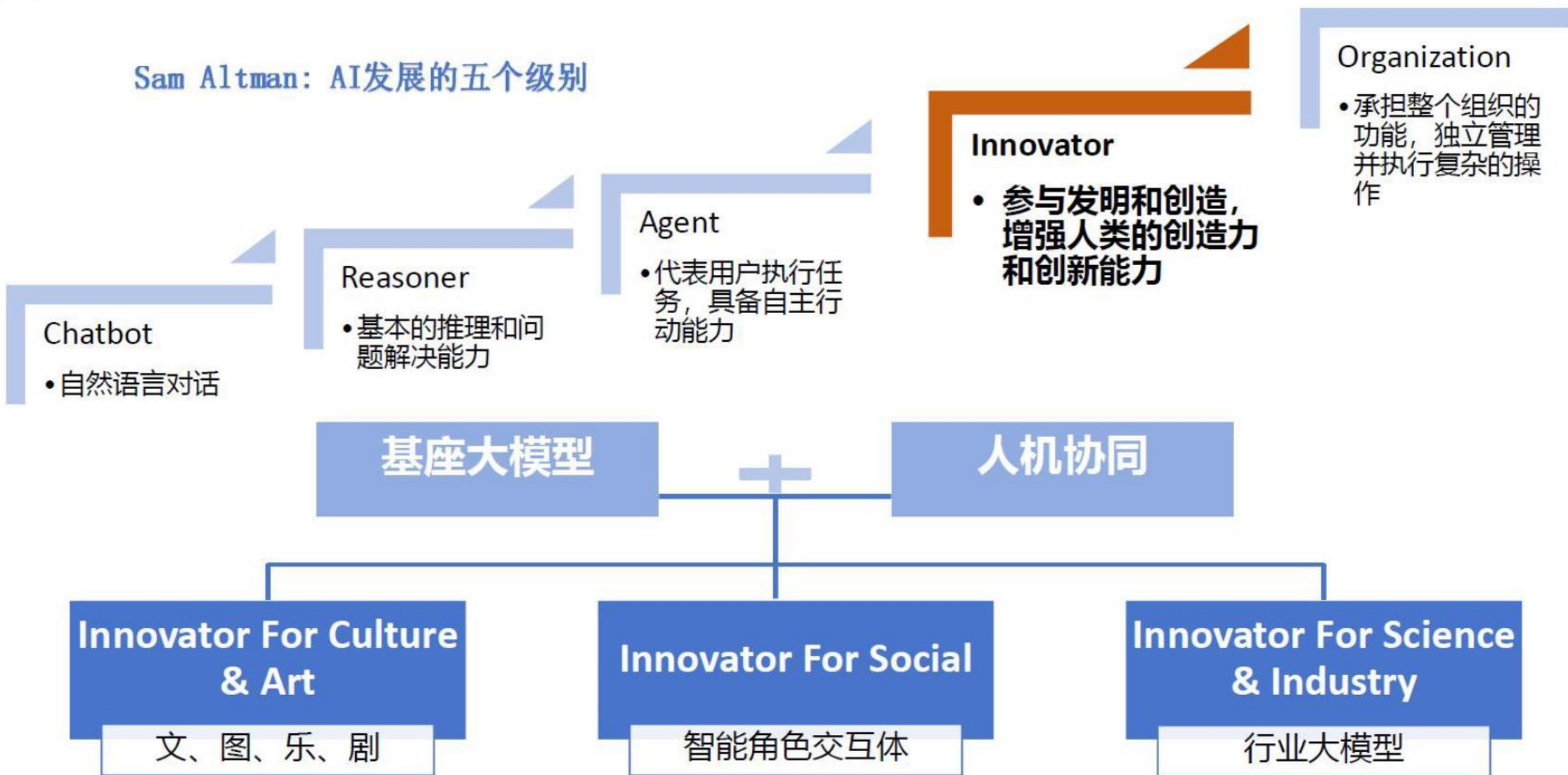
语言模型：自然语言统计建模，简单说，就是预测句子中的下一个单词是什么

大语言模型LLM: 2018 — 2024



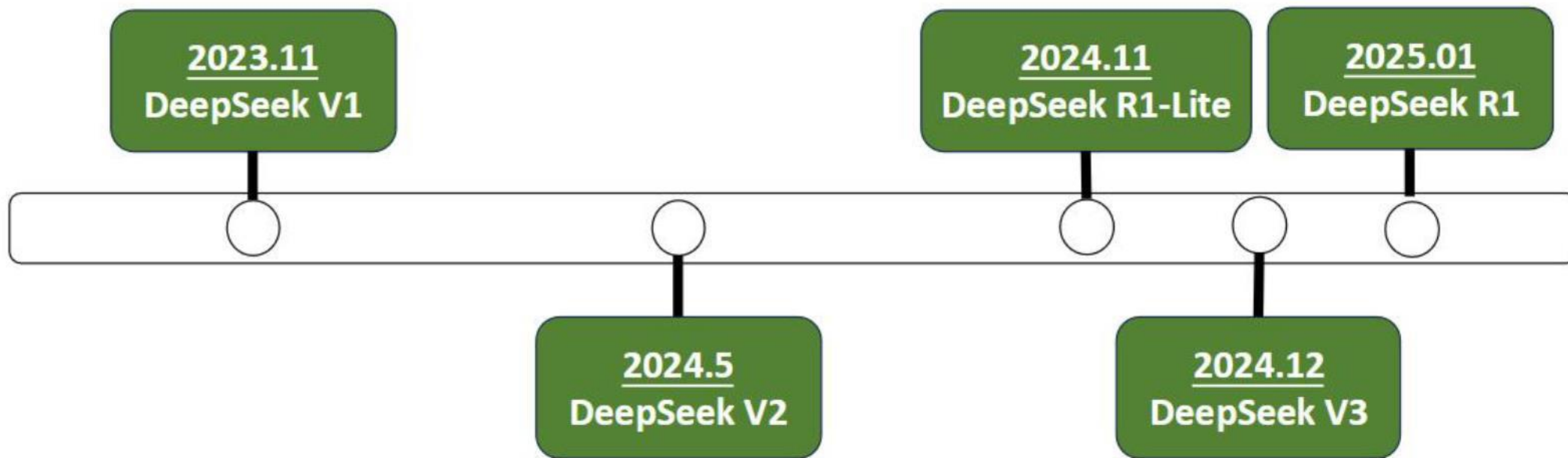
人机协同与共生驾驭AI实现新式智能

Sam Altman: AI发展的五个级别



DeepSeek介绍

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

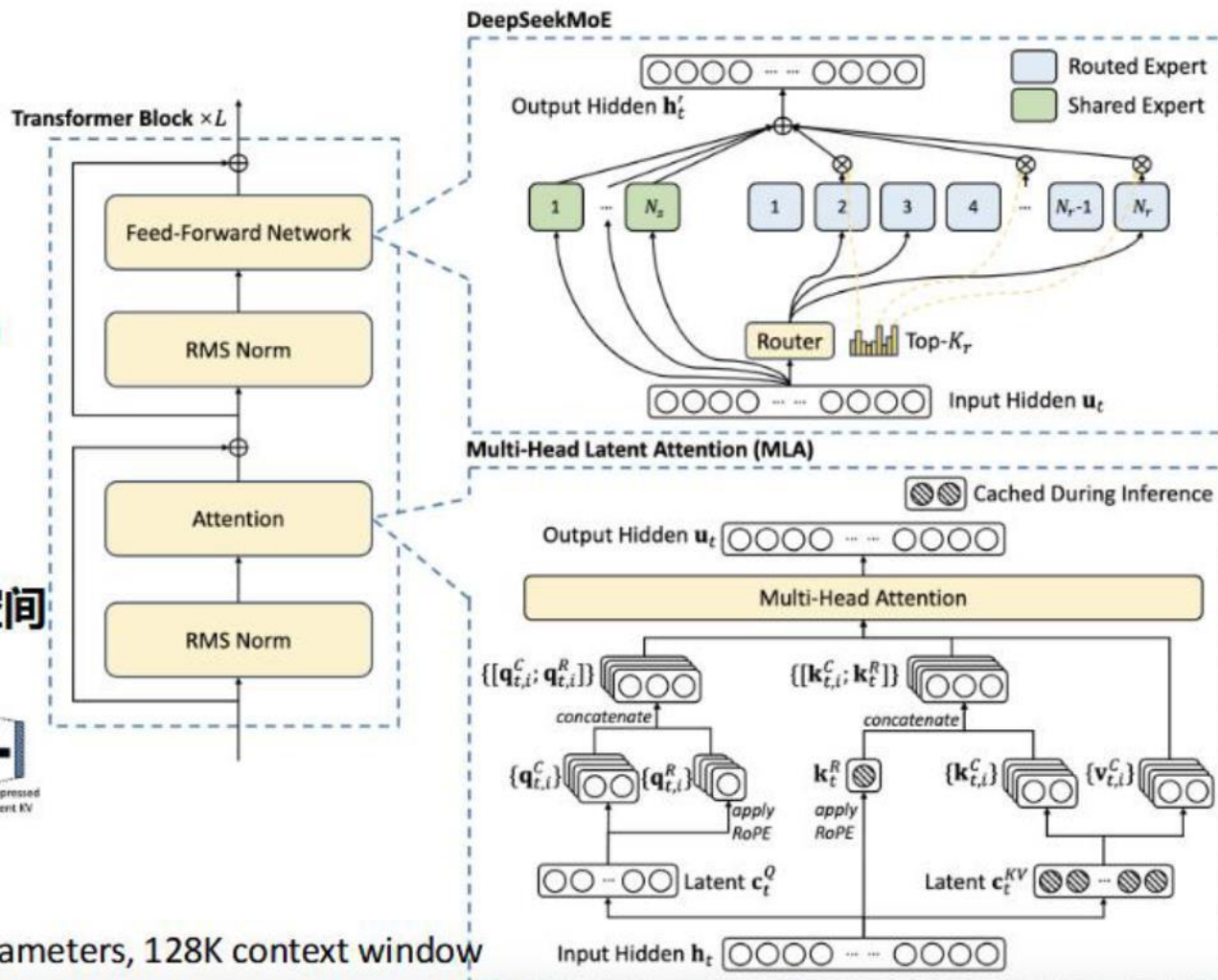
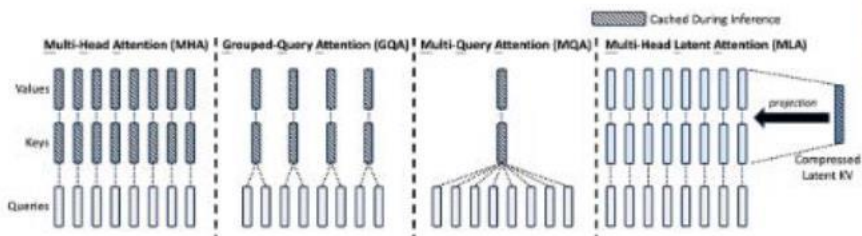
DeepSeek V2主要创新

- **DeepSeekMoE**
- **MLA (Multi-Head Latent Attention)**

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 | V3

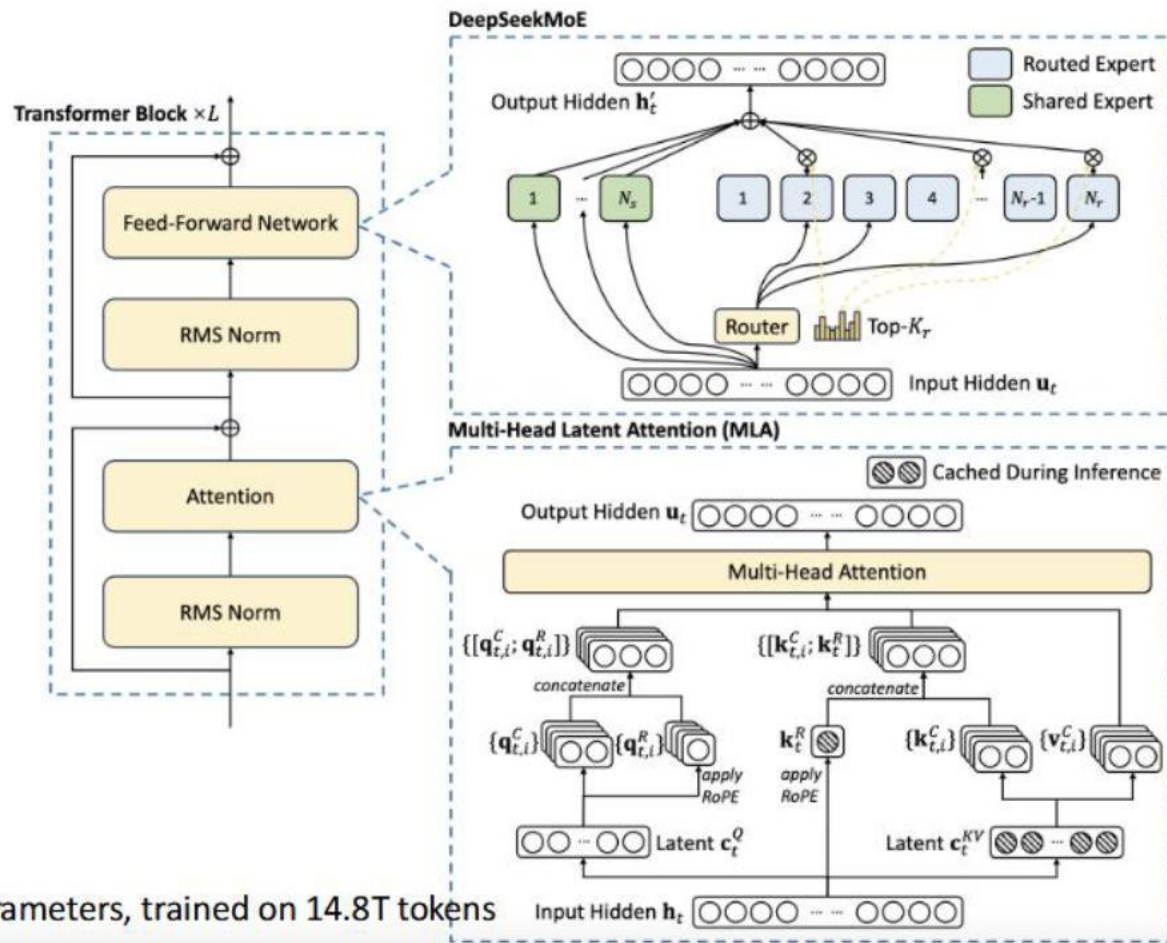
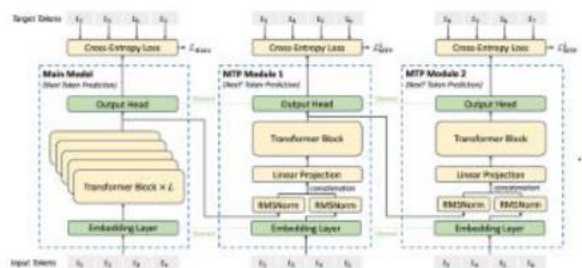
DeepSeek V3主要创新

- **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—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

PPO: Proximal Policy Optimization

GRPO: Group Relative Policy Optimization

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

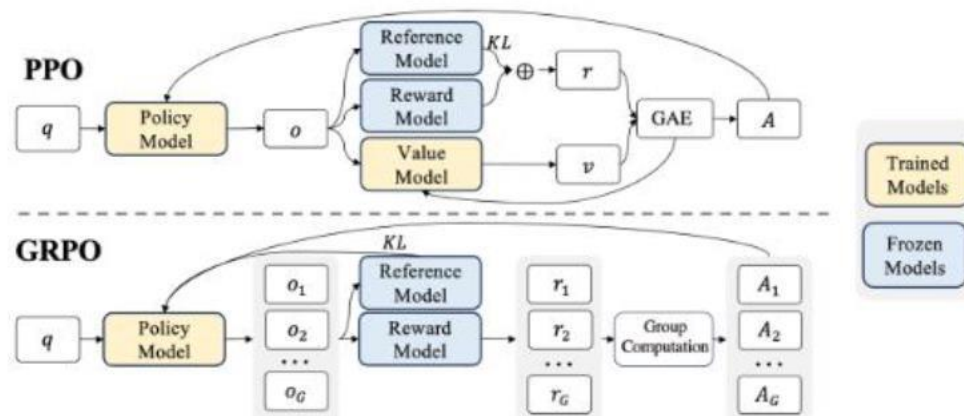


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.



