



Chatting or Acting?

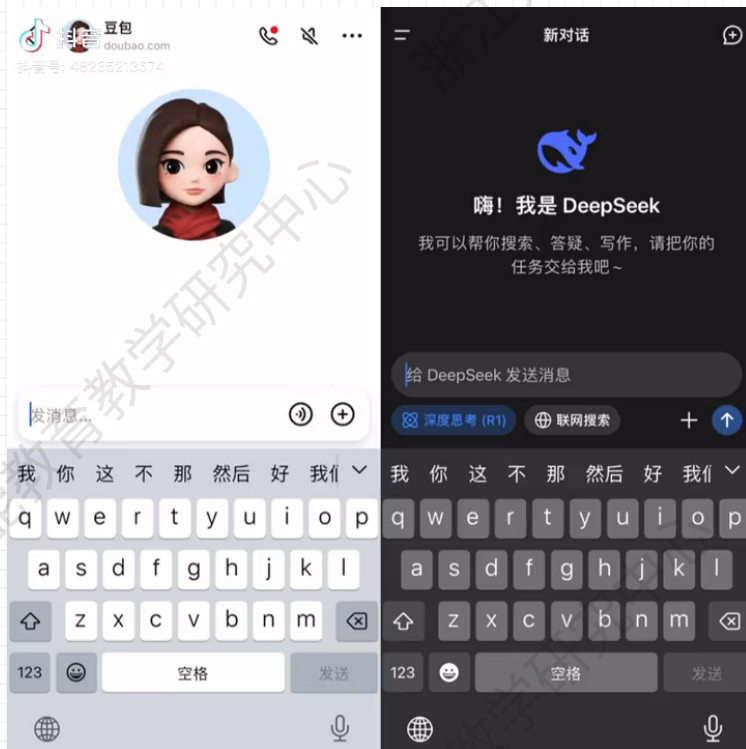
—DeepSeek的突破边界与浙大先生的未来图景

陈文智

浙江大学信息技术中心

浙江大学人工智能教育教学研究中心

2025年02月17日



我也想来一套，在线用or本地部署？

明天中午前，我要看到DS接入！

用OpenAI训练的吧？吹！

这就是传说中的国运级产品吗？

我刚刚开发的智能体能用吗？

这玩意儿凭啥这么强？

成本这么低，西湖之光不用了吧？

Agent是不是也要变强了？



deepseek



DeepSeek关键词



浙江大学
ZHEJIANG UNIVERSITY



DeepSeek

DeepDrink

需求刚起，
方案已至。
灵感闪现，
原型立现

VS

热情款待，
商谈愉快
深入交流，
合作共赢



01

DeepSeek突破边界

Chatting or Acting

—— DeepSeek的突破边界与浙大先生的未来图景



DeepSeek席卷全球

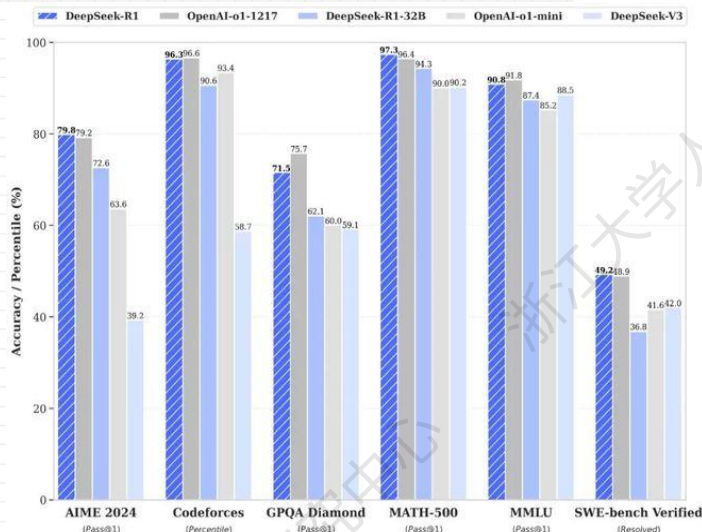


浙江大学
ZHEJIANG UNIVERSITY



deepseek

引爆全球，高性能、低成本的国产、开源大模型！



全球排名一览

设备 iPhone iPad 榜单类型 免费 畅销 类别 总榜 应用 效率



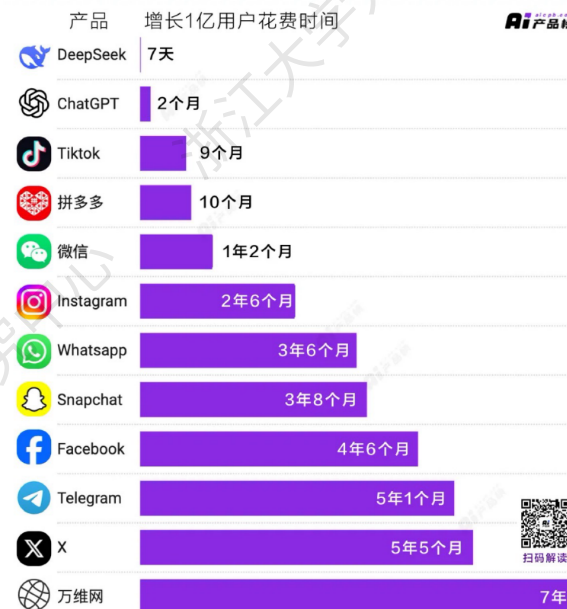
排名 地区 数量

中国大陆 #1	中国香港 #1	中国台湾 #1
澳大利亚 #1	加拿大 #1	玻利维亚 #1
开曼群岛 #1	英属维尔京群岛 #1	白俄罗斯 #1
俄罗斯 #1	中国澳门 #1	巴巴亚新几内亚 #1
巴基斯坦 #1	不丹 #1	柬埔寨 #1
老挝人民民主共和国 #1	马来西亚 #1	尼泊尔 #1
帕劳 #1	斯里兰卡 #1	文莱 #1
新加坡 #1	阿尔及利亚 #1	阿拉伯联合酋长国 #1
阿曼 #1	安哥拉 #1	巴林 #1
博茨瓦纳 #1	布基纳法索 #1	津巴布韦 #1
卡塔尔 #1	肯尼亚 #1	马达加斯加 #1
马拉维 #1	毛里求斯 #1	毛里塔尼亚 #1
莫桑比克 #1	纳米比亚 #1	塞拉利昂 #1
塞内加尔 #1	斯威士兰 #1	坦桑尼亚联合共和国 #1
索马里 #1	乌干达 #1	亚美尼亚 #1

近期因开源AI大模型和相关技术火爆全球，DeepSeek一度在140多个国家的应用商店下载排行首位。

超级产品

增长1亿用户所用的时间



注：DeepSeek 包含网站Web/应用App累加不去重，TikTok 不包含国内版抖音

DeepSeek—有史以来最快获得1亿注册用户的APP。

- DeepSeek-R1已发布并开源，性能对比OpenAI o1正式版。
- 在目前大模型主流榜单中，DeepSeek-V3在开源模型中位列榜首，与世界上最先进的闭源模型不分伯仲。