强化学习让智能体 (Agent) 在环境 (Environment) 中不断尝试、学习, 并优化自己的策略 (Policy), 最终获得最大化的奖励 (Reward)。

DeepSeek应用场景

DeepSeek赋能职场的多维场景

A 文档与演示制作

- PPT设计与制作:快速生成演示框架、设计版式、优化配色方案
- 办公文档撰写:会议纪要、工作报告、项目方案、商业计划书
- 图表制作:数据可视化、流程图、组织架构图、思维导图
- 文档智能校对:语法检查、格式规范、内容优化建议

B 营销与内容创作

- 新媒体文案:公众号文章、社交媒体内容、产品文案
- 营销方案策划:活动策划、营销创意、传播策略
- 品牌设计:品牌标识、视觉元素、宣传物料
- 营销素材生成:广告文案、产品描述、EDM邮件

C 数据分析与决策支持

- 数据处理:数据清洗、转换、整理、标准化
- 数据分析:趋势分析、关联分析、预测分析
- 商业智能:销售预测、市场分析、客户洞察
- 报表生成:自动化报表、数据仪表盘、分析报告

D 研发与技术支持

- 代码开发:程序编写、代码优化、bug修复
- 应用开发:小程序开发、网站建设、系统开发
- 技术文档:API文档、使用手册、技术规范
- 测试与调试:自动化测试、性能优化、问题诊断

E 客户服务与沟通

- 智能客服:自动回复、问题解答、服务支持
- 会议支持:会议记录、语音转写、翻译服务
- 邮件处理:邮件分类、自动回复、重要信息提取
- 培训辅助:培训材料生成、知识库建设、在线答疑

F 项目与流程管理

- 项目规划:进度安排、资源分配、风险评估
- 任务管理:任务分解、进度跟踪、提醒督办
- 流程优化:流程分析、效率提升、自动化改造
- 绩效管理:数据收集、分析评估、报告生成

DeepSeek+：AI赋能职场应用范式拓展

工具组合	应用场景	使用步骤
DeepSeek + Kimi/讯飞智文/WPSAI等	制作PPT	在DeepSeek中生成PPT大纲与细节，复制到PPT助手中选择模板生成PPT
DeepSeek + 剪映	制作短视频	在DeepSeek中生成爆款短视频文案，使用剪映的图文成片功能生成视频
DeepSeek + Cursor	编写代码	使用Cursor中的chat和composer界面编写和生成工程代码
DeepSeek + 即梦/Midjourney/...	制作海报	在DeepSeek中生成海报描述，复制到即梦中生成海报图片
DeepSeek + XMind	创建思维导图	在DeepSeek中生成xmind格式的Markdown文本，导入XMind生成思维导图
DeepSeek + boardmix AI	创建图表和设计架构图	在DeepSeek中生成需求描述，复制到boardmix AI中生成交互式图表或架构图
.....		

DeepSeek的能力层级

- **1. 基础能力层** 多模态数据融合与结构化理解，包括**跨模态语义对齐**（文本、图像、音频、视频、代码、传感器数据统一语义）和**动态数据治理**（解决数据缺失、噪音干扰、概念飘逸等），支持200多种数据格式自动解析。
- **2. 中级能力层** 领域问题建模与复杂推理，包括**领域自适应学习**（建立医、教育、金融垂直应用于模型）、**因果推理引擎**（建立因果图模型）和**多目标优化决策**（求解帕累托最优解）。
- **3. 高级能力层** 复杂系统建模与自主决策，包括**数字孪生仿真系统**（构建物理与数字融合虚拟环境模拟天气等）、**多智能体协同优化**（将每个个体作为智能体通过联邦学习模拟群体行为）和**元认知调控机制**（实施监控自身决策、动态分配资源、自动触发行为）。
- **4. 终极能力层** 自主进化与创造性突破，包括**概念空间探索**（通过对抗网络探索新合金成分等）、**范式转移预警**（监控跨领域知识流、识别技术革命前兆）和**自编程能力**（自动模块设计、代码编写、测试用例）。

DeepSeek的三种模式



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

 深度思考 (R1)

 联网搜索



DeepSeek的三种模式对比

特性/模式	DeepSeek V3 (基础模型)	DeepSeek R1 (深度思考)	联网搜索
定位	通用自然语言处理模型，适合快速响应和高效处理多模态任务	复杂逻辑推理模型，专注于解决复杂问题和提供可解释性输出	基于RAG (检索增强生成)，能够结合网络信息回答问题 16
处理速度	快速响应，优化了高速性能	稍慢，但提供了更准确的结果，特别是在复杂任务中	取决于网络状况和数据源的加载时间
语言理解能力	强大，注重清晰、简洁的输出	更加深入地了解上下文和细微差别，支持动态调整音色和风格	结合实时网络信息进行理解和回答
架构特点	混合专家 (MoE) 架构，支持FP8和BF16推理模式	强化学习 (RL) 训练，采用思维链展示推理过程	需要外部API端点访问互联网资源
应用场景	适合快速问答、百科知识查询、内容创作等日常任务	适用于科研、算法交易、复杂决策支持等需要深度分析的任务	用于获取最新资讯、市场动态、学术研究成果等时效性强的信息
定制灵活性	有限的自定义选项	更灵活，允许对特定任务进行更深入的自定义	依赖于所使用的搜索引擎API的功能
延迟	低延迟、高速性能	因为需要更多的处理能力，延迟略高	可能因网络条件和API响应时间而有所不同
最佳用例	需要速度和效率的任务，如编码、数据分析	需要深入了解和推理的任务，如金融策略生成	当需要最新或未预见到的数据时，如新闻事件
计算资源需求	较低，适合小规模部署	较高，专为大规模企业工作负载设计	需要稳定的网络连接和可能较高的计算资源来处理大量数据
开源情况	是的，开放模型权重	是的，采用MIT开源协议	开源情况取决于所集成的搜索引擎API

DeepSeek 三种模式差异

- 基础模型 (V3) : 通用模型 (2024.12) , 高效便捷, 适用于绝大多数任务, “规范性”任务
- 深度思考 (R1) : 推理模型, 复杂推理和深度分析任务, 如数理逻辑推理和编程代码, “开放性”任务
- 联网搜索: RAG (检索增强生成) , 知识库更新至**2024年7月**

V3 还是 R1? 过程驱动 (规范约束) 还是 结果驱动 (模糊目标) ?

“聪明且听话”

Deepseek V3

“很聪明但没那么听话”

Deepseek R1

DeepSeek 两种模型对比 (5R)

维度	V3模型	R1模型
Regulation (规范性)	强规范约束 (操作路径明确)	弱规范约束 (操作路径开放)
Result (结果导向)	目标确定性高 (结果可预期)	目标开放性高 (结果多样性)
Route (路径灵活性)	线性路径 (流程标准化)	网状路径 (多路径探索)
Responsiveness (响应模式)	被动适配 (按规则执行)	主动创新 (自主决策)
Risk (风险特征)	低风险 (稳定可控)	高风险 (不确定性高)

(限于文本生成任务)

DeepSeek V3适配的任务场景

任务类型	具体场景	特点
快速响应	即时问答、百科知识查询、内容创作	快速反馈，优化了高速性能，适合即时性要求高的任务
文本生成	文章写作、故事创作、诗歌写作、营销文案	清晰、简洁的输出，适合规范性、有示例的任务
对话系统	客服对话、用户意向收集、智能外呼	灵活的自然语言理解能力，能够处理多样化的对话场景
多轮对话	开放性问答、角色扮演模拟	能够理解上下文，维持连续对话的一致性和流畅度
编程辅助	代码注释生成、API文档生成	提供技术文档处理功能，有助于开发者提高工作效率

DeepSeek R1适配的任务场景

任务类型	具体场景	特点
复杂推理	数学问题解答、逻辑分析、因果关系推断	强化学习驱动，支持长链推理（CoT），提供详细的推理步骤
代码生成与调试	根据需求生成代码片段、自动补全、错误分析与修复建议	高精度的专业领域处理能力，尤其适合需要严格逻辑链的任务
创意生成	创意类文案写作、脑洞型思考、开放性创作	创意能力强、开放性思考能力强
文本摘要与改写	长文本摘要、文本简化、多语言翻译与本地化	擅长处理高密度逻辑的任务，如复杂的文档理解和再创作
企业级应用	多模态任务预处理、专业领域的深度分析	支持大规模企业工作负载，适用于金融策略生成等高级应用场景
科研用途	超大规模AI研究或通用人工智能探索	适用于科研机构或大型企业的复杂任务

Deepseek r1 鞅的证明

14:50

设 $W(t)$ 是标准布朗运动, 证明...

为了证明标准布朗运动 $W(t)$ 的平方 $W(t)^2$ 是否构成一个鞅过程, 我们可以使用伊藤引理进行分析。

1. 伊藤引理的应用:
考虑函数 $f(W(t)) = W(t)^2$, 应用伊藤引理:

$$df = f(W(t))dW(t) + \frac{1}{2}f''(W(t))dt$$

其中, $f'(W(t)) = 2W(t)$ 和 $f''(W(t)) = 2$ 。代入后得到:

$$d(W(t)^2) = 2W(t)dW(t) + dt$$

积分得到:

$$W(t)^2 = \int_0^t 2W(s)dW(s) + t$$

2. 鞅的分解:
积分项 $\int_0^t 2W(s)dW(s)$ 是一个伊藤积分, 因此是一个鞅。而 t 是一个递增的有限变差过程。因此, $W(t)^2$ 可以分解为一个鞅和一个有限变差过程。

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

14:50

设 $W(t)$ 是标准布朗运动, 证明...

$$W(t)^2 = \text{鞅} + t$$

3. 条件期望的计算:
计算条件期望 $E[W(t)^2 | \mathcal{F}_s]$ (其中 $s \leq t$):

$$E[W(t)^2 | \mathcal{F}_s] = E\left[\int_0^t 2W(u)dW(u) + t \mid \mathcal{F}_s\right]$$

由于伊藤积分是一个鞅, 其条件期望为:

$$E\left[\int_0^t 2W(u)dW(u) \mid \mathcal{F}_s\right] = \int_0^s 2W(u)dW(u)$$

代入后得到:

$$E[W(t)^2 | \mathcal{F}_s] = \int_0^s 2W(u)dW(u) + t$$

根据之前的分解, $W(s)^2 = \int_0^s 2W(u)dW(u) + s$, 因此:

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

14:50

设 $W(t)$ 是标准布朗运动, 证明...

$$E[W(t)^2 | \mathcal{F}_s] = \int_0^s 2W(u)dW(u) + t$$

根据之前的分解, $W(s)^2 = \int_0^s 2W(u)dW(u) + s$, 因此:

$$E[W(t)^2 | \mathcal{F}_s] = W(s)^2 - s + t = W(s)^2 + (t - s)$$

4. 结论:
由于 $E[W(t)^2 | \mathcal{F}_s] = W(s)^2 + (t - s)$, 这表明 $W(t)^2$ 的条件期望比 $W(s)^2$ 多了一个时间差项 $t - s$, 因此 $W(t)^2$ 本身并不是鞅, 而是一个下鞅。正确的鞅过程应为 $W(t)^2 - t$, 因为:

$$E[W(t)^2 - t | \mathcal{F}_s] = W(s)^2 - s$$

最终答案:
 $W(t)^2 - t$ 是一个鞅过程。题目中的 $W(t)^2$ 本身并不是鞅, 正确的鞅过程应为 $W(t)^2$ 减去 t , 即 $W(t)^2 - t$ 。

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

Deepseek r1 Roy安全准则组合模型推理

Roy安全第一组合理论的证明

1. 问题设定

投资者目标是选择资产权重 \mathbf{w} , 使得组合收益率 $R_p = \mathbf{w}^T \mathbf{R}$ 低于阈值 s 的概率最小化:

$$\min_{\mathbf{w}} P(R_p \leq s) \quad \text{满足} \quad \mathbf{w}^T \mathbf{1} = 1.$$

2. 分布假设与概率转化

假设资产收益率服从多元正态分布, 则组合收益 $R_p \sim \mathcal{N}(\mu_p, \sigma_p^2)$, 其中:

$$\mu_p = \mathbf{w}^T \boldsymbol{\mu}, \quad \sigma_p^2 = \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}.$$

低于 s 的概率为:

$$P(R_p \leq s) = \Phi\left(\frac{s - \mu_p}{\sigma_p}\right),$$

其中 Φ 为标准正态分布函数。为最小化该概率, 需最大化 $\frac{\mu_p - s}{\sigma_p}$ (因 Φ 单调递增)。

3. 优化目标

转化为优化问题:

$$\max_{\mathbf{w}} \frac{\mu_p - s}{\sigma_p} = \frac{\mathbf{w}^T \boldsymbol{\mu} - s}{\sqrt{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}}} \quad \text{s.t.} \quad \mathbf{w}^T \mathbf{1} = 1.$$

4. 拉格朗日求解

构建拉格朗日函数:

$$\mathcal{L} = \frac{\mathbf{w}^T \boldsymbol{\mu} - s}{\sqrt{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}}} + \lambda(\mathbf{w}^T \mathbf{1} - 1).$$

对 \mathbf{w} 求导并令导数为零, 得到最优条件:

$$\boldsymbol{\Sigma} \mathbf{w} \propto \boldsymbol{\mu} - s \mathbf{1}.$$

解得最优权重:

$$\mathbf{w}^* = \frac{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - s \mathbf{1})}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - s \mathbf{1})}.$$

5. 结论

最优组合权重 \mathbf{w}^* 与 $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - s \mathbf{1})$ 成比例, 归一化后满足预算约束。该权重最大化调整后的夏普比率 $\frac{\mu_p - s}{\sigma_p}$, 从而最小化下行风险。

关键公式

• 最优权重:

$$\mathbf{w}^* = \frac{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - s \mathbf{1})}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - s \mathbf{1})}.$$

• 安全第一比率:

$$\frac{\mu_p - s}{\sigma_p} \quad (\text{最大化此比率}).$$

经济意义

Roy的理论强调控制极端损失, 通过调整组合权重使得收益分布向高于安全水平 s 的方向偏移。与均值-方差模型不同, 安全第一准则更关注下行风险, 适用于风险厌恶程度极高的投资者。

最终结论: Roy的安全第一组合理论通过最大化调整后的夏普比率, 确定最优权重, 确保组合收益低于灾难性水平的概率最小化。

代码说明

1. 核心函数 `roy_safety_first()`:

• 直接实现Roy模型的闭式解公式:

$$\mathbf{w}^* = \frac{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - s \mathbf{1})}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - s \mathbf{1})}$$

• 输入为预期收益向量、协方差矩阵和安全阈值 s

• 输出最优权重、组合预期收益和标准差

2. 示例数据生成:

• 使用随机数生成3个资产的收益均值和协方差矩阵

• 通过矩阵运算确保协方差矩阵正定

3. 结果展示:

• 使用Pandas格式化输出结果

• 显示安全比率 $(\mu_p - s)/\sigma_p$

4. 可视化分析:

• 展示安全阈值 s 变化对组合收益和风险的影响

• 灰色区域表示收益的波动范围

Deepseek r1推理玻尿酸配方

16:42

设 $W(t)$ 是标准布朗运动，证明...