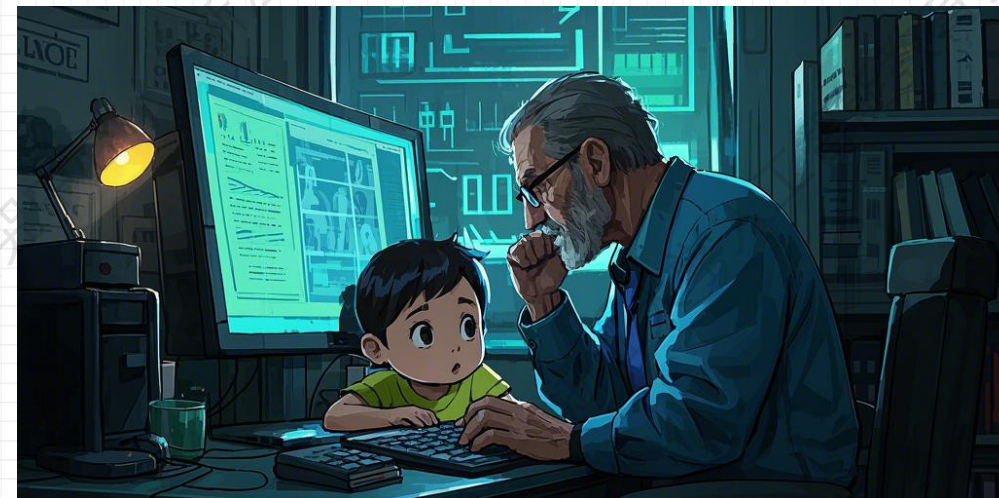
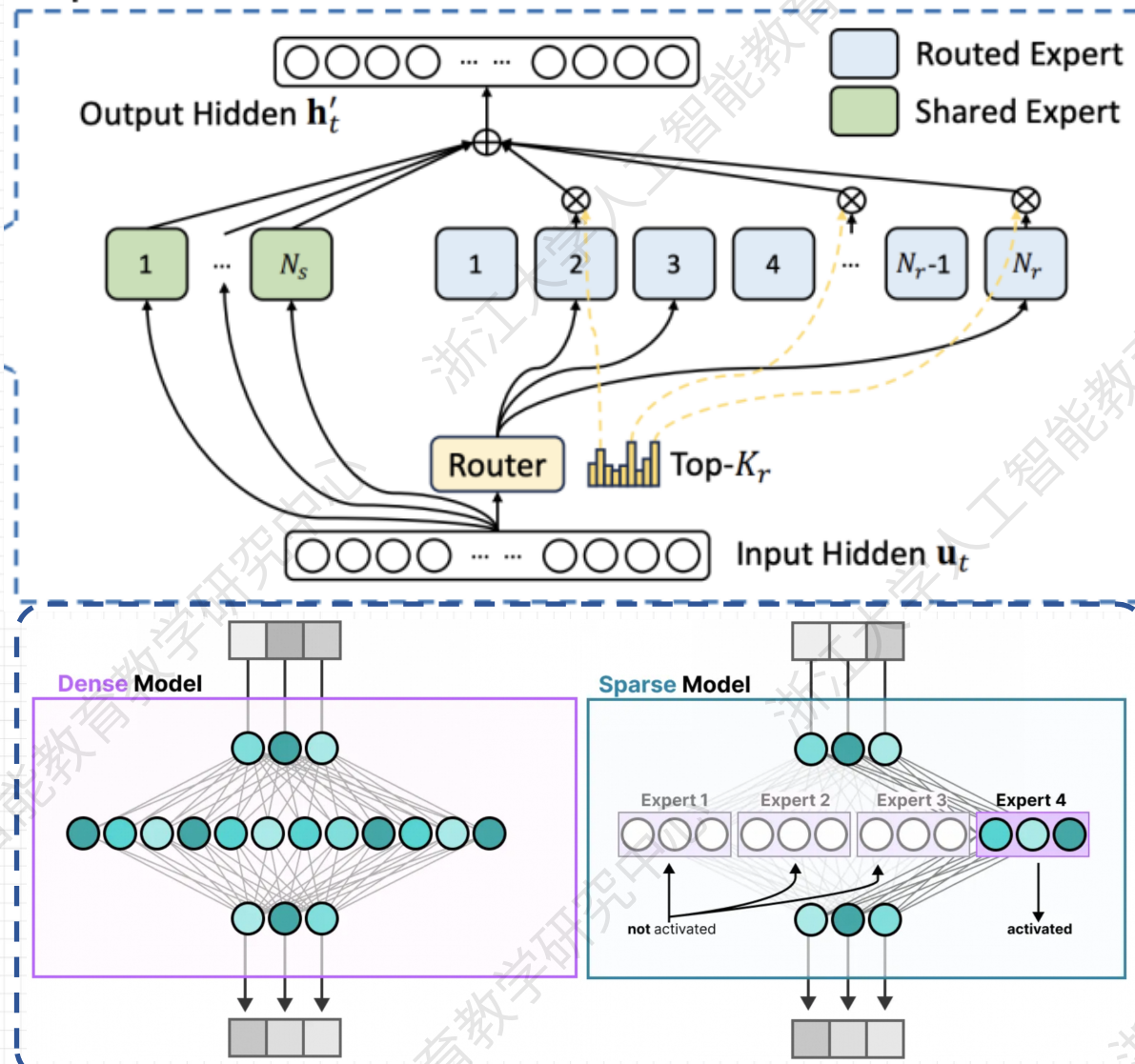
DeepSeek模型架构创新



浙江大学
ZHEJIANG UNIVERSITY

——采用MoE架构 并解决路由崩溃难题

DeepSeekMoE



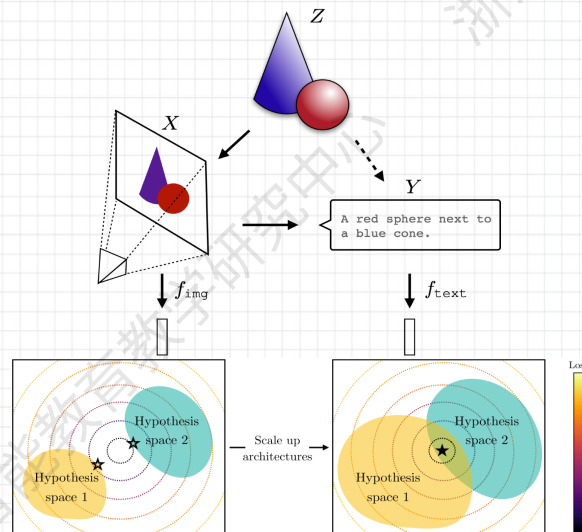
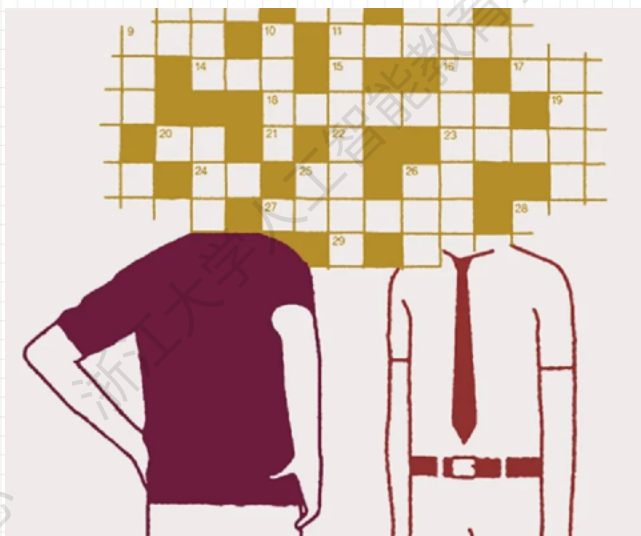
MOE

DeepSeek模型架构创新

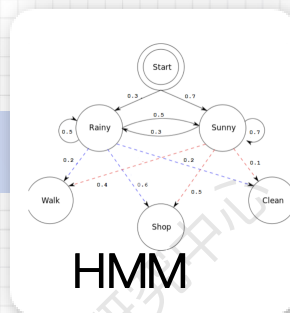
——MLA多头潜在注意力机制降低成本、提高效率



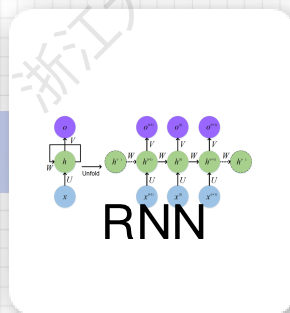
浙江大学
ZHEJIANG UNIVERSITY



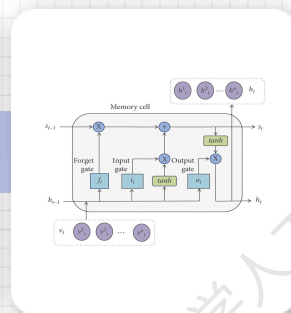
柏拉图表征假说



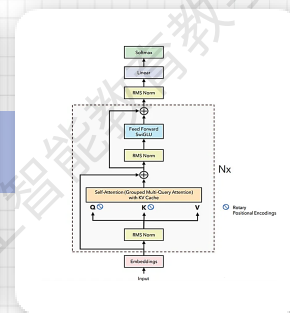
隐马尔可夫链
(HMM)



神经网络时代
(RNN)



神经网络时代
(LSTM)



Transformer时代
(Attention)

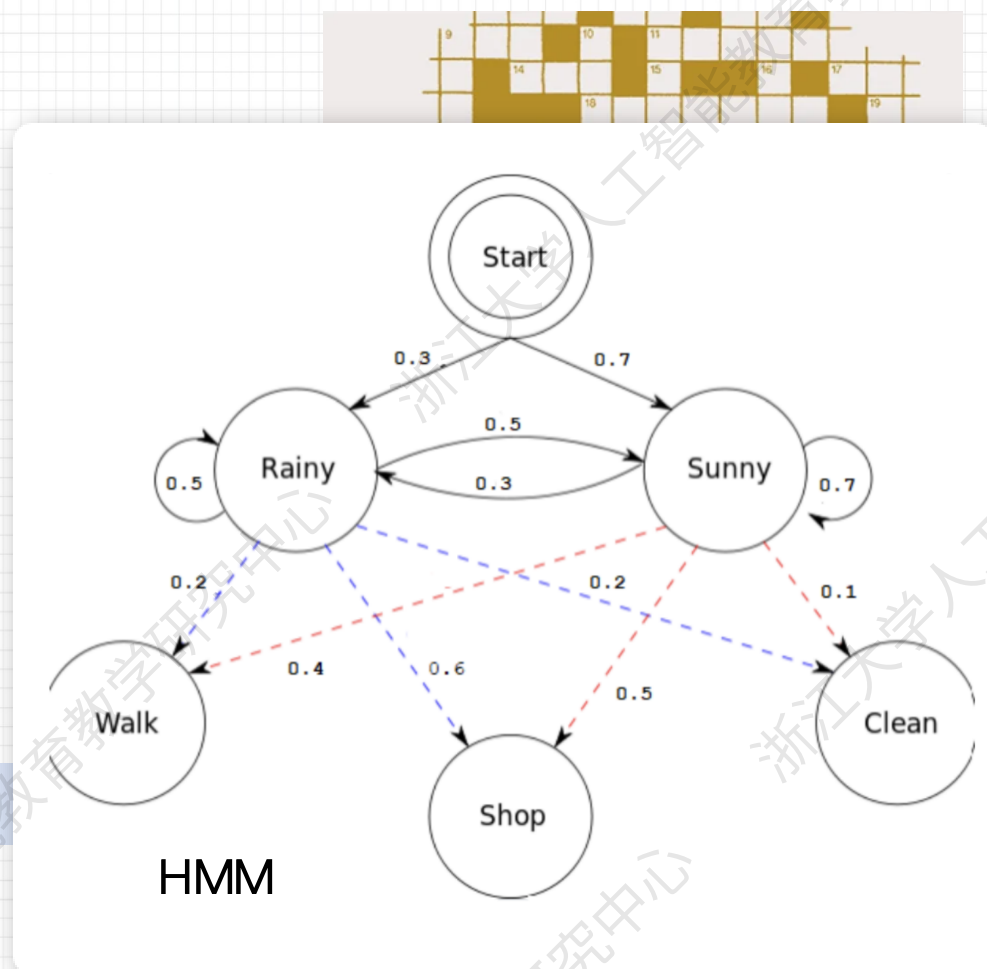


DeepSeek模型架构创新

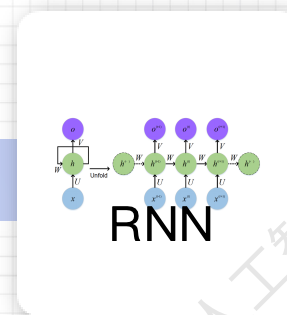
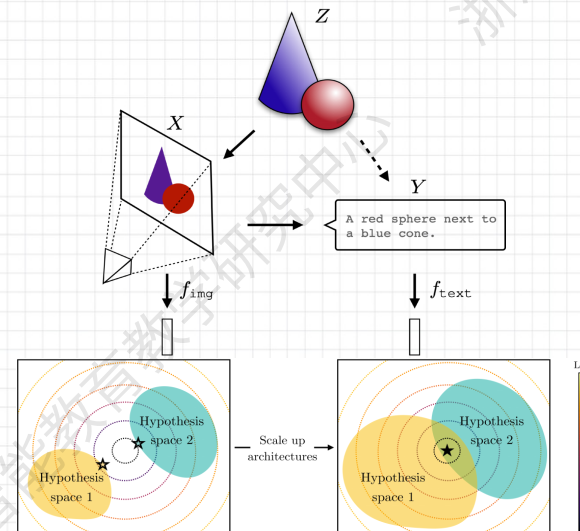
——MLA多头潜在注意力机制降低成本、提高效率



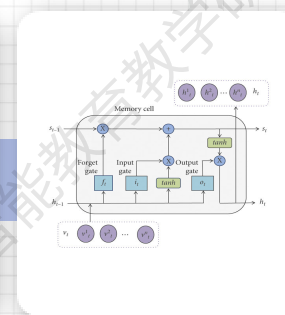
浙江大学
ZHEJIANG UNIVERSITY



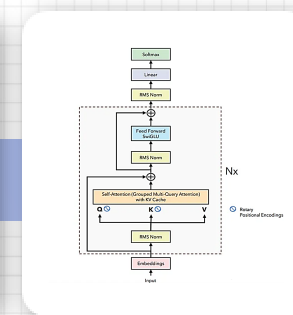
隐马尔卡夫链
(HMM)



神经网络时代
(RNN)



神经网络时代
(LSTM)



Transformer时代
(Attention)

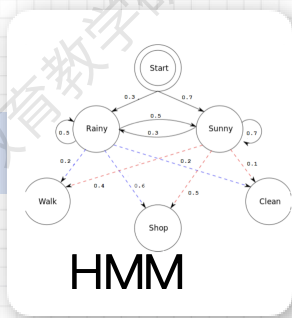
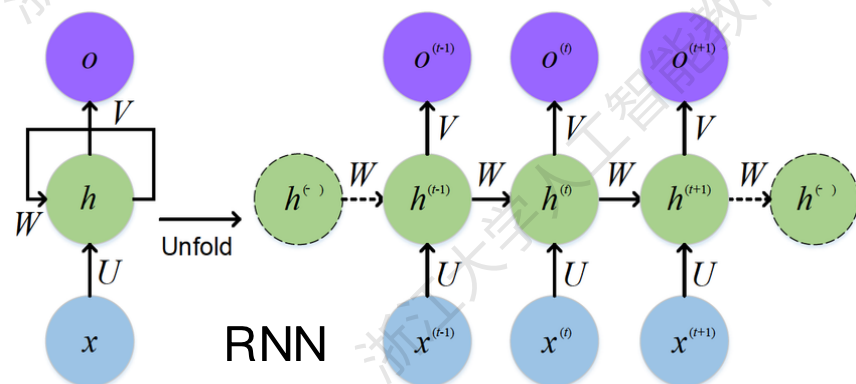
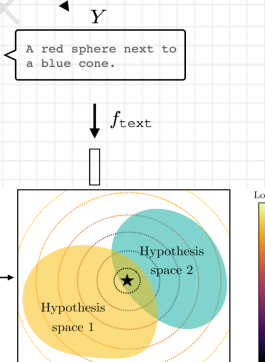
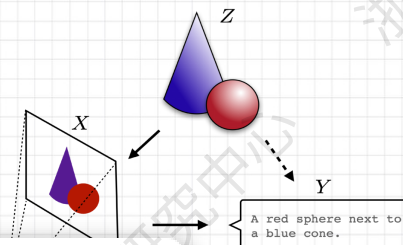


DeepSeek模型架构创新

——MLA多头潜在注意力机制降低成本、提高效率

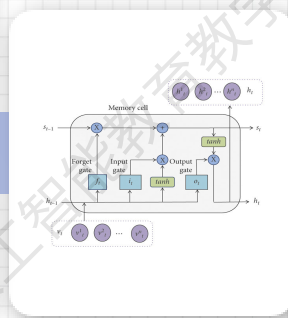


浙江大学
ZHEJIANG UNIVERSITY

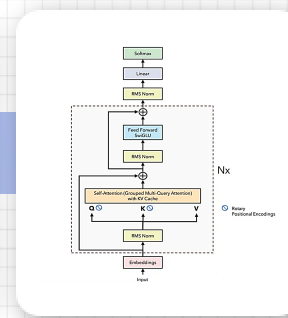


隐马尔卡夫链
(HMM)

神经网络时代
(RNN)



神经网络时代
(LSTM)



Transformer时代
(Attention)

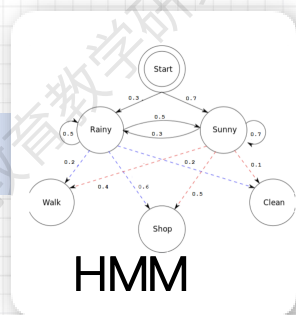
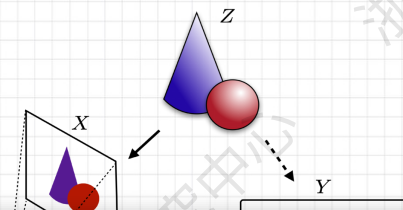
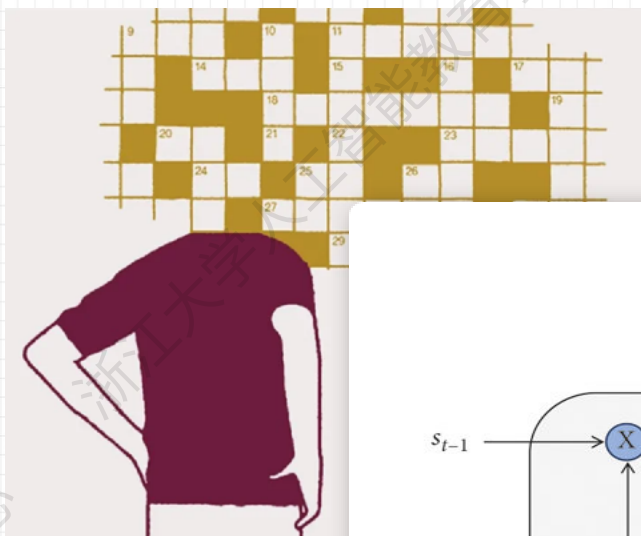