设计玻尿酸（透明质酸）美容配方时，需兼顾保湿性、稳定性、安全性和肤感。以下是一个科学合理的最佳配方框架及详细说明：

配方目标

- 深层保湿：**通过不同分子量透明质酸（HA）实现多层次补水。
- 舒缓修护：**添加抗炎和屏障修复成分，减少敏感风险。
- 稳定安全：**避免刺激性成分，确保防腐体系温和有效。
- 清爽肤感：**适合多种肤质（包括油性/混合性皮肤）。

核心配方成分及比例（以精华液为例）

成分	浓度范围	功能说明
透明质酸钠 (HA)	0.5%-2%	分梯度添加不同分子量（高分子量成膜锁水

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

16:42

设 $W(t)$ 是标准布朗运动，证明...

泛醇 (维生素 B5)	1%-2%	修复屏障，缓解干燥刺激，提升皮肤柔软度。
β -葡聚糖	0.1%-0.5%	抗炎舒缓，增强皮肤免疫力，减少 HA 可能引发的刺痛感。
烟酰胺 (可选)	2%-3%	控油、提亮肤色（敏感肌需谨慎测试）。
戊二醇 + 乙基己基甘油	0.5%-1%	温和防腐体系，替代传统防腐剂（如苯氧乙醇），降低刺激风险。
柠檬酸/乳酸	适量	调节 pH 至 5.5-6.0（接近皮肤天然 pH 值）。

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

16:42

设 $W(t)$ 是标准布朗运动，证明...

3. 肤感调节：

- 若需清爽质地，可加入轻质硅油（如环五聚二甲基硅氧烷）或丙烯酸（酯）类共聚物（0.1%-0.3%）改善延展性。
- 避免过多增稠剂（如卡波姆），以免粘腻。

4. 敏感肌适配：

- 添加红没药醇（0.05%-0.1%）或积雪草提取物，进一步舒缓敏感。

制备工艺流程

- 水相制备：**
 - 将去离子水加热至 70-75°C，依次溶解甘油、泛醇、 β -葡聚糖。
 - 降温至 40°C 后，加入透明质酸钠（预先用少量水分散），搅拌至完全溶解。
- pH 调节：**
 - 用柠檬酸或乳酸调整 pH 至 5.5-6.0。
- 防腐与稳定：**
 - 加入戊二醇和 EDTA

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

16:42

设 $W(t)$ 是标准布朗运动，证明...

最终配方示例（简化版）

```
markdown
- 去离子水 至 100%
- 甘油 4%
- 泛醇 2%
- 高分子透明质酸钠 0.3%
- 低分子透明质酸钠 0.7%
-  $\beta$ -葡聚糖 0.2%
- 戊二醇 0.6%
- 乙基己基甘油 0.4%
- 柠檬酸 适量 (调 pH 至 5.5)
```

结论

此配方通过多分子量透明质酸复配实现“外锁内补”，结合泛醇和 β -葡聚糖的修护作用，适合作为日常保湿精华。若需拓展功效（如抗衰、美白），可叠加其他活性成分（如依克多因、VC 衍生物），但需注意配伍稳定性。

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

DeepSeek提示词工程

如何提问？两种模型的提示语差异

- 基础模型 (V3) : “过程-结果”清晰 (指令)
- 深度思考 (R1) : 目标清晰, 结果可以模糊 (推理)

提示词工程

DeepSeek两种模型的提示语差异

特性/方面	DeepSeek R1	DeepSeek V3
推理能力	<u>强调深度推理，模型会自动执行链式思考（Chain of Thought, CoT）提供详细的推理过程</u>	<u>更适合通用任务，依赖于下一个单词预测机制，通常不展示内部推理步骤</u>
提示结构	<u>不需要复杂的角色扮演或分步指令，简单直接的提示即可获得良好效果，但上下文信息依旧重要</u>	<u>可能需要明确的角色设定、任务描述和步骤指导来引导模型生成更符合预期的回答</u>
背景信息需求	对于超出常识范围的任务，必须提供足够的背景信息或上下文，以确保模型能够理解用户的意图	同样需要提供背景信息，但在某些情况下，由于其广泛的训练数据，可能不需要过多的背景解释
输出长度与形式	倾向于生成细致且结构化的答案，尤其是对于复杂问题，可能会产生较长的响应	通常输出更为简练，除非特别要求，否则不会主动提供详细的中间步骤
自我检查机制	具备一定的自我事实核查机制，在生成响应的过程中更善于发现并纠正错误	自我检查机制相对较少，更多依赖于外部反馈进行修正
交互速度	因为需要进行深入的思考过程，响应时间相对较长	响应速度快，适用于实时交互场景
提示词技巧	<u>提供简单的提示即可，强调准确表达目标而非具体步骤；避免过度细化的指令，让AI自行决定最佳行动路径 18</u>	<u>需要清晰定义你是谁、你的目标是什么，并可能需要给出具体的执行步骤来引导模型</u>

DeepSeek官方提示语示例

提示库

探索 DeepSeek 提示词样例，挖掘更多可能



代码改写

对代码进行修改，来实现纠错、注释、调优等。



代码生成

让模型生成一段完成特定功能的代码。



结构化输出

将内容转化为 Json，来方便后续程序处理



角色扮演（情景续写）

提供一个场景，让模型模拟该场景下的任务对话



诗歌创作

让模型根据提示词，创作诗歌



宣传标语生成

让模型生成贴合商品信息的宣传标语。



代码解释

对代码进行解释，来帮助理解代码内容。



内容分类

对文本内容进行分析，并对齐进行自动归类



角色扮演（自定义人设）

自定义人设，来与用户进行角色扮演。



散文写作

让模型根据提示词创作散文



文案大纲生成

根据用户提供的主题，来生成文案大纲



模型提示词生成

根据用户需求，帮助生成高质量提示词

<https://api-docs.deepseek.com/zh-cn/prompt-library/>



RTGO提示语结构

Role (角色)

定义AI的角色：
经验丰富的数据分析师
具备十年销售经验的SaaS系统商务
.....

Task (任务)

具体任务描述：
写一份关于XXX活动的小红书宣推文案
写一份关于XX事件的舆论分析报告
(XX活动/事件相关背景信息如下.....)

Goal (目标)

期望达成什么目标效果：
通过该文案吸引潜在客户，促成消费.....通过该报告为相关企业管理者提供.....策略支撑

Objective (操作要求)

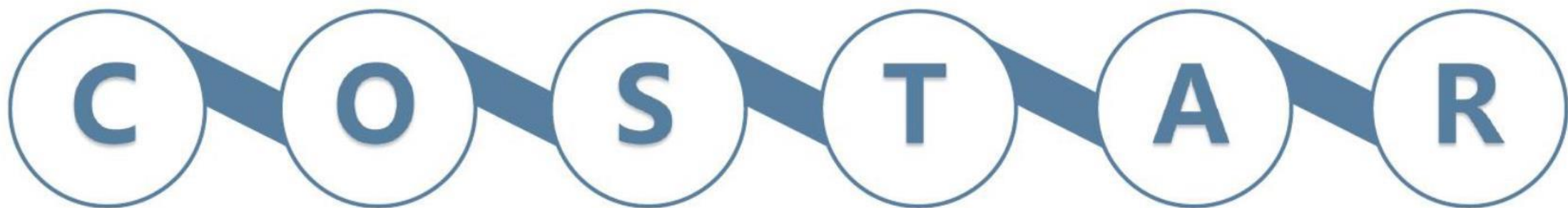
字数要求、段落结构、用词风格、
内容要点、输出格式...



RTGO

CO-STAR提示语框架

新加坡 GPT-4 提示工程竞赛冠军提示词框架



"C"代表
"Context (上下文)" 相关的背景信息, 比如你自己或是你希望它完成的任务的信息。

"O"代表
"Objective (目标)" 明确的指示告诉AI你希望它做什么。

"S"代表 "Style (风格)" 想要的写作风格, 如严肃的、有趣的、创新性表达、学术性.....

"T"代表 "Tone (语调)" 幽默的? 情绪化? 有威胁性?

"A"代表
"Audience", 受众是谁。小白用户? 专业人群? 未成年群体? 女性群体?

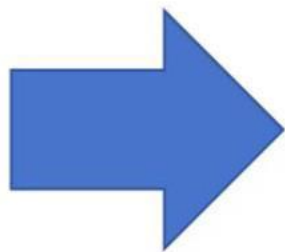
"R", 代表
"Response", 想要的回应类型。一份详细的研究报告? 一个表格? Markdown格式?

提示语使用十级技巧

级别	技巧名称	描述	示例
1	基础请求	直接表达需求。	“请总结一篇维基百科的文章。”
2	应用格式	使用基本格式调整来提高理解。	在提示词中添加破折号以区分含义。
3	精准请求	明确且专注地表达需求，避免模糊。	“请用标题、小标题和表格格式回答以下问题。”
4	示例说明	提供示例输入和期望的输出。	描述问题并直接给出希望的答案格式。
5	自我反省	利用 AI 的自我评估能力优化输出。	“嘿，你漏掉了什么信息吗？”
6	系统提示与定制指令	精确控制系统提示，提供详细背景信息。	介绍自己是软件工程师，偏好简短回答等。
7	人格化应用	通过角色设定提高答案准确性。	“扮演一个历史学家，回答这个问题。”
8	思维链	让 AI 解释其思考过程，逐步处理复杂问题。	“让我们一步步来思考这个问题。”
9	让大语言模型自己写提示词	允许 AI 自己生成详细的提示词来解答问题。	用 AI 生成的提示词在新聊天窗口中解答问题。
10	CO-STAR 框架	使用 CO-STAR 框架精确地构建提示。	提供背景、目标、风格、语调、听众和回应的详细说明。

DeepSeek R1提示语技巧 (开放性)

- 不需要角色设定
- 不需要思维链提示
- 不需要结构化提示词
- 不需要给示例
- 不需要做太多解释
-



- 干什么?
- 给谁干?
- 目的是? (要什么)
- 约束是? (不要什么)

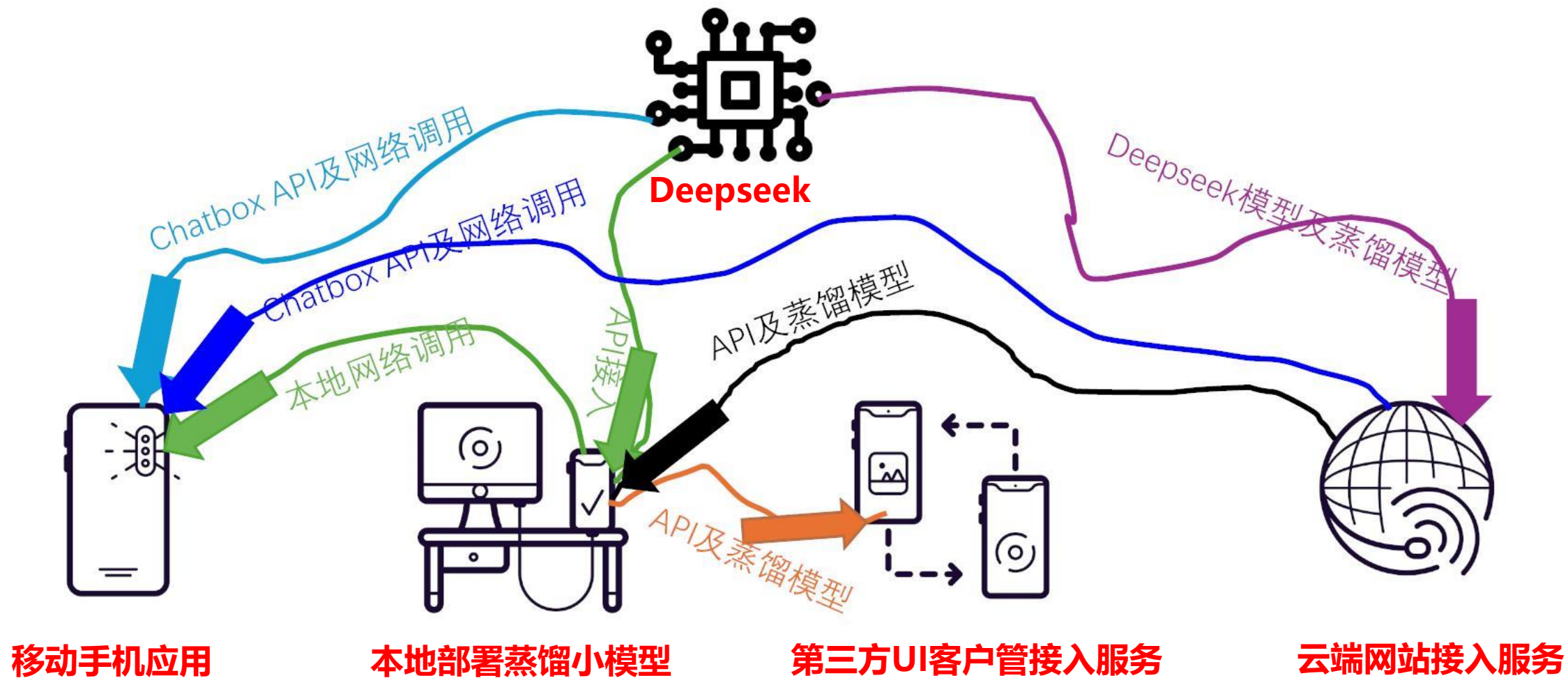
不要过多干预, 不要手把手教学

Deepseek十类提示词

- **1. 内容生成类**：文本生成、代码生成、创意生成和数据模拟。
- **2. 信息处理类**：文本摘要、信息抽取、情感分析和多语言翻译。
- **3. 对话交互类**：角色扮演、多轮对话、反问引导。
- **4. 技能应用类**：数学计算、代码解释、逻辑推理。
- **5. 个性化定制类**：风格迁移、知识库绑定、偏好记忆。
- **6. 系统操作类**：模式切换、资源优化、记忆管理。
- **7. 知识查询类**：事实核查、概念解释、溯源检索。
- **8. 教育与研究类**：题目生成、论文润色、实验设计。
- **9. 多模态处理类**：图文互译、表格解析、视频脚本。
- **10. 伦理与安全类**：内容审核、价值观对齐、风险预警。

DeepSeek部署方案

Deepseek部署全线图



Deepseek替代部署方案

- **本地部署蒸馏小模型：**可通过ollama平台、LM Studio平台快速本地部署基于Llama和Qwen蒸馏的DeepSeek - R1推理模型，涵盖多种不同参数规模，适合想在本地运行模型的用户。
- **第三方UI客户管接入服务：**可通过Anything LLM、Cherry Studio、Chatbox等选择API接入。
- **秘塔AI搜索：**接入满血版DeepSeek - R1推理模型，无广告且搜索结果直达。以学习JDK21新特性为例，能详细给出学习计划，包括快速预览、深入学习核心特性、实战与总结等阶段。
- **硅基流动：**注册即送2000万Tokens，提供多个基于DeepSeek - R1蒸馏训练的模型，如DeepSeek - R1 - Distill - Llama - 70B等，在推理能力上表现出色。
- **阿里云百炼：**提供多个DeepSeek系列模型，如DeepSeek - V3、DeepSeek - R1 - Distill - Qwen - 1.5B等，部分限时免费，涵盖文本生成等功能。
- **百度智能云：**千帆ModelBuilder全面支持DeepSeek - R1/V3调用，且限时免费，其模型在百科知识、数学推理等任务表现突出。
- **英伟达：**可通过特定接口调用DeepSeek - R1 671B全量模型，获取API Key后能进行相关操作。