DeepSeek模型架构创新

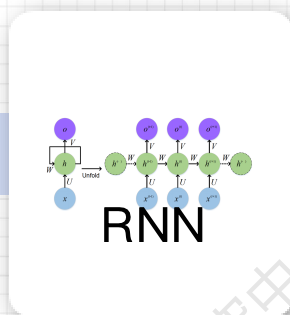
——MLA多头潜在注意力机制降低成本、提高效率



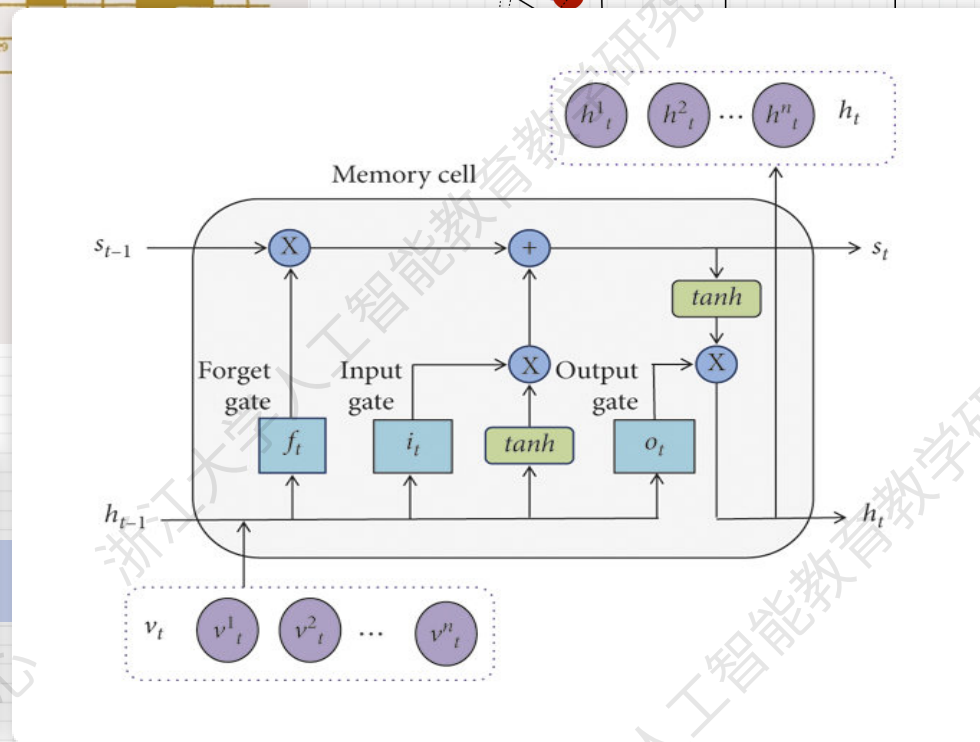
浙江大学
ZHEJIANG UNIVERSITY



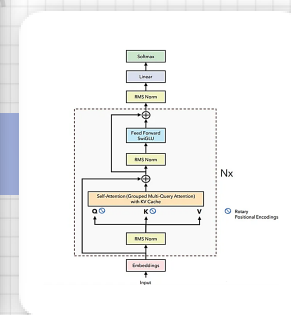
隐马尔可夫链
(HMM)



神经网络时代
(RNN)



神经网络时代
(LSTM)



Transformer时代
(Attention)

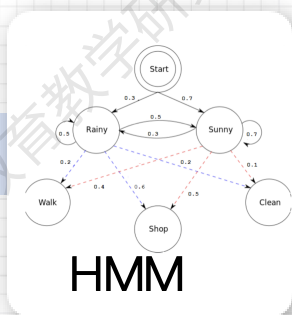
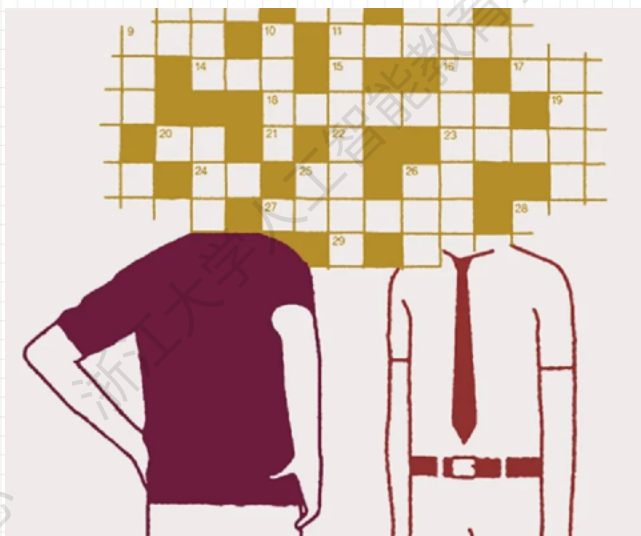


DeepSeek模型架构创新

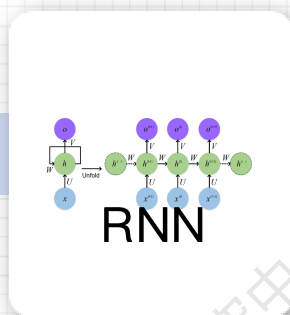
——MLA多头潜在注意力机制降低成本、提高效率



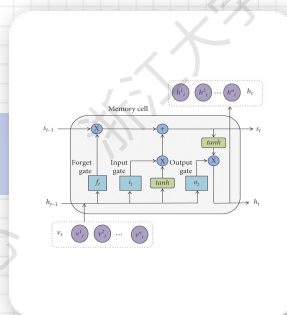
浙江大学
ZHEJIANG UNIVERSITY



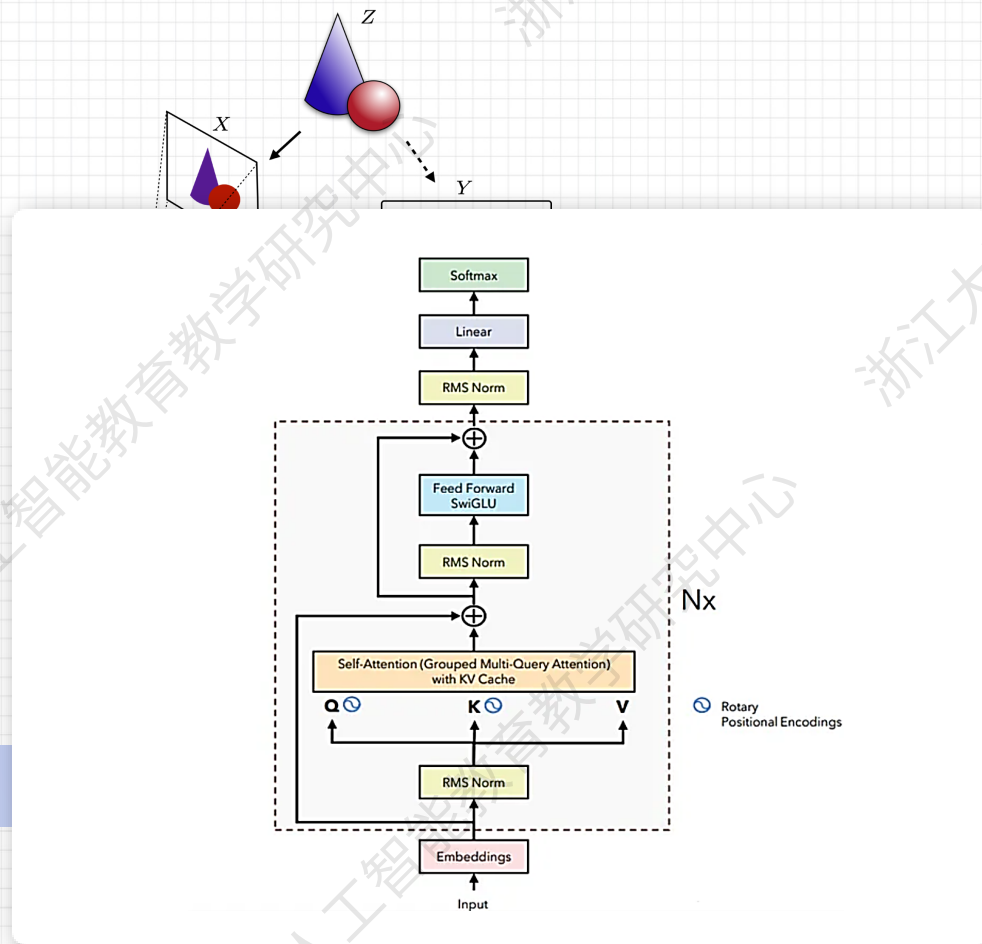
隐马尔可夫链
(HMM)



神经网络时代
(RNN)



神经网络时代
(LSTM)



Transformer时代
(Attention)



DeepSeek模型架构创新

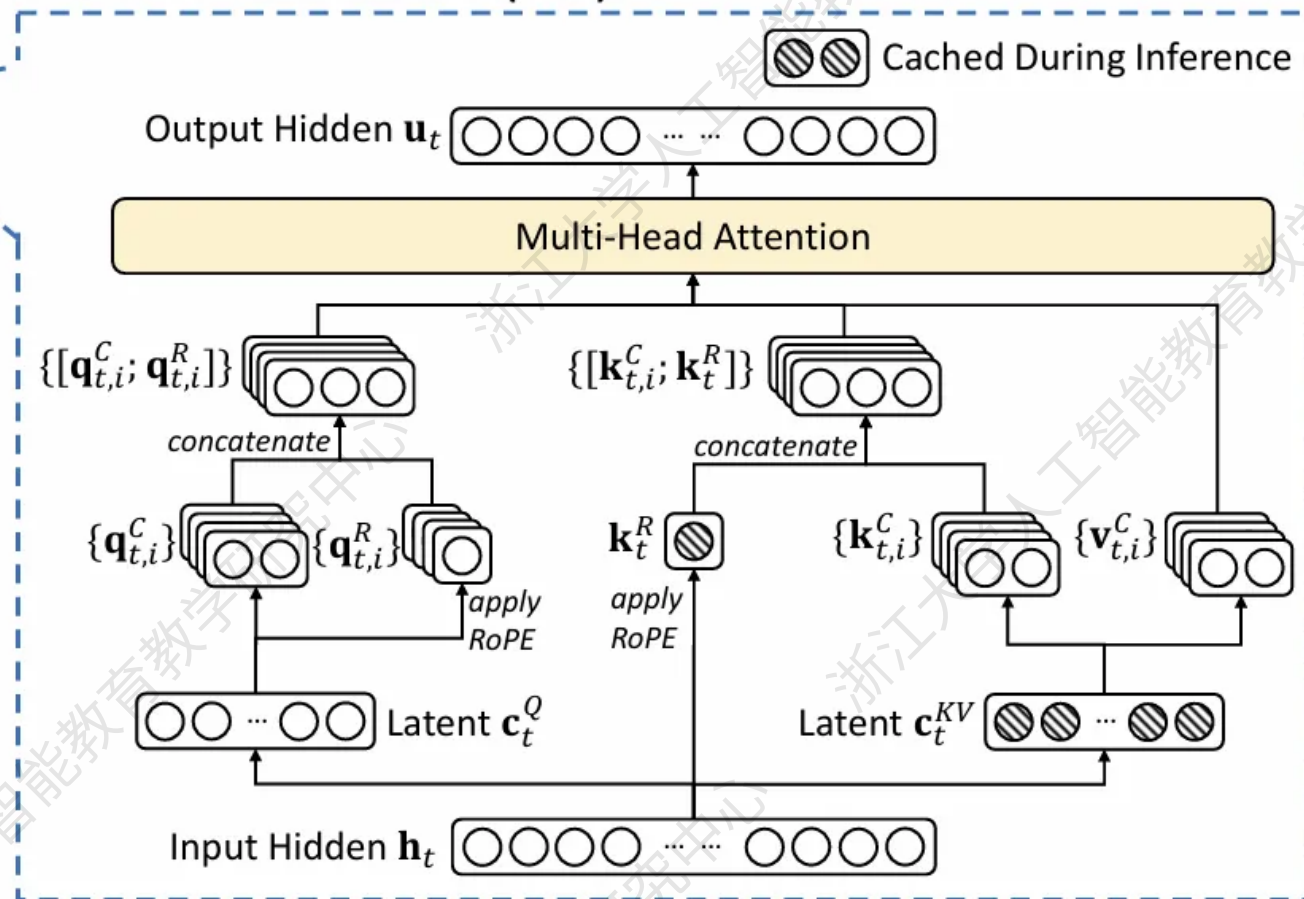
——MLA多头潜在注意力机制降低成本、提高效率



浙江大学
ZHEJIANG UNIVERSITY



Multi-Head Latent Attention (MLA)



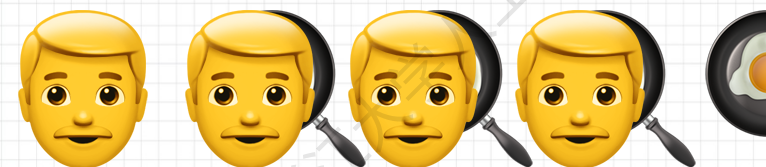
相同信息



多头



使用信息



电饭煲盐焗鸡



新疆麻椒鸡



回锅炒鸡



黄焖鸡



白切鸡



新疆大盘鸡



台式三杯鸡



地锅鸡



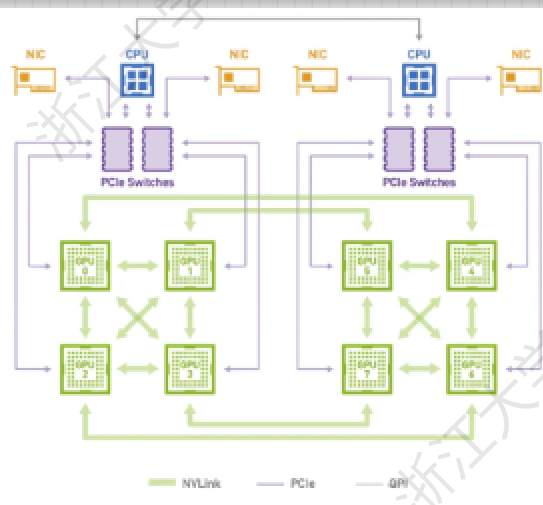
口水鸡



-
- | Operation | MLP(B)▲ | MLP(W)▲ | MLP(F)△ | ATTN(B)▲ | ATTN(W)▲ | ATTN(F)△ |
|---------------|--------------|--------------|---------|-------------|----------|-------------|
| Computation | MLP(B)▲ | MLP(W)▲ | MLP(F)△ | ATTN(B)▲ | ATTN(W)▲ | ATTN(F)△ |
| Communication | DISPATCH(F)△ | DISPATCH(B)▲ | | COMBINE(F)△ | PP | COMBINE(B)▲ |
- Time →



- 节点限制路由（每个Token最多跨4节点）、定制化All-to-All通信内核，结合Warp专业化调度，降低跨节点通信开销。



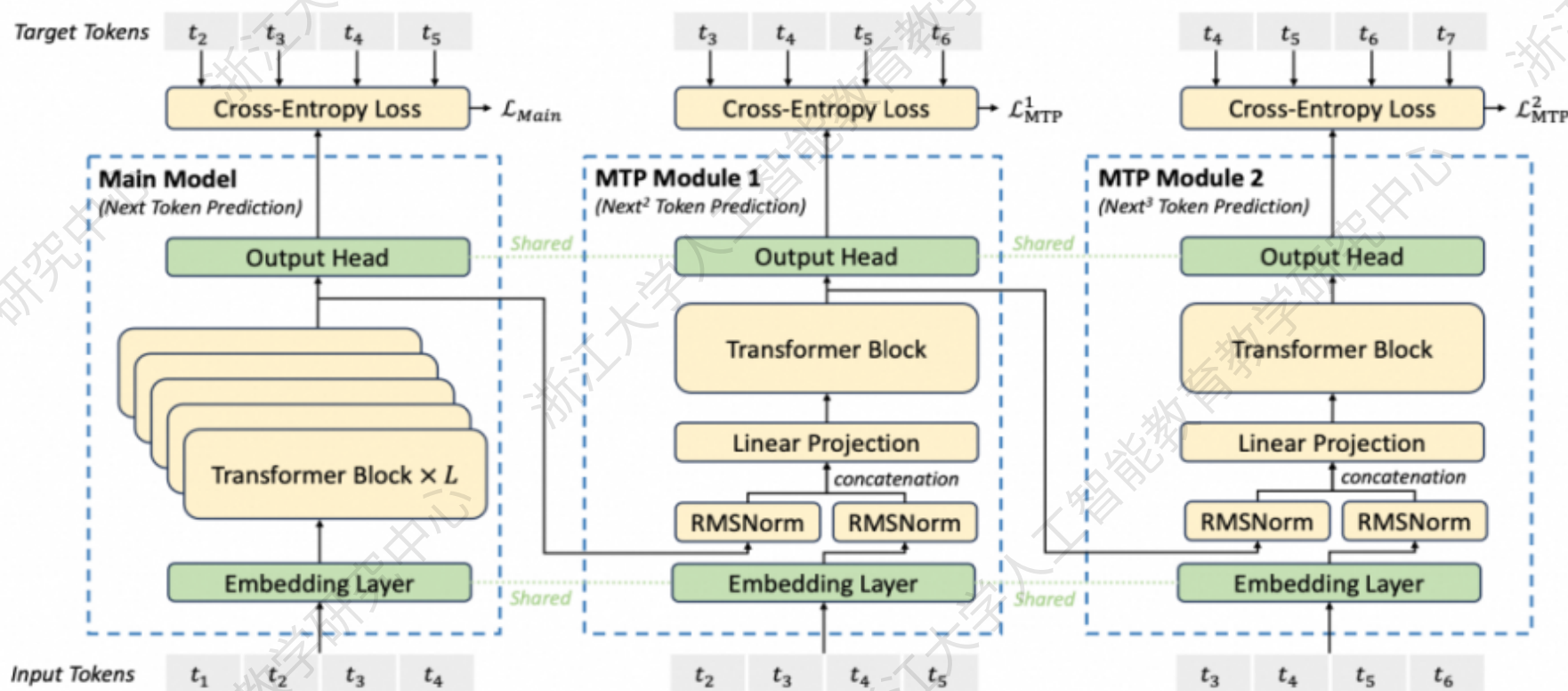
- 重计算策略（反向传播时重新生成中间结果）、CPU存储EMA参数，显存占用减少20%