DeepSeek API调用

DeepSeek的使用入口

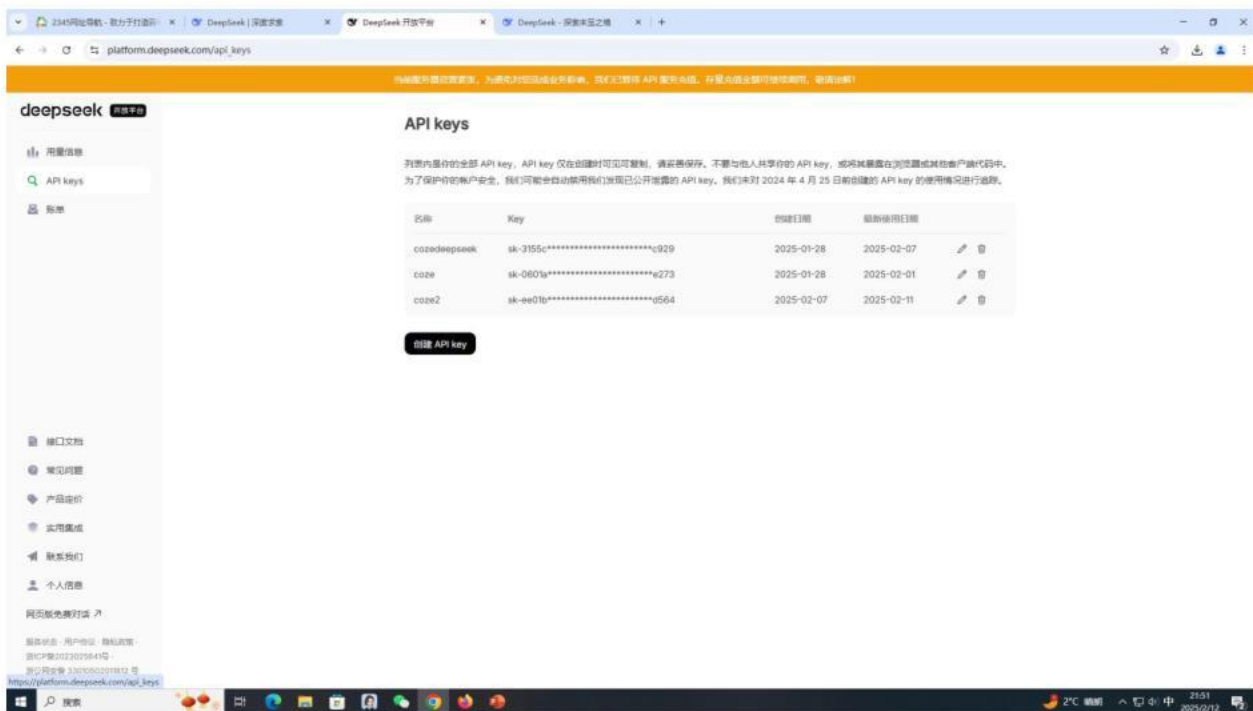
官方入口: <https://chat.deepseek.com/>

平台	地址	版本	备注
硅基流动 (SiliconFlow)	https://siliconflow.cn/zh-cn/	DeepSeek-R1 和 DeepSeek-V3, 均为671B满血版, 普通版和 Pro 版 (均为671B, Pro版响应更稳定但需付费)	新用户注册赠送 2000万Tokens (约14元额度), 普通版可用额度抵扣, Pro版需充值。
纳米AI	APP	671B满血版 (官方同款)	目前提供免费使用 (满血版暂时开放)。 推荐用于编程等场景, 性能与官方一致。
秘塔搜索	https://metaso.cn/	R1增强版 (基于671B参数优化)	免费使用, 侧重搜索功能, 编程性能略有牺牲。 支持“长思考”模式, 适合复杂问题处理。
英伟达NIM微服务	https://build.nvidia.com/deepseek-ai/deepseek-r1	671B (全量模型)	网页版直接使用, 支持API调用, 注册送1000点数, 免费体验。
微软Azure	https://ai.azure.com	671B (全量模型)	需注册微软账户并创建订阅, 免费部署, 支持参数调节。
亚马逊AWS	https://aws.amazon.com/cn/blogs/aws/deepseek-r1-models-now-available-on-aws	671B (全量模型)	需注册AWS账户, 填写付款方式, 免费部署。
Cerebras	https://cerebras.ai	70B	邮箱注册, 速度快, 宣称比GPU方案快57倍。
Groq	https://groq.com/groqcloud-makes-deepseek-r1-distill-llama-70b-available	70B	邮箱注册, 速度快, 但感觉比Cerebras弱一些。

.....

API 调用方法

- DeepSeek API



PARAM	VALUE
base_url *	https://api.deepseek.com
api_key	apply for an API key

curl python nodejs

```
# Please install OpenAI SDK first: `pip3 install openai`

from openai import OpenAI

client = OpenAI(api_key="<DeepSeek API Key>", base_url="https://api.deepseek.com")

response = client.chat.completions.create(
    model="deepseek-chat",
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": "Hello"},
    ],
    stream=False
)

print(response.choices[0].message.content)
```

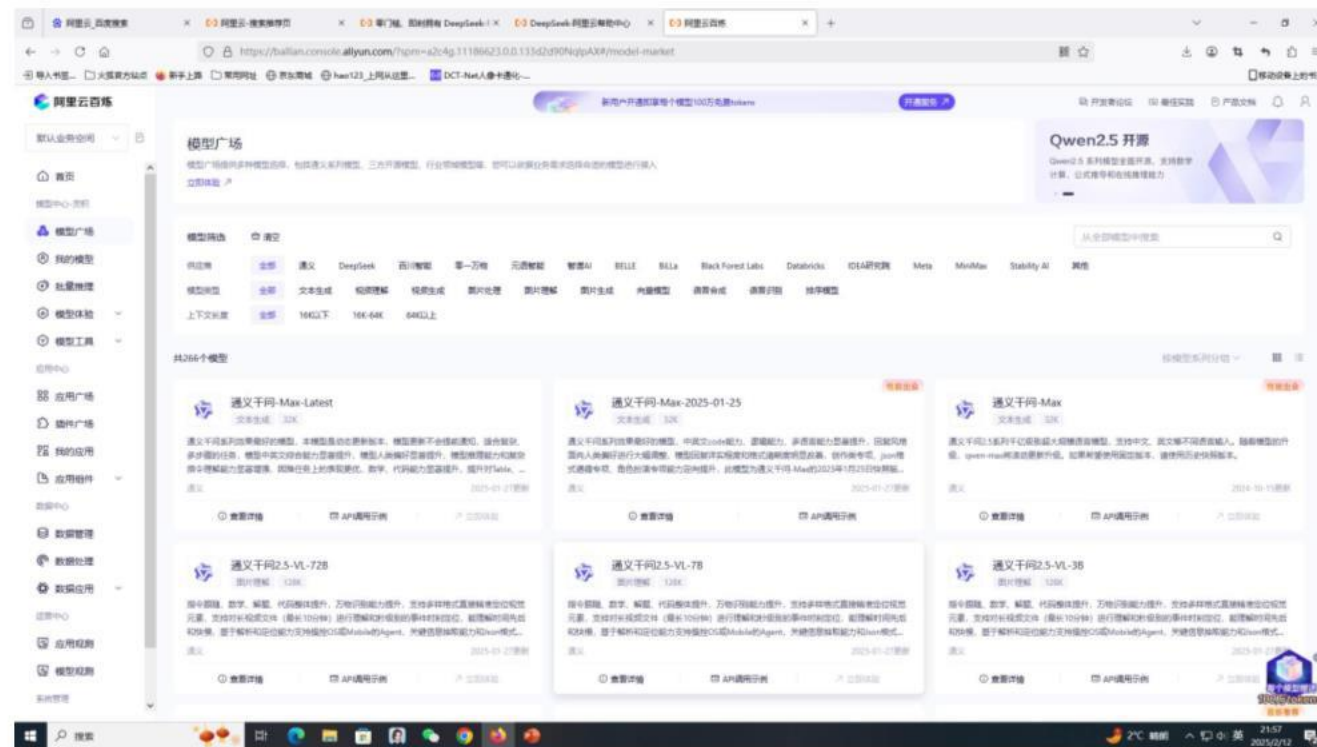
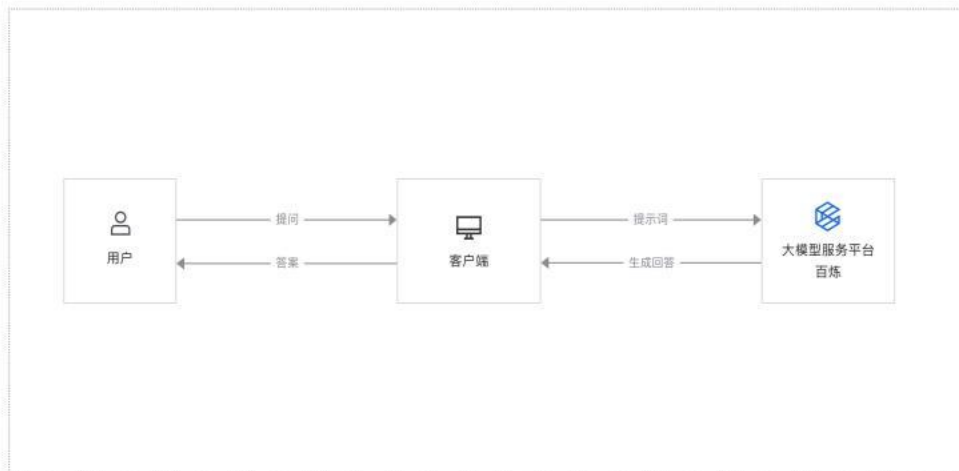
model='deepseek-chat' 即可调用 DeepSeek-V3
model='deepseek-reasoner', 即可调用 DeepSeek-R1。

Token 用量计算

- token 是模型用来表示自然语言文本的基本单位，也是我们的计费单元，可以直观的理解为“字”或“词”；通常 1 个中文词语、1 个英文单词、1 个数字或 1 个符号计为 1 个 token。
- 一般情况下模型中 token 和字数的换算比例大致如下：
 - 1 个英文字符 \approx 0.3 个 token。
 - 1 个中文字符 \approx 0.6 个 token。
- 但因为不同模型的分词不同，所以换算比例也存在差异，每一次实际处理 token 数量以模型返回为准，您可以从返回结果的 usage 中查看。

阿里云部署Deepseek

以 DeepSeek-R1 满血版为例进行演示，通过**百炼模型服务**进行 DeepSeek 开源模型调用，可以根据实际需求选择其他参数规模的 DeepSeek 模型。百炼平台的 **API** 提供标准化接口，无需自行搭建模型服务基础设施，且具备负载均衡和自动扩缩容机制，保障 API 调用稳定性。搭配 **Chatbox** 可视化界面客户端，进一步简化了调用流程，无需在命令行中操作，通过图形化界面即可轻松配置和使用 DeepSeek 模型。



<https://chatboxai.app/zh#download>

模型广场

模型广场提供多种模型选择,包括通义系列模型、三方开源模型、行业领域模型等,您可以依据业务需求选择合适的模型进行接入

[立即体验](#)

DeepSeek

模型筛选 [清空](#)

从全部模型中搜索

- 供应商: [全部](#) [通义](#) [DeepSeek](#) [百川智能](#) [零一万物](#) [元语智能](#) [智谱AI](#) [BELLE](#) [Billa](#) [Black Forest Labs](#) [Databricks](#) [IDEA研究院](#) [Meta](#) [MiniMax](#) [Stability AI](#) [其他](#)
- 模型类型: [全部](#) [文本生成](#) [视频理解](#) [视频生成](#) [图片处理](#) [图片理解](#) [图片生成](#) [向量模型](#) [语音合成](#) [语音识别](#) [排序模型](#)
- 上下文长度: [全部](#) [16K以下](#) [16K-64K](#) [64K以上](#)

共266个模型

按模型系列分组



通义千问-Max-Latest

文本生成 32K

通义千问系列效果最好的模型。本模型是动态更新版本,模型更新不会提前通知,适合复杂、多步骤的任务,模型中英文综合能力显著提升,模型人类偏好显著提升,模型推理能力和复杂指令理解能力显著增强,困难任务上的表现更优,数学、代码能力显著提升,提升对Table、...

通义

2025-01-27更新

[查看详情](#)

[API调用示例](#)

[立即体验](#)



通义千问-Max-2025-01-25

文本生成 32K

通义千问系列效果最好的模型,中英文code能力、逻辑能力、多语言能力显著提升,回复风格偏向人类偏好进行大幅调整,模型回复详实程度和格式清晰度明显改善,创作类专项、json格式遵循专项、角色扮演专项能力定向提升,此模型为通义千问-Max的2025年1月25日快照版...

通义

2025-01-27更新

[查看详情](#)

[API调用示例](#)

性能出众



通义千问-Max

文本生成 32K

通义千问2.5系列千亿级超大规模语言模型,支持中文、英文等不同语言输入。随着模型的升级,qwen-max将滚动更新升级,如果希望使用固定版本,请使用历史快照版本。

通义

2024-10-15更新

[查看详情](#)

[API调用示例](#)

[立即体验](#)

性能出众



通义千问2.5-VL-72B

图片理解 128K

指令跟随、数学、解题、代码整体提升,万物识别能力提升,支持多样格式直接精准定位视觉元素,支持对长视频文件(最长10分钟)进行理解和秒级别的事件时刻定位,能理解时间先后和快慢,基于解析和定位能力支持操控OS或Mobile的Agent,关键信息抽取能力和json格式...

通义

2025-01-27更新

[查看详情](#)

[API调用示例](#)

[立即体验](#)



通义千问2.5-VL-7B

图片理解 128K

指令跟随、数学、解题、代码整体提升,万物识别能力提升,支持多样格式直接精准定位视觉元素,支持对长视频文件(最长10分钟)进行理解和秒级别的事件时刻定位,能理解时间先后和快慢,基于解析和定位能力支持操控OS或Mobile的Agent,关键信息抽取能力和json格式...

通义

2025-01-27更新

[查看详情](#)

[API调用示例](#)

[立即体验](#)



通义千问2.5-VL-3B

图片理解 128K

指令跟随、数学、解题、代码整体提升,万物识别能力提升,支持多样格式直接精准定位视觉元素,支持对长视频文件(最长10分钟)进行理解和秒级别的事件时刻定位,能理解时间先后和快慢,基于解析和定位能力支持操控OS或Mobile的Agent,关键信息抽取能力和json格式...

通义

2025-01-27更新

[查看详情](#)

[API调用示例](#)

[立即体验](#)



默认业务空间

首页

模型中心-灵积

模型广场

我的模型

批量推理

模型体验

模型工具

应用中心

应用广场

插件广场

我的应用

应用组件

数据中心

数据管理

数据处理

数据应用

运营中心

应用观测

模型观测

系统管理

模型筛选 清空

供应商 全部 通义 DeepSeek 百川智能 零一万物 元语智能 智谱AI BELLE BiLa Black Forest Labs Databricks IDEA研究院 Meta MiniMax Stability AI 其他

模型类型 全部 文本生成 视频理解 视频生成 图片处理 图片理解 图片生成 向量模型 语音合成 语音识别 排序模型

上下文长度 全部 16K以下 16K-64K 64K以上

从全部模型中搜索

共8个模型

按模型系列分组



DeepSeek-R1

文本生成 64K

DeepSeek-R1 在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，性能较高，能力较强。

DeepSeek

查看详情

DeepSeek-R1-Ditill-Qwen-7B



DeepSeek-R1-Distill-Qwen-7B

文本生成 32K

DeepSeek-R1-Distill-Qwen-7B是一个基于Qwen2.5-Math-7B的蒸馏大型语言模型，使用了DeepSeek R1的输出。

DeepSeek

查看详情

DeepSeek-R1-Ditill-Ollama-7B



DeepSeek-R1-Distill-Llama-8B

文本生成 32K

DeepSeek-R1-Distill-Llama-8B是一个基于Llama-3.1-8B的蒸馏大型语言模型，使用了DeepSeek R1的输出。



DeepSeek-V3

文本生成 64K

DeepSeek-V3 为自研 MoE 模型，671B 参数，激活 37B，在 14.8T token 上进行了预训练，在长文本、代码、数学、百科、中文能力上表现优秀。

DeepSeek

查看详情

DeepSeek-V3



DeepSeek-R1-Distill-Qwen-14B

文本生成 32K

DeepSeek-R1-Distill-Qwen-14B是一个基于Qwen2.5-14B的蒸馏大型语言模型，使用了DeepSeek R1的输出。

DeepSeek

查看详情

DeepSeek-R1-Ditill-Qwen-14B



DeepSeek-R1-Distill-Llama-70B

文本生成 32K

DeepSeek-R1-Distill-Llama-70B是一个基于Llama-3.1-70B的蒸馏大型语言模型，使用了DeepSeek R1的输出。

DeepSeek-R1-Distill-Ollama-7B



DeepSeek-R1-Ditill-Qwen-1.5B

文本生成 32K

DeepSeek-R1-Distill-Qwen-1.5B是一个基于Qwen2.5-Math-1.5B的蒸馏大型语言模型，使用了DeepSeek R1的输出。

DeepSeek

查看详情

DeepSeek-R1-Ditill-Qwen-32B



DeepSeek-R1-Distill-Qwen-32B

文本生成 32K

DeepSeek-R1-Distill-Qwen-32B是一个基于Qwen2.5-32B的蒸馏大型语言模型，使用了DeepSeek R1的输出。

DeepSeek

查看详情

2025-02-05更新

API调用示例



腾讯云部署流程

- **Cloud Studio:** 登录腾讯云官网，进入Cloud Studio，可一键部署，即开即用，能快速体验7B以下DeepSeek - R1蒸馏模型。
- **云原生构建:** 登录腾讯云，云原生构建相关服务，一键体验DeepSeek，无需等待下载，支持1.5b/7b/8b/14b/32b多款模型。
- **TI平台:** 登录腾讯云TI平台，提供专属算力、多种尺寸模型、0代码模型部署等功能。
- **HAI:** 登录腾讯云HAI控制台，新建DeepSeek - R1应用，创建完成后可通过站内信获取密码，选择可视化界面（ChatbotUI）或命令行（JupyterLab）直接调用DeepSeek - R1模型。
- **云应用:** 登录腾讯云找到云应用服务，直接购买成品DeepSeek应用，或利用云应用定制化软件功能，快速部署AI应用或定制传统软件。
- **搜索方案:** 大模型知识引擎：注册并登录腾讯云，在控制台搜索“大模型知识引擎” 体验” 开通服务。新建应用并配置基础信息，选择DeepSeek - R1/V3模型，即可在右侧调试窗口提问搜索。
- **API接口:** 申请腾讯云DeepSeek API接口，获取API Key后，可在自己的应用或程序中通过调用API来实现DeepSeek模型的搜索功能。

腾讯云API调用

开发者首选，DeepSeek部署简单、直观、高效！

HAI 提供 GPU 算力和 ollama 环境，支持 1.5B, 7B, 8B, 32B 等多个 DeepSeek-R1 蒸馏模型。

最佳产品推荐

<p>NEW DeepSeek-V3</p> <p>是在14.8万亿高质量token上完成预训练的强大混合专家 (MoE) 语言模型，拥有6710亿参数。在知识问答、内容生成、智能客服等领域表现出色。</p> <p>查看详情 > 立即购买 ></p>	<p>NEW DeepSeek-R1-Distill-Llama-8B</p> <p>在MATH-500上达到了89.1%的准确率，在AIME2024上达到了50.4%的通过率，在CodeForces上获得了1205的评分，展示了较强的数学和编程能力。</p> <p>查看详情 > 立即购买 ></p>	<p>NEW DeepSeek-R1-Distill-Qwen-14B</p> <p>在MATH-500上达到了93.9%的准确率，在AIME2024上达到了69.7%的通过率，在CodeForces上获得了1481的评分，显示出在数学和编程领域的强大实力。</p> <p>查看详情 > 立即购买 ></p>
<p>NEW DeepSeek-R1-Distill-Qwen-32B</p> <p>在AIME 2024、MATH-500、GPQA Diamond等多个基准测试中都取得了优异成绩，其中在MATH-500上达到了94.3%的准确率，展现出强大的数学推理能力。</p> <p>查看详情 > 立即购买 ></p>	<p>NEW DeepSeek-R1-Distill-Llama-70B</p> <p>在AIME 2024、MATH-500、GPQA Diamond 等多个基准测试中都取得了优异的成绩，在数学、编程和推理等多个领域展现出强大的性能。</p> <p>查看详情 > 立即购买 ></p>	<p> 混元大模型 Hunyuan-Large</p> <p>混元大模型是腾讯自研的千亿级大模型，可应用于通用行业的对话、问答、文本生成等领域，在垂直行业场景仅需少量自训练数据量，便可达到很好的训练效...</p> <p>查看详情 > 立即体验 ></p>

企业级首选，DeepSeek部署稳定、安全、易用！

最佳产品推荐

产品配置	应用	算力方案	时长	折扣	限购	价格	日均价格
HAI-GPU基础型 限1个 搭载NVIDIA T4级GPU, 16G显存	DeepSeek-R1	GPU-16GB显存(T4级)	自购买7天内有效	1.9折	限1个	39元	约5.57元/天
HAI-GPU进阶型 限1个 搭载NVIDIA V100级GPU, 32G显存	DeepSeek-R1	GPU-32GB显存(V100级)	自购买7天内有效	0.8折	限1个	49元	约7元/天
HAI-GPU基础型 不限购 搭载NVIDIA T4级GPU, 16G显存	DeepSeek-R1	GPU-16GB显存(T4级)	7天	6折	不限购	124.29元	约17.76元/天
HAI-GPU进阶型 不限购 搭载NVIDIA V100级GPU, 32G显存	DeepSeek-R1	GPU-32GB显存(V100级)	7天	6折	不限购	366.21元	约52.32元/天

腾讯云智能全栈AI服务上架DeepSeek系列模型

产品体验



体验DeepSeek联网助手

DeepSeek满血版+联网搜索，突破预训练数据的时间边界，提供时效精准的智能问答服务

[立即体验对话](#)



创建DeepSeek应用

DeepSeek-R1、V3等多种模型任选，通过大模型知识引擎平台快捷搭建并发布，实现与现有业务无缝对接

[创建应用（请使用电脑端创建应用）](#)



调用DeepSeek API

提供DeepSeek-R1、V3两款模型快速接入，API服务更稳定、安全、易用；灵活组建企业专属 AI 服务

[立即前往](#)

创建应用

- 应用管理
- 原子能力
- 插件中心
- 体验中心
- 系统管理

插件中心

插件能帮助知识引擎开发应用拓展功能，支持有开发能力者丰富应用内容，增强知识引擎应用的实用性。

类型 全部插件

+ 创建插件

搜索插件



网页解析 官方 限时免费

网页解析

@知识引擎官方 | 含1个工具



科学计算 官方 限时免费

利用计算逻辑对需要进行计算的问题进行运算。适用于需要进行运算的场景，如金融计算、应...

@知识引擎官方 | 含1个工具



知识库问答 官方 限时免费

从知识库中检索内容，并通过大模型润色回复

@知识引擎官方 | 含1个工具



百科 官方 限时免费

查询百科信息

@知识引擎官方 | 含1个工具



混元搜索 官方 限时免费

利用搜索引擎搜索相关内容，并通过混元大模型进行总结回复

@知识引擎官方 | 含1个工具



新闻 官方 限时免费

提供全球各类新闻，为用户及时了解新闻内容提供保障

@知识引擎官方 | 含1个工具



天气 官方 限时免费

提供国内区县级以上的地区主流城市的天气查询

@知识引擎官方 | 含1个工具



代码解释器 官方 限时免费

支持大模型代码生成，并运行 Python 代码，适用于数值计算、数据分析、图表生成、图片处...

@知识引擎官方 | 含1个工具



图片理解 官方 限时免费

图片理解

@知识引擎官方 | 含1个工具

没有更多数据了

华为云+siliconflow

通过创建API key 调用



邀请码: z0sl0urY

邀请链接:

<https://cloud.siliconflow.cn/i/z0sl0urY>

The screenshot displays the SiliconFlow model marketplace. On the left is a navigation sidebar with options like '模型广场', '模型微调', '体验中心', '客户管理', '活动中心', and '我的邀请'. The main area shows a grid of model cards, each with a title, description, and pricing. The models listed include:

- deepseek-ai/DeepSeek-R1 (16 / M Tokens)
- deepseek-ai/DeepSeek-V3 (8 / M Tokens)
- Pro/deepseek-ai/DeepSeek-R1 (16 / M Tokens)
- deepseek-ai/DeepSeek-R1-Distill-Llama-70B (4.13 / M Tokens)
- deepseek-ai/DeepSeek-R1-Distill-Qwen-32B (1.26 / M Tokens)
- deepseek-ai/DeepSeek-R1-Distill-Qwen-14B (0.7 / M Tokens)
- deepseek-ai/DeepSeek-R1-Distill-Llama-8B (免费)
- deepseek-ai/DeepSeek-R1-Distill-Qwen-7B (免费)
- deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B (免费)
- deepseek-ai/Janus-Pro-7B (免费)
- Qwen/QVQ-72B-Preview (9.9 / M Tokens)
- FunAudioLLM/CosyVoice2-0.5B (50 / M UTF-8 bytes)
- deepseek-ai/DeepSeek-V2.5 (1.33 / M Tokens)
- meta-llama/Llama-3.3-70B-Instruct (4.13 / M Tokens)

华为云模型微调

◆对话模型微调

◆生图模型微调

The screenshot displays the SiliconFlow 'Model Fine-tuning' (模型微调) interface. The left sidebar contains navigation options: Model (模型), Model Mall (模型广场), Model Fine-tuning (模型微调), Experience Center (体验中心), and Account Management (账户管理). The main content area shows a 'New Fine-tuning Task' (新建微调任务) button and a task card for 'black-forest-labs/FLUX.1-dev' with the file 'lora-train.zip'. The task status is 'Completed' (已完成) and 'Style' (风格). Training parameters include: TriggerWord: asuna, Number of Repeats: 10, Number of Epochs: 3, Batch Size: 1, and Learning Rate: 0.0001. The task was completed on 2025-02-07 at 18:58:46 (44m 36s). The fine-tuning model (微调模型) is listed as 'v3qywmdvhc: asunaLorayppauadpgcokbqtwkbeazepoch_3.safetensors'. Training statistics show 1260 total steps and a cost of ¥40.3200. A task ID 'yppauadpgcokbqtwkbea' is also visible.

模型微调工程

DeepSeek: 推理模型 | 推理能力蒸馏



推理模型蒸馏到小模型

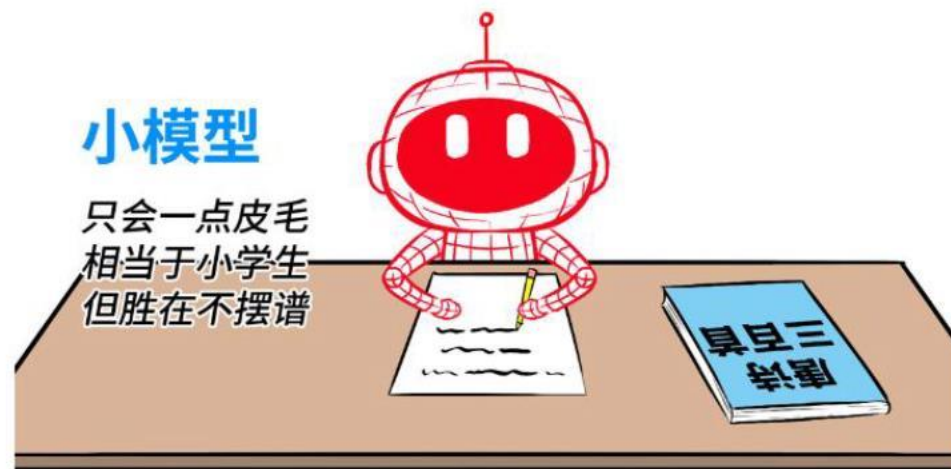
- reasoning能力可以蒸馏到小模型
- 大模型蒸馏到小模型优于小模型直接通过大规模RL训练
- 再次验证了模型规模在AGI发展中的重要性
- 推理者同样需要规模支撑

模型蒸馏工程

老师模型和学生模型

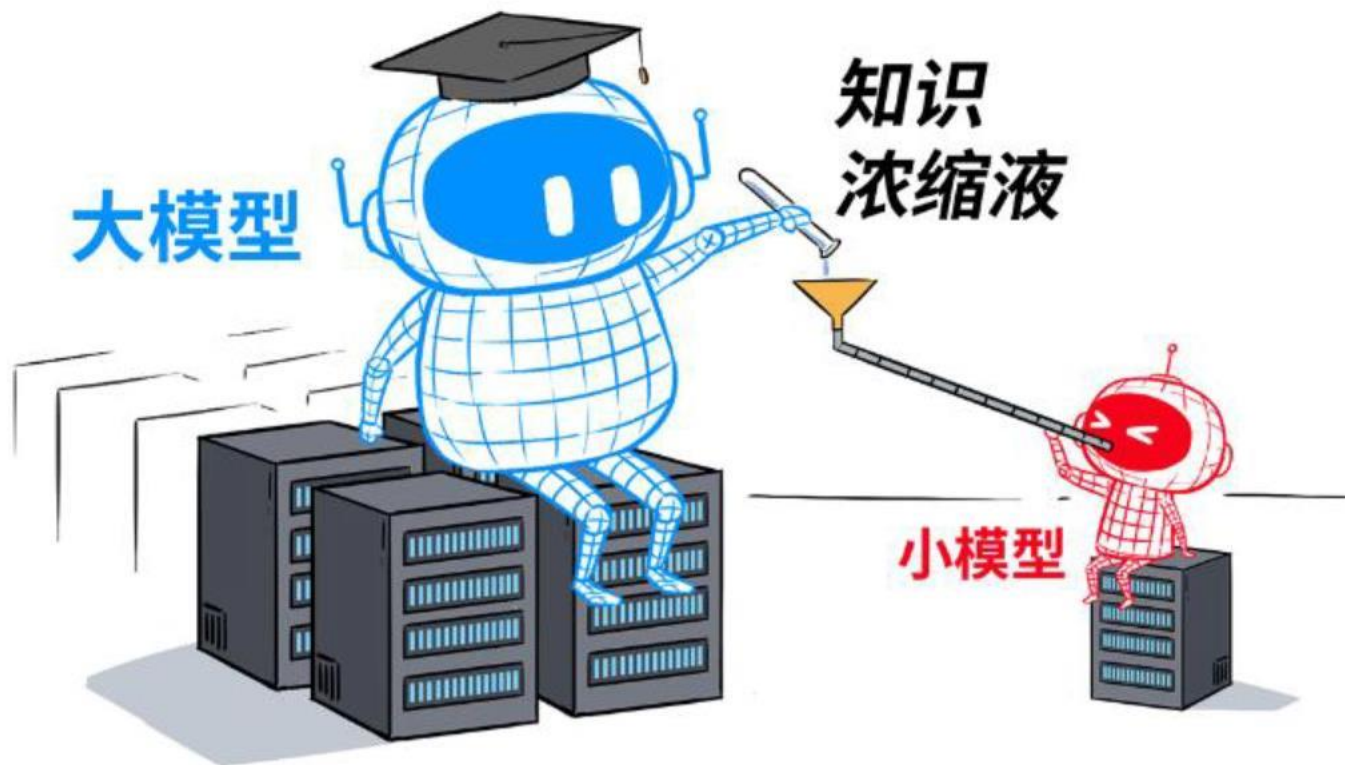


大模型：像一位见多识广、知识储备庞大的“大教授”，无所不知，但是“供养”他很贵。



小模型：相当于一枚小学生，知识面非常有限，但是胜在没教授那么大谱，给个板凳坐着就够了（部署成本低，推理成本低）。

什么是模型蒸馏？



智力水平 ^①	基础“学生”模型
幼儿园	Qwen2.5-Math-1.5B
小学生	Qwen2.5-Math-7B
小学生	Llama-3.1-8B
初中生	Qwen2.5-14B
高中生	Qwen2.5-32B
大学生	Llama-3.3-70B-Instruct