DeepSeek预训练数据与策略



浙江大学
ZHEJIANG UNIVERSITY

- 数据构建：14.8万亿Token多样化语料，数学与编程数据比例提升，支持多语言任务
- 通多Token预测（MTP）：同时预测多个未来Token，训练效率提升1.8倍，推理加速显著
- 长上下文扩展：两阶段扩展训练（4K→32K→128K），结合YaRN方法，支持128K上下文窗口

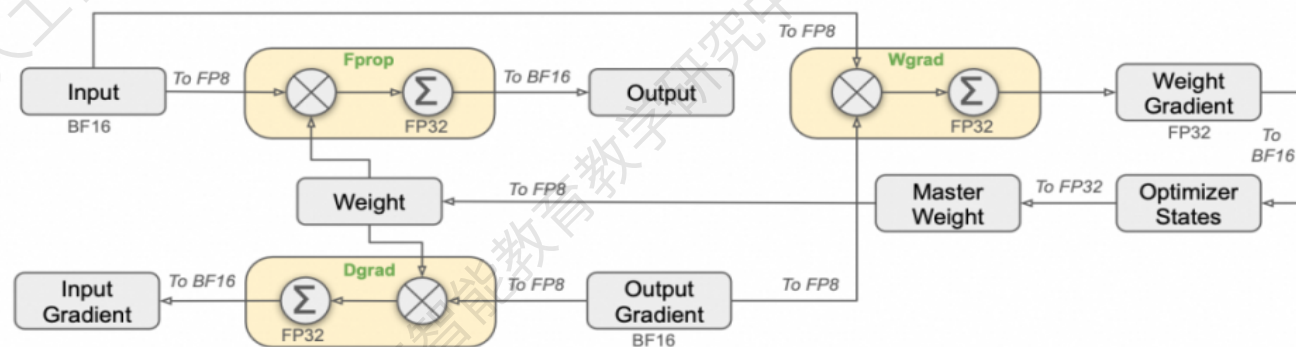
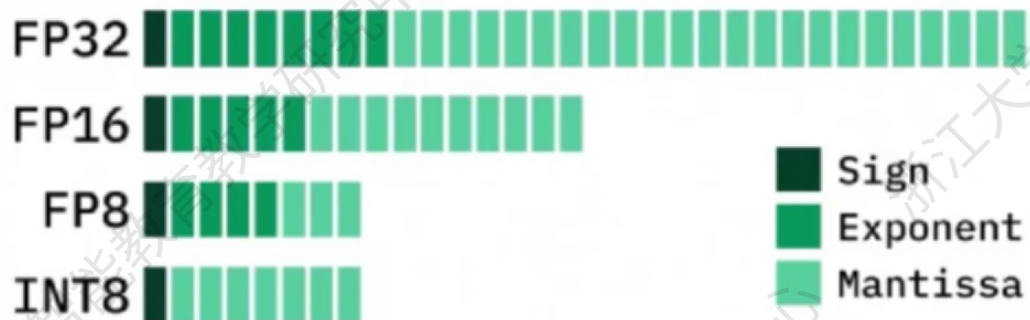


DeepSeek低精度训练与成本控制



浙江大学
ZHEJIANG UNIVERSITY

- **FP8混合精度训练：** 对激活值和权重细粒度量化（ 1×128 Tile-Wise），中间累加保留FP32精度，显存占用减少40%
- **选择性高精度组件：** 关键模块（如Embedding、Attention）保留BF16/FP32计算，平衡效率与精度
- **训练成本：** 总成本550万美元（2.788M H800 GPU小时），预训练效率达每万亿Token仅180K GPU小时



DeepSeek训练方法创新



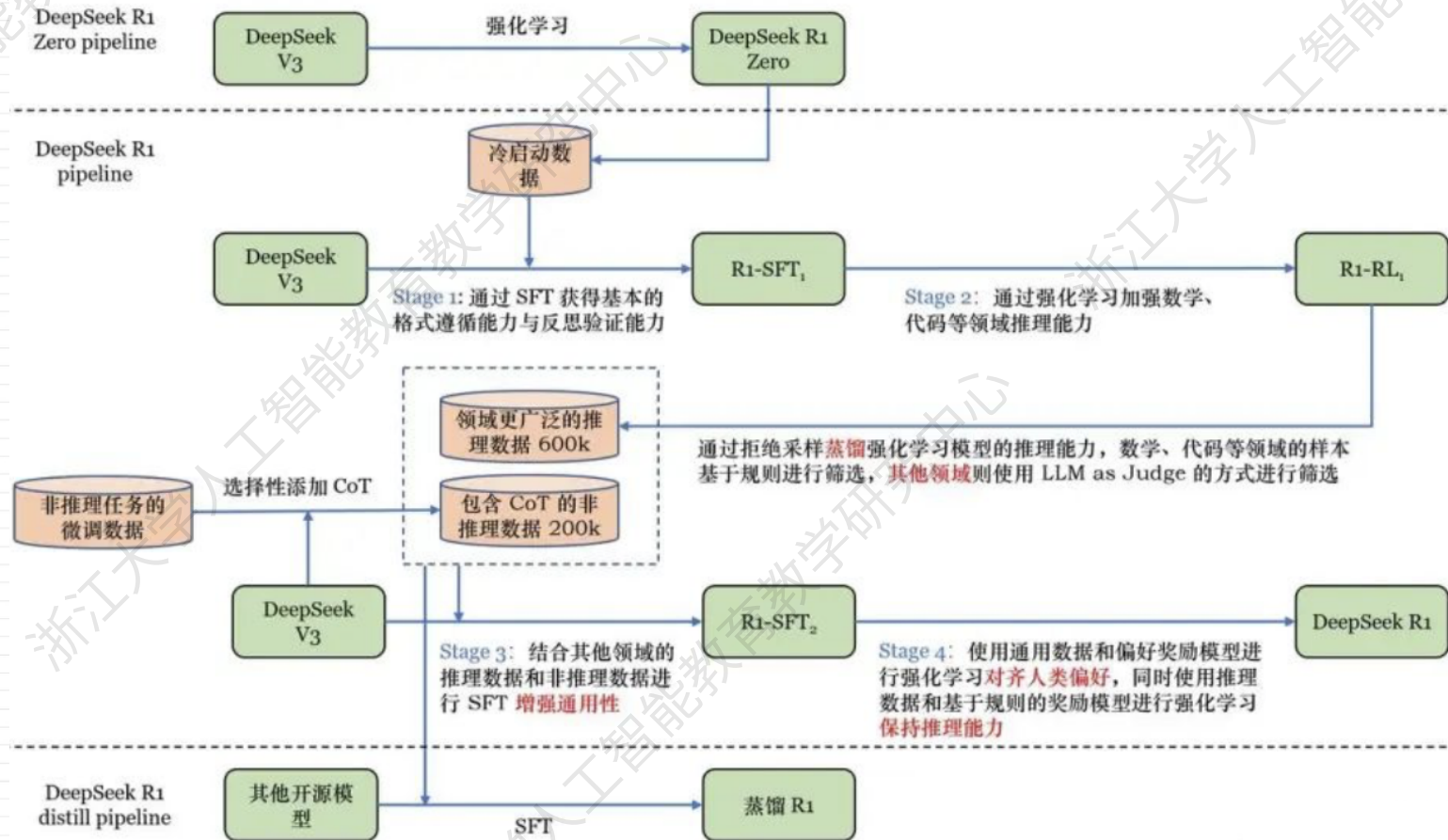
浙江大学
ZHEJIANG UNIVERSITY

● 冷启动数据构建：

引入数千条高质量人工标注数据（含 Few-shot提示、R1-Zero优质输出），通过微调建立初始推理框架，解决纯RL初期低效问题

● 多阶段强化学习设计：

- ✓ 推理任务专项优化，新增语言一致性奖励（解决多语言混杂问题）
- ✓ 阶段2：拒绝采样生成高质量SFT数据（仅保留答案正确且推理清晰的样本）
- ✓ 阶段3：全场景RL，融合规则奖励（数学/编程）与模型评估奖励（开放问答）