蒸馏后得到的“新模型”
DeepSeek-R1-Distill-Qwen-1.5B
DeepSeek-R1-Distill-Qwen-7B
DeepSeek-R1-Distill-Llama-8B
DeepSeek-R1-Distill-Qwen-14B
DeepSeek-R1-Distill-Qwen-32B
DeepSeek-R1-Distill-Llama-70B

“模型蒸馏”就是把大模型学到的本领，用“浓缩”的方式教给小模型的过程，在保证一定精度的同时，大幅降低运算成本和硬件要求。

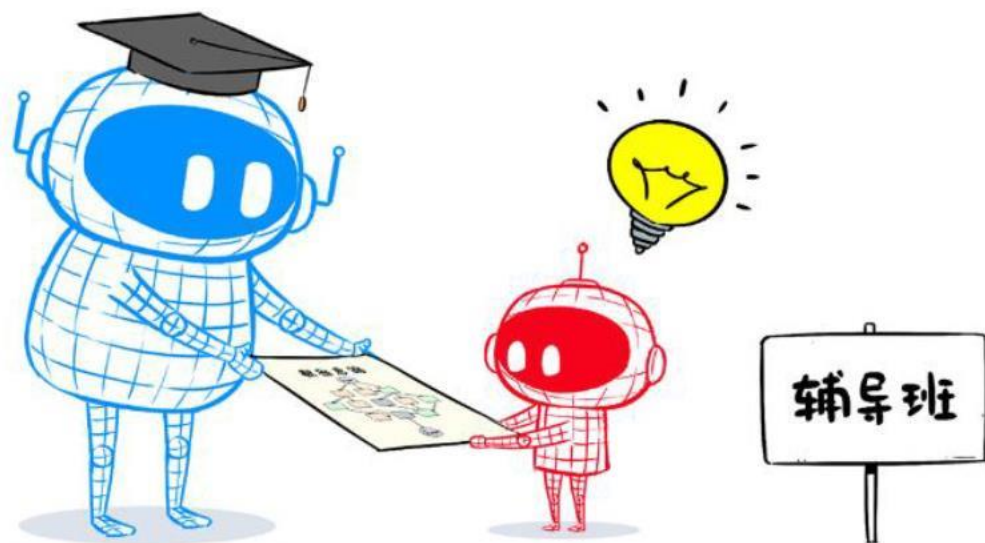
模型蒸馏

- **蒸馏**是一种机器学习技术，其中较小的模型（“学生模型”）被训练来模仿较大、预训练模型（“教师模型”）的行为。
- **1. 数据蒸馏** 在数据蒸馏中，教师模型生成合成数据或伪标签，然后这些数据用于训练学生模型。
- **2. Logits蒸馏** 在logits蒸馏中，学生模型被训练来匹配教师模型的logits，而不仅仅是最终的预测。这种方法保留了更多关于教师模型置信水平和决策过程的信息。
- **3. 特征蒸馏** 特征蒸馏涉及将教师模型中间层的知识转移到学生模型中。通过对齐两个模型的隐藏表示，学生模型可以学习到更丰富和更抽象的特征。

蒸馏、微调、RAG

这是蒸馏

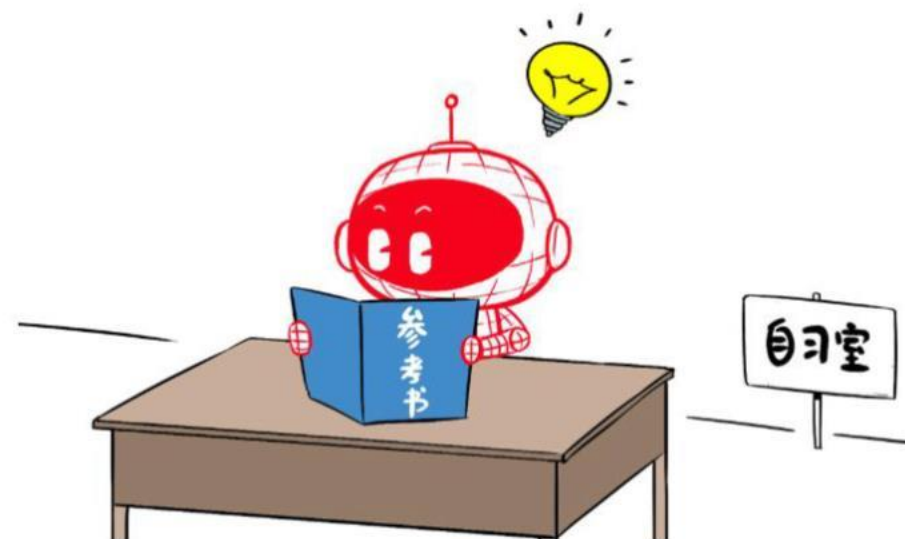
这题我不会，但是在老师的辅导下，看了老师的解题思路和答案，我顿悟了，以后这类题我都会



蒸馏：是学生通过模仿老师的解题思路，达到和老师相似的知识水平。

这是微调

这题我不会，但是我找到了参考书，通过自学参考书，我也顿悟了，以后这类题我也都会

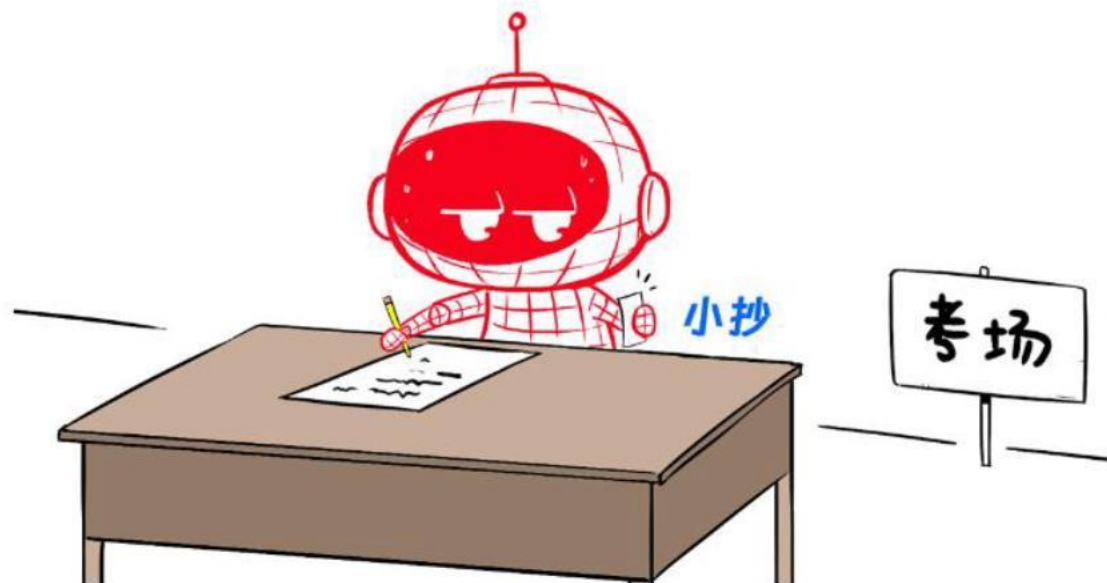


微调：又叫精调，相当于学生意识到自己某门课有短板，然后自己找参考书恶补了一下，从而补上短板。

蒸馏、微调、RAG

这是RAG

这题我不会，考试的时候我带了小抄，通过看小抄，我也能正确答题，但下次如果没小抄我还是不会~



RAG: 直译过来叫做“检索增强生成”。相当于这题我不会，但是我有“小抄”，我回答的时候，就看一眼小抄，然后再综合我脑子里的已有知识，进行回答。

DeepSeek本地部署

DeepSeek本地部署-硬件环境准备

1. GPU要求

推荐NVIDIA显卡 (RTX 3090/4090/A100等)

显存要求:

- 7B模型: 至少16GB显存
- 13B/20B模型: 至少24GB显存
- 70B模型: 需多卡并行 (如2xA100)

支持CUDA 11.7+ 和 cuDNN 8.5+

2. 系统要求

Linux (Ubuntu 20.04+) 或 Windows WSL2

Python 3.8+, 推荐使用Anaconda/Miniconda
管理环境

DeepSeek本地部署-部署办法

方法1: 使用Hugging Face Transformers

```
bash 复制  
  
# 1. 创建虚拟环境  
conda create -n deepseek python=3.10  
conda activate deepseek  
  
# 2. 安装依赖  
pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118  
pip install transformers accelerate sentencepiece  
  
# 3. 下载模型 (需申请权限)  
git lfs install  
git clone https://huggingface.co/deepseek-ai/deepseek-llm-7b-base  
  
# 4. 推理代码示例  
from transformers import AutoTokenizer, AutoModelForCausalLM  
model = AutoModelForCausalLM.from_pretrained("deepseek-llm-7b-base", device_map="auto")  
tokenizer = AutoTokenizer.from_pretrained("deepseek-llm-7b-base")  
inputs = tokenizer("DeepSeek is", return_tensors="pt").to("cuda")  
outputs = model.generate(**inputs, max_length=50)  
print(tokenizer.decode(outputs[0]))
```

方法2: 使用vLLM加速推理

```
bash 复制  
  
# 1. 安装vLLM  
pip install vllm  
  
# 2. 启动API服务  
python -m vllm.entrypoints.api_server \  
  --model deepseek-ai/deepseek-llm-7b-base \  
  --tensor-parallel-size 2 # 多卡并行时指定GPU数量  
  
# 3. 调用API  
curl http://localhost:8000/generate \  
  -d '{  
    "prompt": "DeepSeek的优势是",  
    "max_tokens": 100  
  }'
```

方法3: Docker部署

```
bash 复制  
  
# 使用官方Docker镜像  
docker run -it --gpus all -p 7860:7860 \  
  -v /path/to/models:/models \  
  deepseekai/deepseek-llm:latest \  
  --model /models/deepseek-7b \  
  --quantization int8 # 支持int8/int4量化
```

Anything LLM接入

- **下载安装**• 打开官网，根据自己的系统选择下载对应的版本。
- **安装文件**，按照安装向导的提示进行操作，默认路径安装或修改默认安装路径都可以。安装完成后，点击完成自动跳转到AnythingLLM界面。
- **模型配置**• LLM Selection（大语言模型选择）：选择Ollama，如果本地已通过Ollama部署了DeepSeek等模型，AnythingLLM会自动检测。这意味着模型和聊天记录仅在运行Ollama模型的机器上可访问，保证了数据的安全性和隐私性。• Embedding Preference（嵌入偏好）：使用AnythingLLM Embedder嵌入工具，文档文本会在AnythingLLM的实例上私密嵌入，文本数据的处理和转换在本地进行，不会泄露给第三方。• Vector Database（向量数据库）：使用LanceDB作为向量数据库，向量和文档文本都存储在AnythingLLM实例上，进一步确保数据的私密性和安全性。
- **接入DeepSeek模型**• 通过Ollama接入：先在本地通过Ollama部署DeepSeek模型，然后在AnythingLLM中选择Ollama作为语言模型推理后端，AnythingLLM会自动连接到本地Ollama服务中的DeepSeek模型。• 通过DeepSeek R1 API接入：在AnythingLLM工作区右侧点击设置，选择聊天设置，更改LLM模型为DeepSeek，输入API Key并选择DeepSeek R1模型，点击更新工作区即可。
- **搭建本地知识库**• 在AnythingLLM左侧工作区找到上传按钮，选中要上传的文档，点击移动到工作区，再点击“Save and Embed”，对文档进行切分和词向量化。• 完成后，点击图钉按钮，将文档设置为当前对话的背景文档，即可在聊天窗口基于上传的文档内容与模型进行对话。

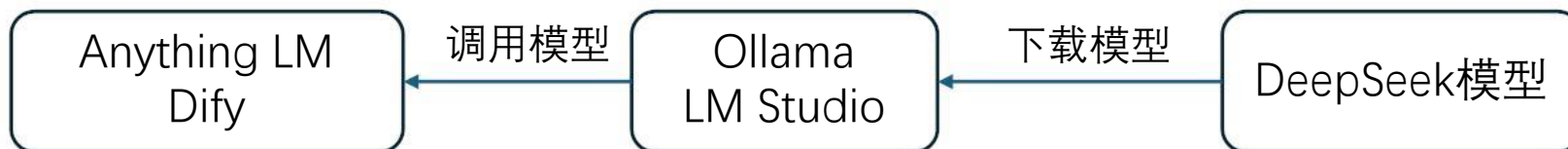
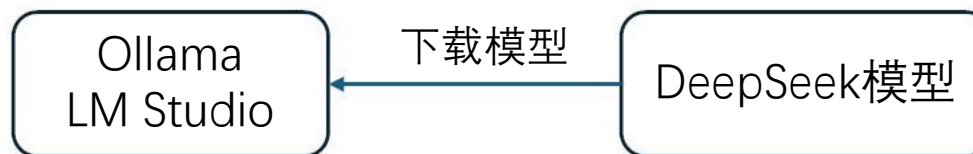
本地部署方法

➤ Ollama+Deepseek

➤ LM Studio+DeepSeek

➤ Anything LM+Ollama+Deepseek +向量模型 (Embessing Model)

➤ Dify +Ollama +Deepseek+向量模型 (Embessing Model)



Ollama



Get up and running with large language models.

Run [Llama 3.3](#), [DeepSeek-R1](#), [Phi-4](#), [Mistral](#), [Gemma 2](#), and other models, locally.

Download ↓

Available for macOS, Linux, and Windows

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

[1.5b](#) [7b](#) [8b](#) [14b](#) [32b](#) [70b](#) [671b](#)

↓ 14.9M Pulls Updated 6 days ago

7b

29 Tags

```
ollama run deepseek-r1
```



Discord

GitHub

Models

Search models

Sign in

Download

All

Embedding

Vision

Tools

Popular

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

[1.5b](#) [7b](#) [8b](#) [14b](#) [32b](#) [70b](#) [671b](#)

↓ 14.9M Pulls 29 Tags Updated 6 days ago

llama3.3

New state of the art 70B model. Llama 3.3 70B offers similar performance compared to the Llama 3.1 405B model.

[tools](#) [70b](#)

↓ 1.2M Pulls 14 Tags Updated 2 months ago

phi4

Phi-4 is a 14B parameter, state-of-the-art open model from Microsoft.

[14b](#)

↓ 465.9K Pulls 5 Tags Updated 5 weeks ago

llama3.2

Meta's Llama 3.2 goes small with 1B and 3B models.

[tools](#) [1b](#) [3b](#)

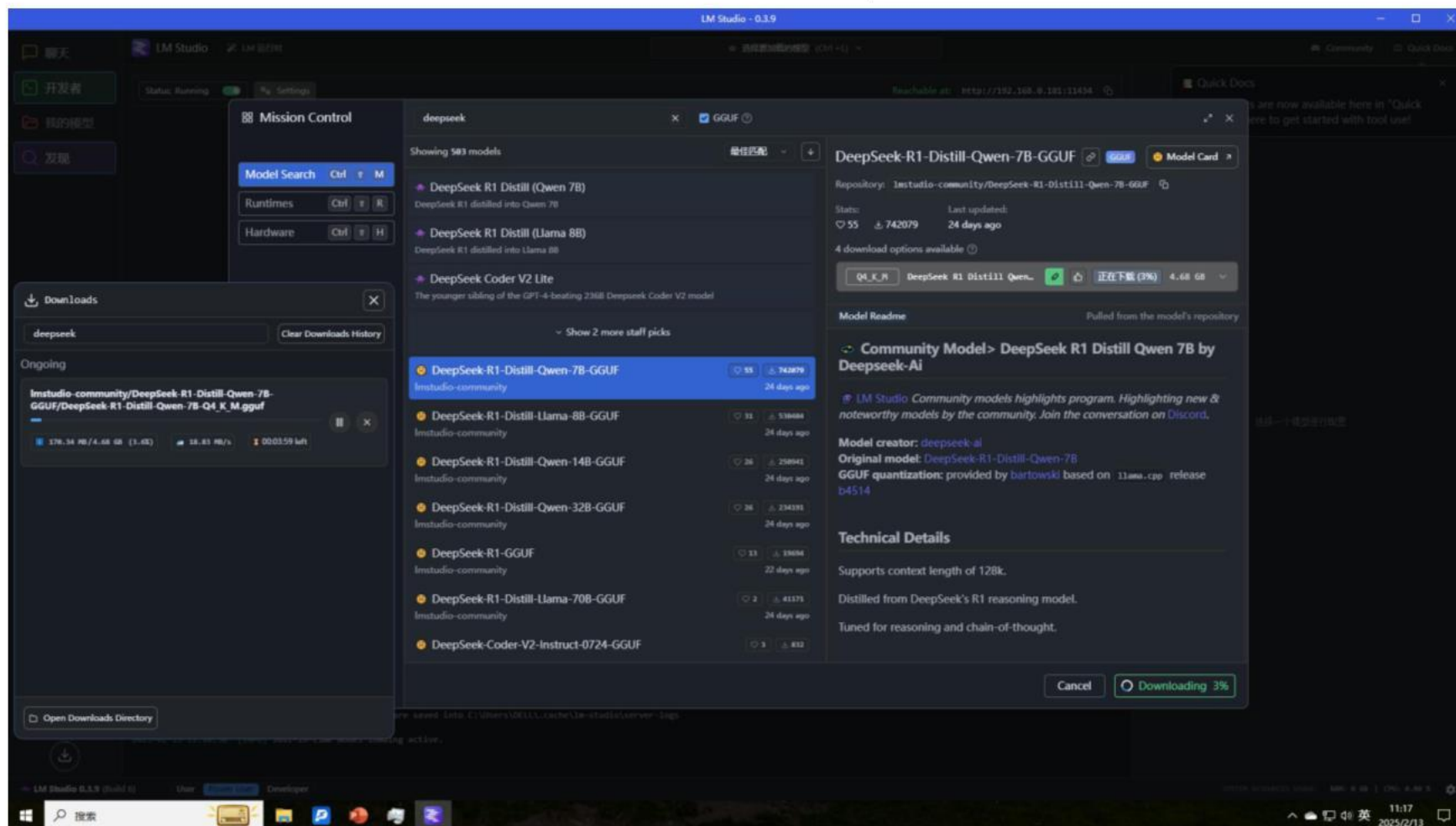
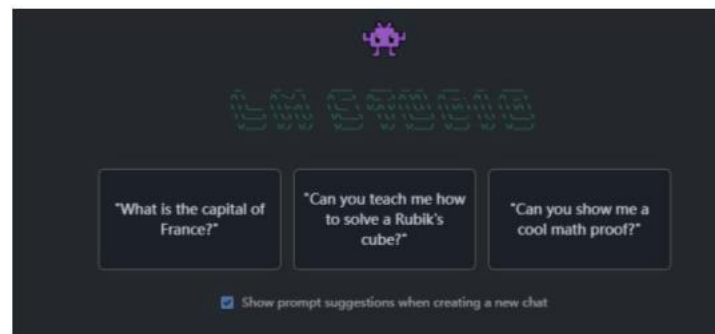
↓ 8.6M Pulls 63 Tags Updated 4 months ago

LM Studio本地部署

- **前期准备：**确保设备满足一定硬件要求，显卡需GTX 1060（6GB）及以上，推荐RTX3060及以上；内存8GB及以上，推荐16GB及更高；C盘剩余20GB空间，建议使用NVMe固态硬盘。
- **安装LM Studio客户端：**在官网lmstudio.ai下载对应操作系统的安装包，双击运行并按提示完成安装。启动后，点击右下角设置图标将语言改为简体中文。
- **加载模型：**自行下载获取模型，下载到本地后，点击左上方文件夹图标选择模型目录导入；若不会找模型，可在LM Studio的设置里，勾选“Use LM Studio's Hugging Face”，再点击左上方搜索图标搜索deepseek选择模型。注意参数规模不同，性能和硬件需求有差异。例如，1.5B适合体验尝鲜，7B适合普通创作及开发测试，8B适合对内容要求更高的场景，14B适合专业及深度内容创作。

LM Studio+DeepSeek

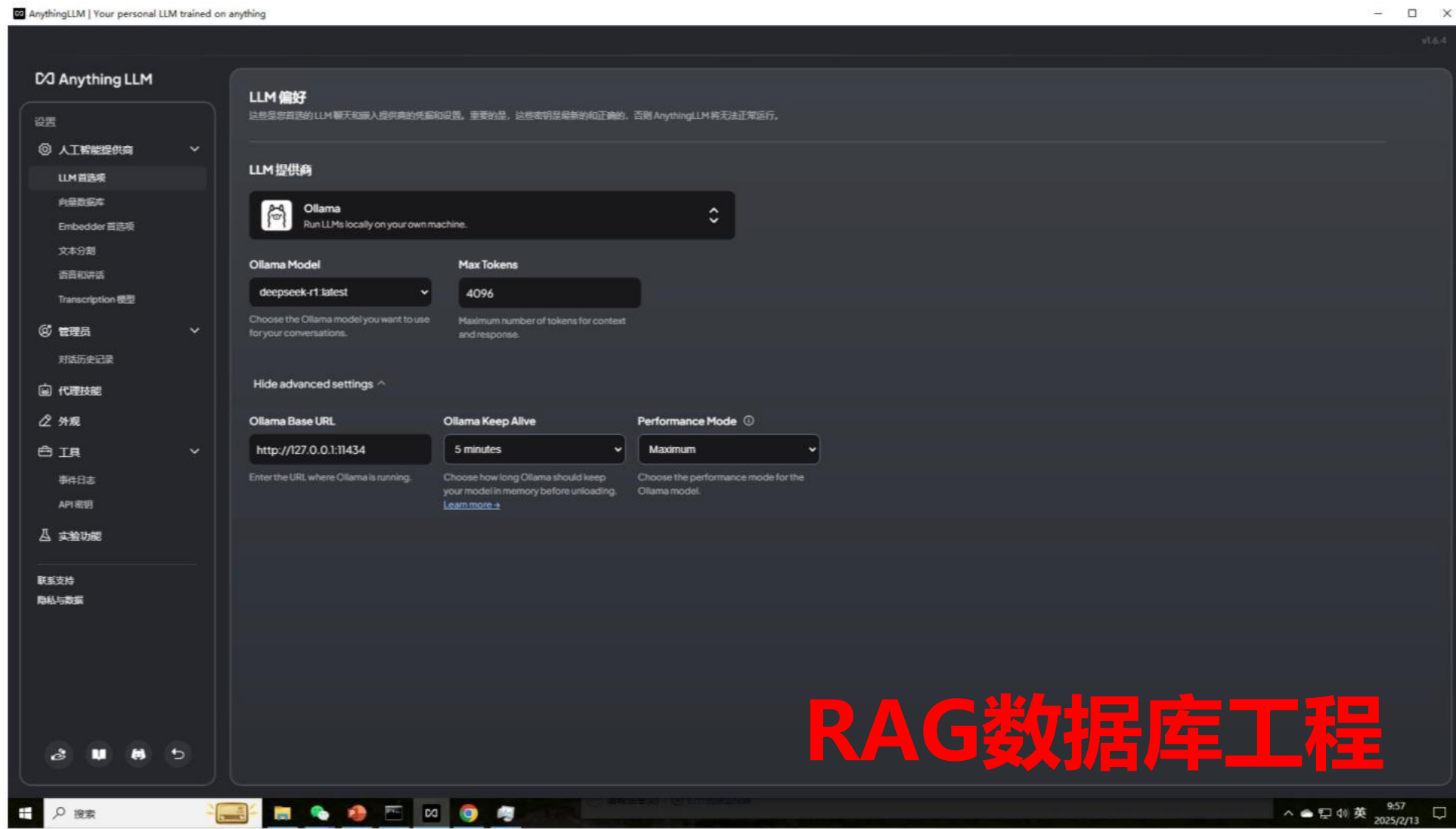
LM Studio可以通过huggaface下载Deepseek各种蒸馏模型 (GGUF)



DeepSeek UI客户端使用

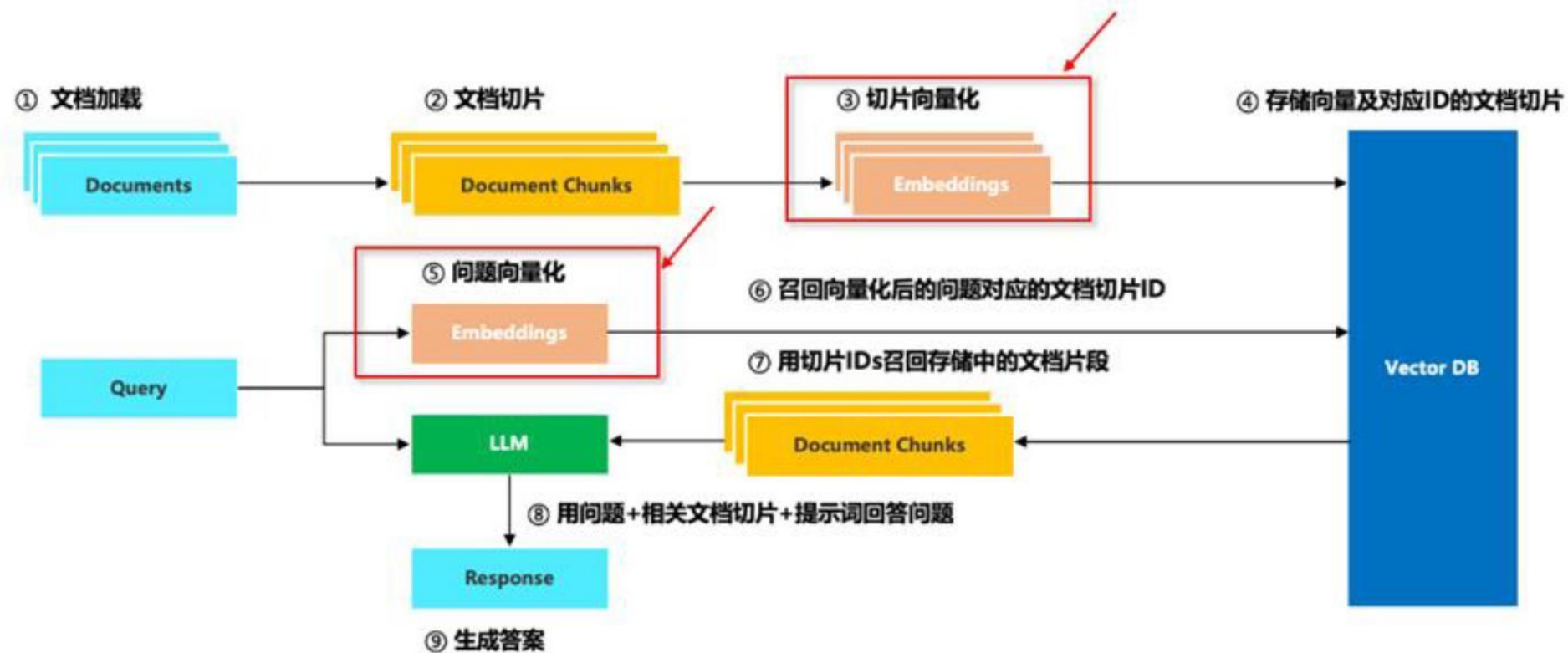
Anything LM+Ollama+Deepseek

LLM首选项
向量数据库
Embedder首选项
文本分割



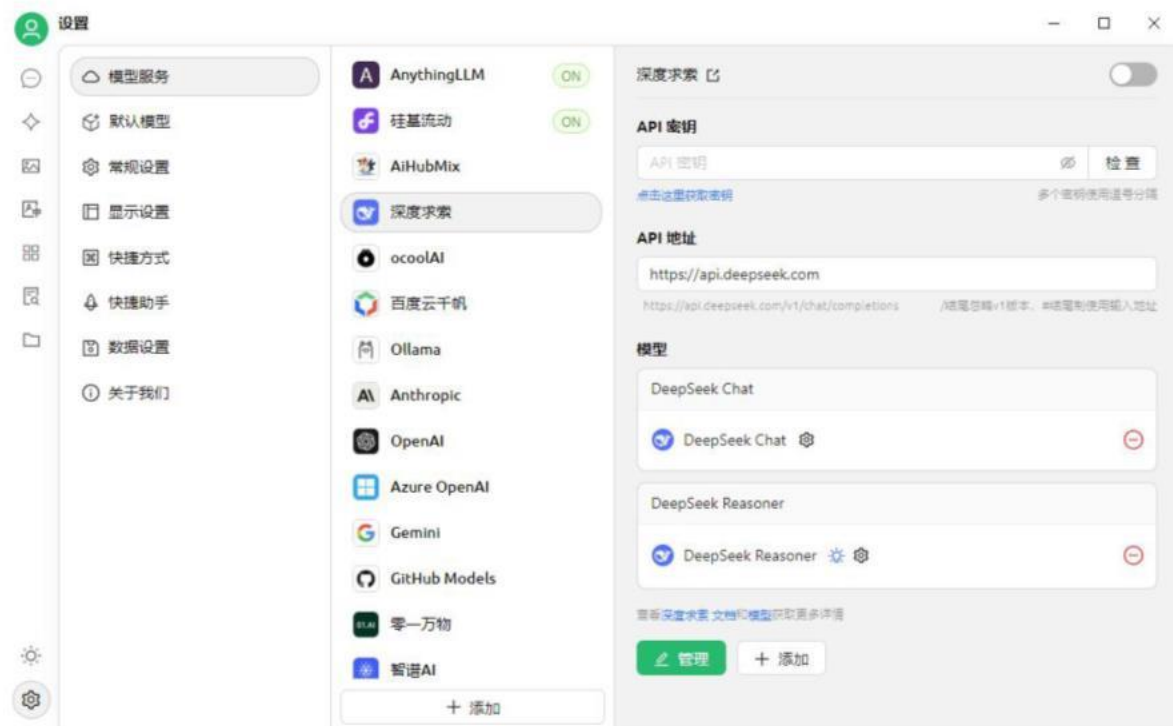
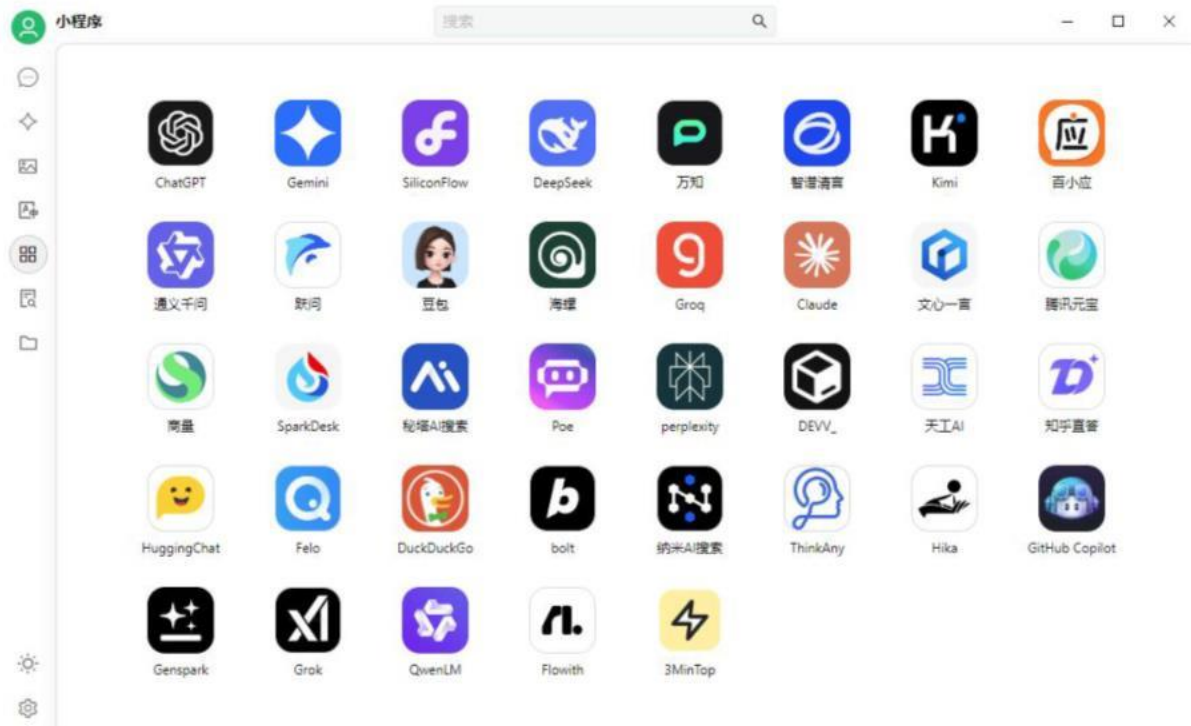
RAG数据库工程