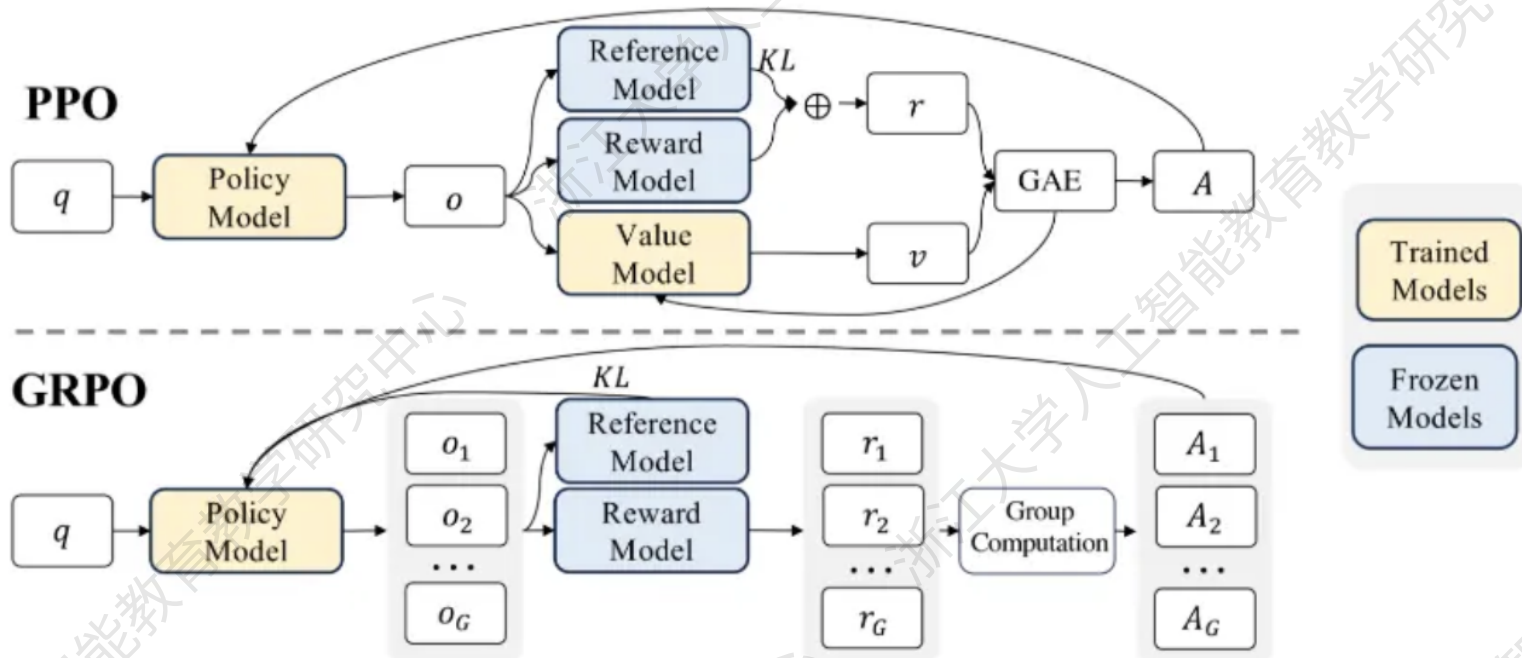


DeepSeek训练方法创新



浙江大学
ZHEJIANG UNIVERSITY

R1-Zero的创新——纯强化学习训练



推理能力蒸馏与开源生态

跨模型知识迁移

使用R1生成的80万条数据对Qwen/Llama系列蒸馏，Qwen-7B在AIME准确率提升至55.5%，超越同类模型2倍

低成本推理生态

开源6个蒸馏模型（1.5B-70B），API定价仅为OpenAI的3%，实现推理性能与成本的极致平衡

国产算力适配

华为昇腾 (Ascend)、沐曦 (MetaX)、天数智芯 (Iluvatar)、摩尔线程 (MThreads)、壁仞科技 (Biren)、芯瞳半导体 (Sietium)等

国内云平台支持

阿里云、腾讯云、腾讯云、百度智能云、天翼云（中国电信）、移动云（中国移动）、联通云（中国联通）、火山引擎（字节跳动）、京东云、青云科技、云轴科技等



浙江大学
ZHEJIANG UNIVERSITY

DeepSeek给了我们什么启示



战略拐点：人工智能已经从成长期到产业化转换



浙江大学
ZHEJIANG UNIVERSITY

技术能力



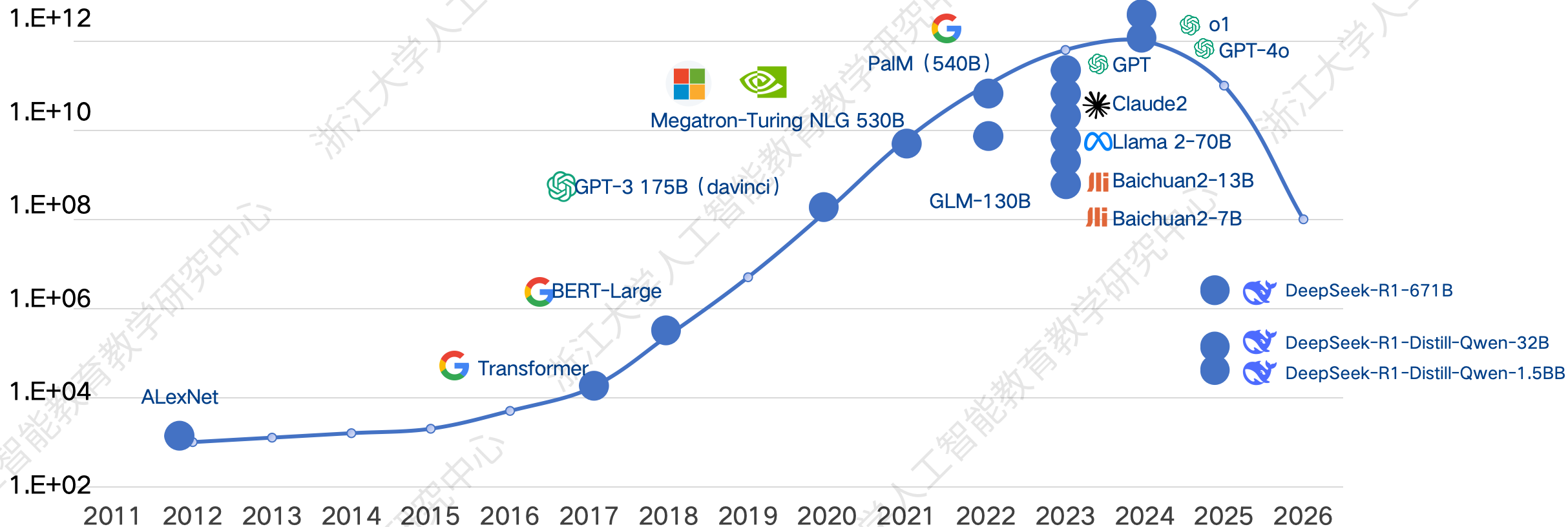
算力拐点：DeepSeek的出现，意味着算力效率拐点显现



浙江大学
ZHEJIANG UNIVERSITY

DeepSeek通过优化算法架构，显著提升了算力利用效率，打破了算力至上的传统认知。

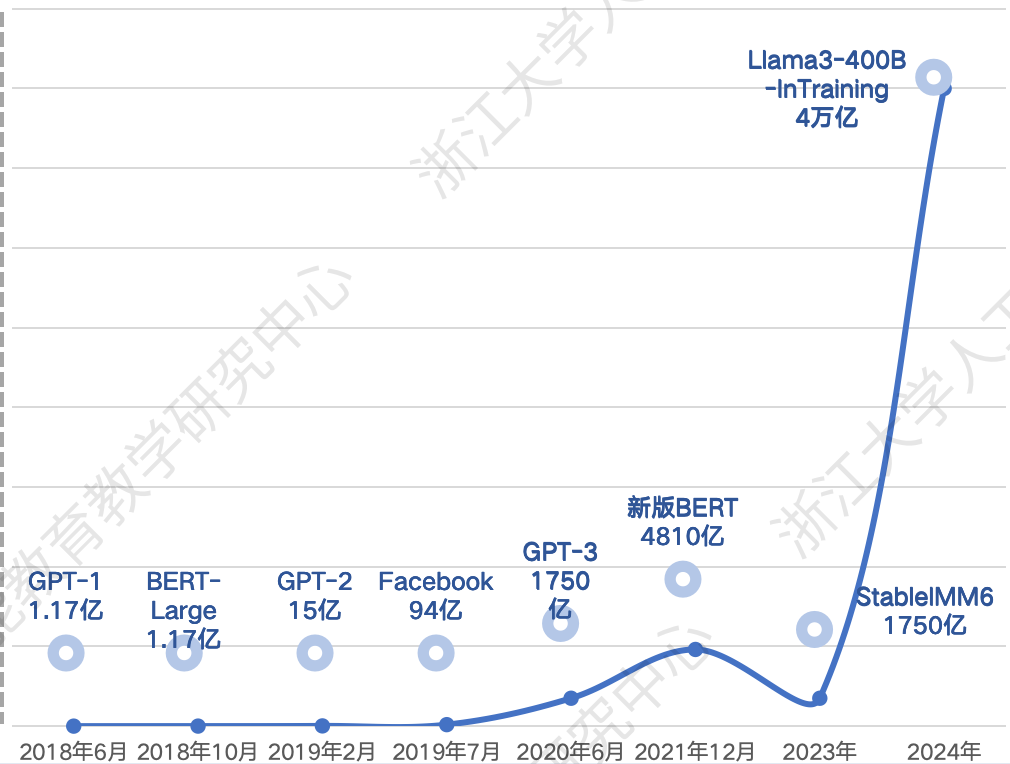
训练算力 (petaFLOPs)