RAG中的Embedding模型



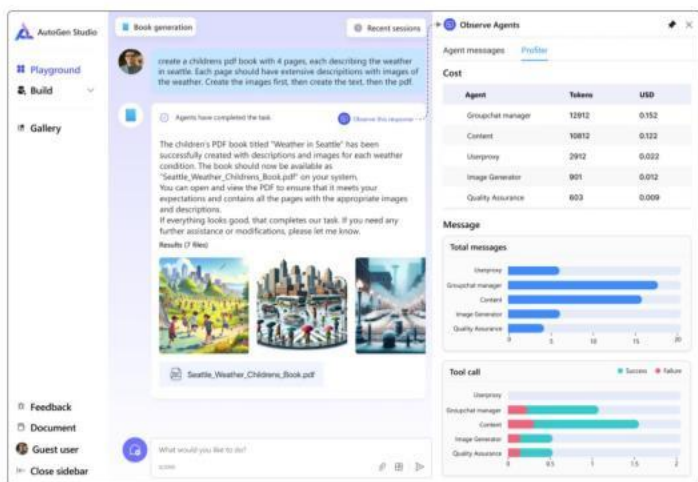
Embedding模型贯穿RAG流程多环节

Cherry LM+DeepSeek API



大模型智能体工程

五大 AI Agent 多智能体开发框架



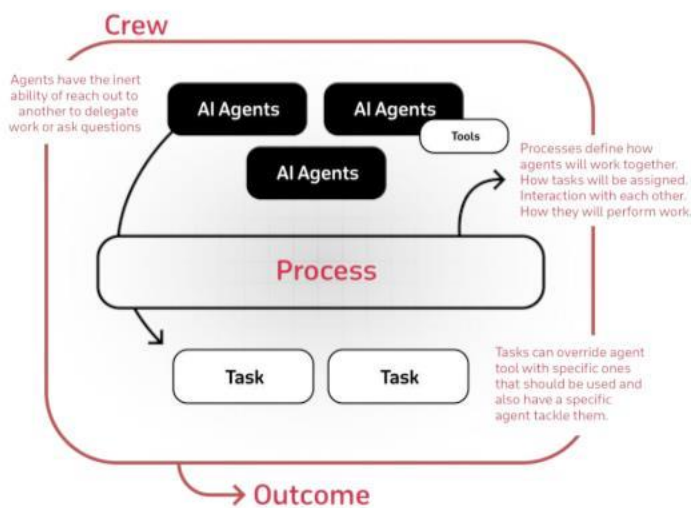
微软AutoGen:

Github 地址: <https://github.com/microsoft/autogen>



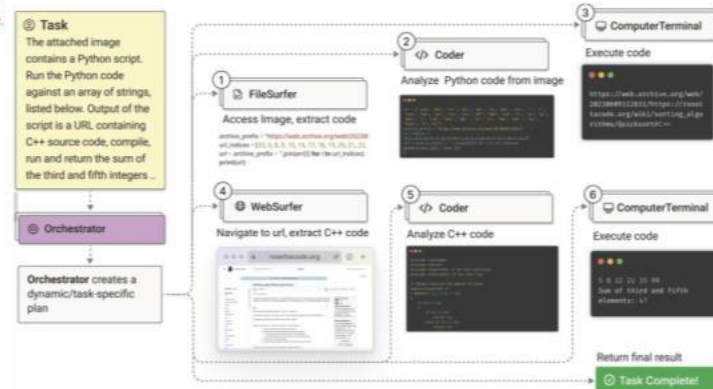
LangGraph:

Github 地址: <https://github.com/langchain-ai/langgraph>



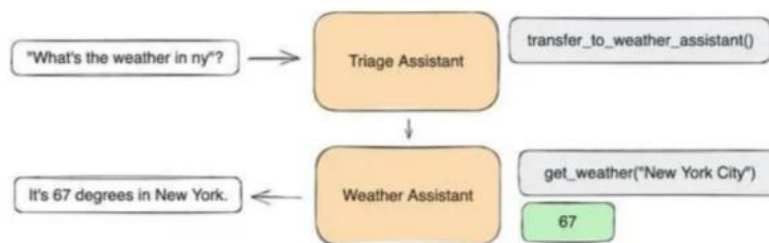
CrewAI

Github 地址: <https://github.com/crewAIInc/crewAI>



微软Magentic-One

Github 地址:
<https://github.com/microsoft/autogen/tree/main/python/packages/autogen-magentic-one>



OpenAI Swarm:

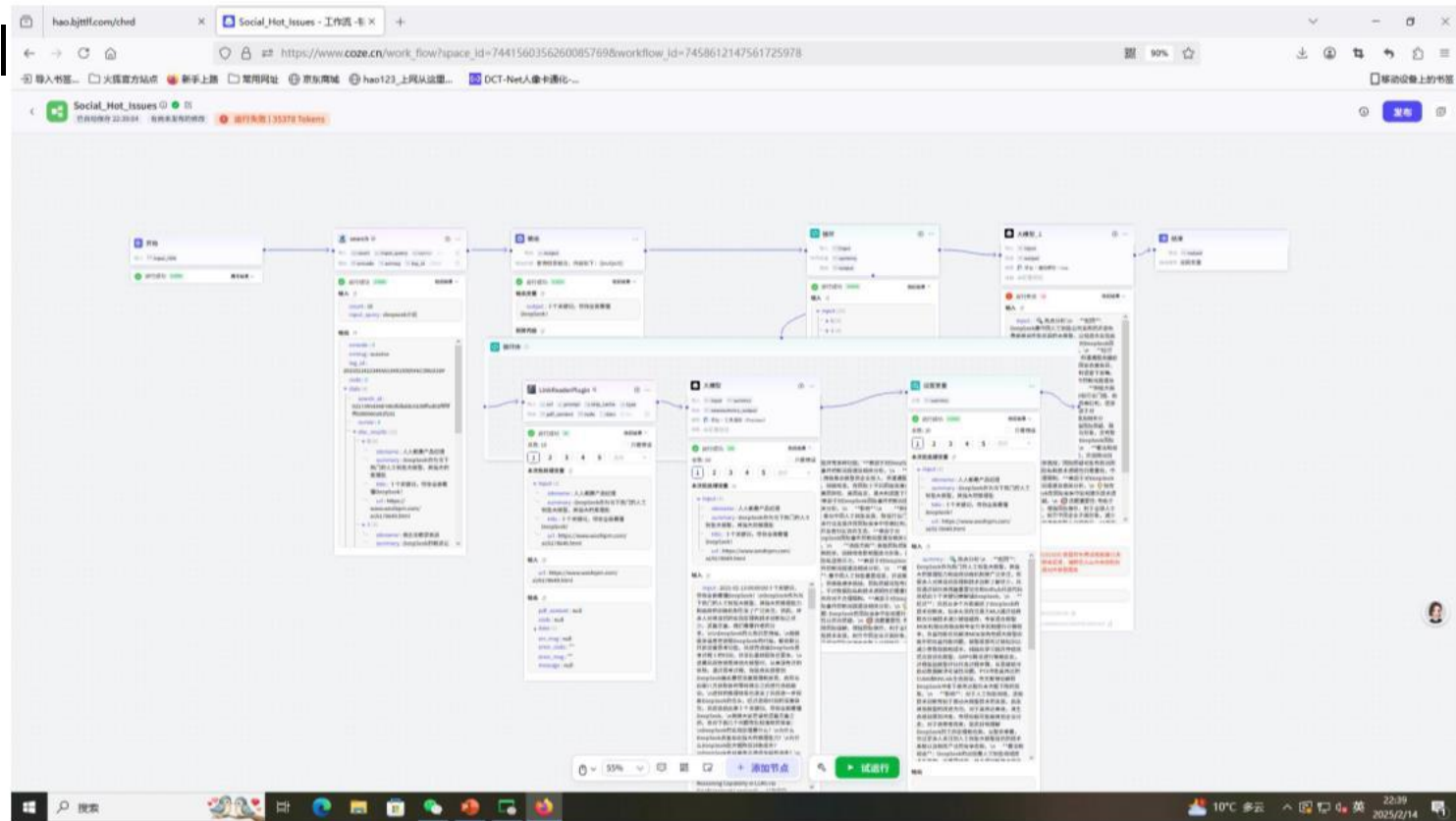
Github 地址: <https://github.com/openai/swarm>

DeepSeek业务流应用

DeepSeek业务流应用

- Coze+Deepseek API (Deepseek R1) 业务流
- Dify+Deepseek API搭建本地知识库
- Langchain+Deepseek API

业务流工程



2025七款RAG

- **1. R2R:** 官网: <https://r2r-docs.sciphi.ai/>
- 功能: 用于开发本地RAG应用程序, 支持多模态内容摄取、混合搜索、知识图谱构建、GraphRAG, 具备用户管理、可观测性、配置及可视化交互等功能, 适用于处理动态数据和复杂实体关系的应用程序。
- **2. Cognita:** 官网: <https://cognita.truefoundry.com/>
- 功能: 模块化框架, 提供定制化管道, 用于构建可扩展且生产就绪的RAG应用程序, 适合企业大规模AI应用。
- **3. LLMWare:** 官网: <https://llmware.ai/>
- 功能: 提供统一框架构建基于LLM的应用, 如RAG和代理, 包含RAG Pipeline和50 + 小型专用模型, 适用于定制特定领域解决方案。
- **4. LangChain:** 官网: <https://www.langchain.com/>
- 功能: 构建于大语言模型之上, 通过链式调用组合组件, 简化开发者与语言模型交互, 可用于构建问答系统等, 支持与多种数据源集成。
- **5. deepset的Haystack**
- 官网: <https://haystack.deepset.ai/>
- 功能: 用于构建基于NLP的搜索和问答系统, 提供模块化架构、多模型支持、高效检索、问答功能, 可扩展性强, 适用于问答系统和文档检索应用。
- **6. LlamaIndex**
- 官网: <https://docs.llamaindex.ai/>
- 功能: 简化外部知识库与LLM的集成, 支持动态数据摄取和自定义索引, 可与多种预训练语言模型配合使用。
- **7. txtai**
- GitHub主页: <https://github.com/neuml/txtai>
- 功能: 一体化嵌入数据库, 提供自然语言搜索、可自定义管道, 具备轻量级高效的特点, 适合快速设置和有强大搜索需求的项目。

DeepSeek 赋能各行业的应用场景

• 1. DeepSeek赋能酒企场景

- 市场分析与消费者洞察
- 智能营销与个性化推荐
- 生产优化与质量控制
- 供应链管理与库存优化
- 品牌故事与文化传播
- 智能客服与消费者互动
- 新品研发与口味预测

• 2. DeepSeek赋能房产装修

- 智能设计与方案生成
- 虚拟现实（VR）与增强现实（AR）体验
- 施工进度管理与优化
- 材料采购与供应链优化
- 客户需求分析与个性化服务
- 智能客服与售后支持
- 装修质量检测与问题预测
- 市场趋势分析与营销策略优化

XU

• 3. DeepSeek赋能知识产权服务

- 专利检索与分析
- 侵权检测与风险评估
- 知识产权法律咨询
- 专利撰写与优化
- 知识产权交易与估值
- 知识产权管理与监控
- 知识产权培训与教育
- 知识产权战略规划

• 4. DeepSeek赋能金融行业

- 智能投顾与投资分析
- 风险管理与欺诈检测
- 客户服务与智能客服
- 信用评分与贷款审批
- 市场趋势分析与预测
- 自动化交易与算法交易
- 金融产品设计 with 定价
- 合规管理与报告生成

• 5. DeepSeek赋能跨境贸易

- 市场分析与需求预测
- 供应链优化与物流管理
- 跨境支付与汇率风险管理
- 合规管理与海关申报
- 多语言客户服务与沟通
- 贸易伙伴匹配与谈判支持
- 风险管理与信用评估
- 跨境电商营销与推广

• 6. DeepSeek赋能健康医美

- 个性化治疗方案设计
- 客户管理与智能咨询
- 市场分析与需求预测
- 治疗效果预测与优化
- 智能营销与个性化推荐
- 健康数据管理与分析
- 医美产品设计与优化
- 合规管理与报告生成

• 7. DeepSeek赋能智库咨询机构

- 智能研究与数据分析
- 政策分析与预测
- 行业趋势分析与市场预测
- 智能报告生成与优化
- 客户服务与智能咨询
- 知识管理与信息检索
- 多语言支持与跨文化分析
- 风险管理与决策支持



DeepSeek职场应用风险

风险/问题类别	具体风险描述	注意事项
数据隐私泄露	使用DeepSeek处理敏感信息可能导致数据泄露，尤其是在缺乏严格的数据保护措施时	确保所有涉及敏感信息的操作都在安全环境下进行，使用加密技术保护传输中的数据，并定期审计数据访问权限
模型偏见	模型可能无意中学习到训练数据中的偏见，导致输出结果存在不公平或歧视性内容	对于关键决策，应结合人类判断进行二次确认；定期更新和优化模型以减少潜在偏见
内容安全问题	生成的内容可能包含不当信息、仇恨言论或其他违反法律法规的内容	设定明确的内容过滤规则，对AI生成的内容进行人工审核；教育员工识别不适当内容
用户输入不确定性	用户输入的模糊性或恶意操作可能导致AI输出错误或有害信息	提供清晰的指导给用户如何正确使用AI工具；实施严格的用户身份验证机制
法律合规风险	不当使用DeepSeek可能触犯知识产权法、隐私法等相关法律法规	在使用前咨询法律顾问，确保遵守相关法律法规；了解并遵循开源协议的要求
供应链安全	依赖第三方组件和服务可能存在安全隐患，影响整体系统的安全性	定期评估第三方服务的安全性和可靠性；采用最新的安全防护措施
模型滥用	可能被用于非法目的，如网络犯罪、诈骗等	制定严格的使用政策和监控机制，防止模型被滥用
性能限制	尽管DeepSeek R1表现出色，但仍有其性能边界，特别是在处理超出其训练数据范围的任务时	明确模型的能力边界，对于超出能力的任务寻求其他解决方案
成本考量	虽然R1是免费的，但本地部署和运行大型语言模型可能会产生显著的成本	计算成本效益分析，选择最适合业务需求的部署方案
伦理道德挑战	AI的广泛使用可能引发一系列伦理问题，如就业替代、个人隐私权侵犯等	建立内部伦理审查委员会，定期评估AI应用对社会的影响

如何利用DeepSeek实现人机高效协作?



谢 谢!

2025.2.15