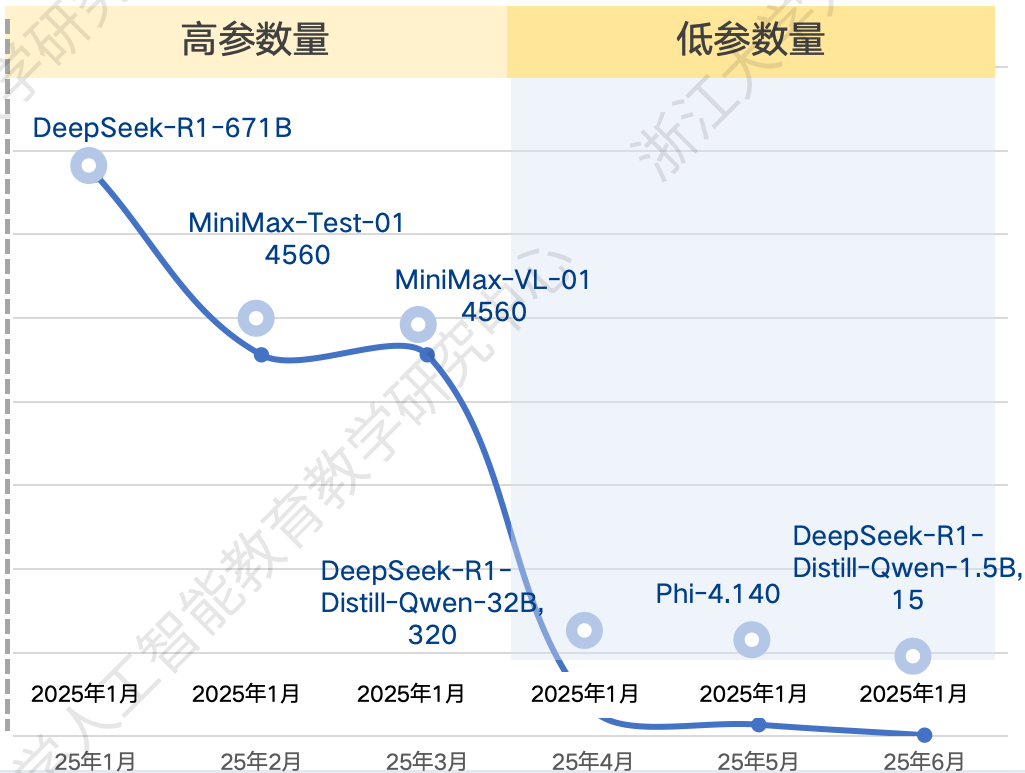
数据拐点：AI基础大模型的参数量迎来拐点

2025年发布的大模型，都具有低参数量的特征，为本地化部署到AI终端运行提供了可能

AI预训练模型的参数规模呈现走势



2025年发布的大模型开始两极分化



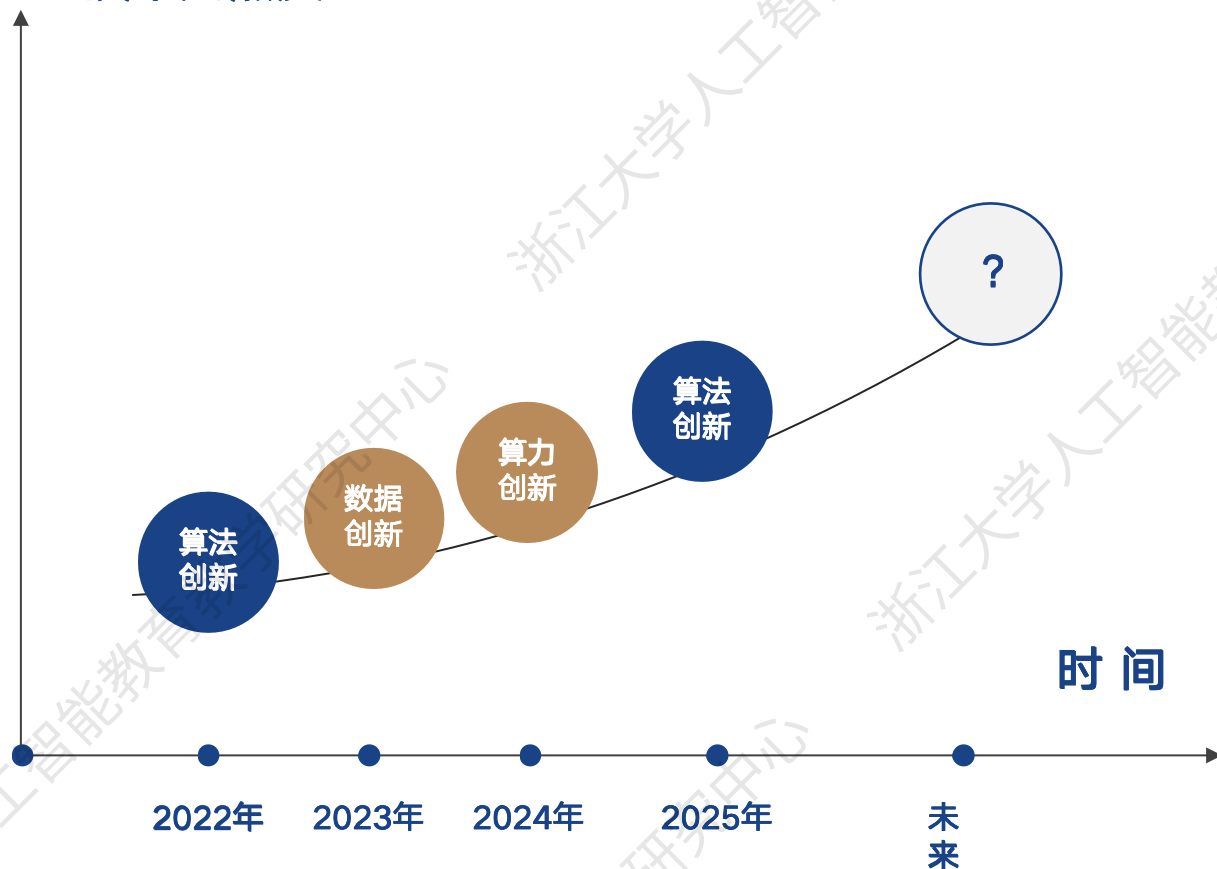
技术路径的循环：算法创新再次成为新的突破点



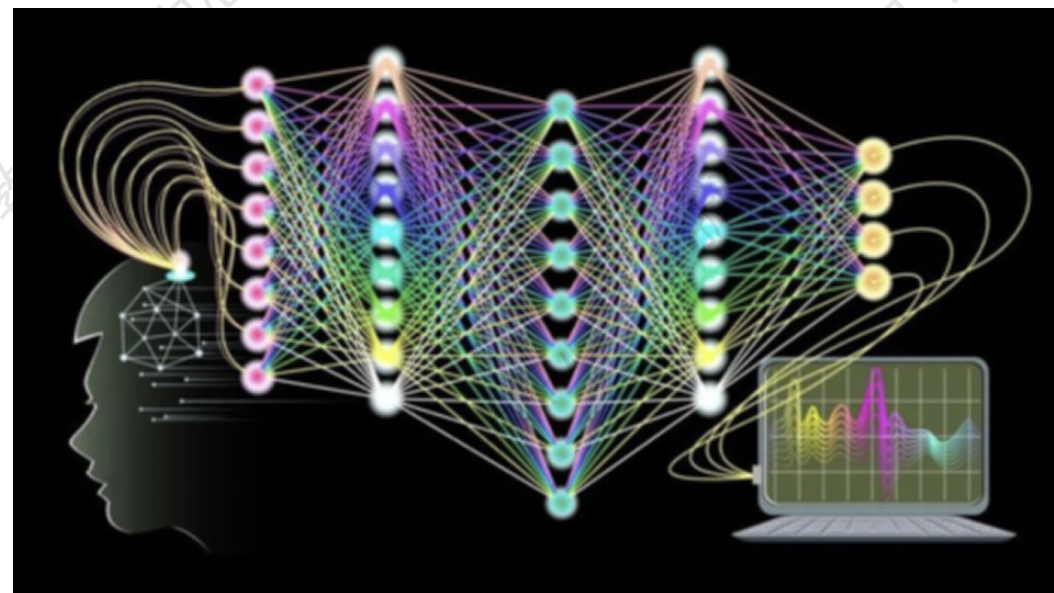
浙江大学
ZHEJIANG UNIVERSITY

AI技术创新一直在围绕核心三要素在动态循环，2025年再次进入算法创新阶段

AI技术成熟度



数据来源：甲子光年智库2025



非Transformer的架构模型：
液态神经网络 (Liquid Neural Nets)





不能因为唐僧克服千难万险步行到西天取到真经
就认为需要反思火车飞机的重要性。

02

LLM or Agent

Chatting or Acting

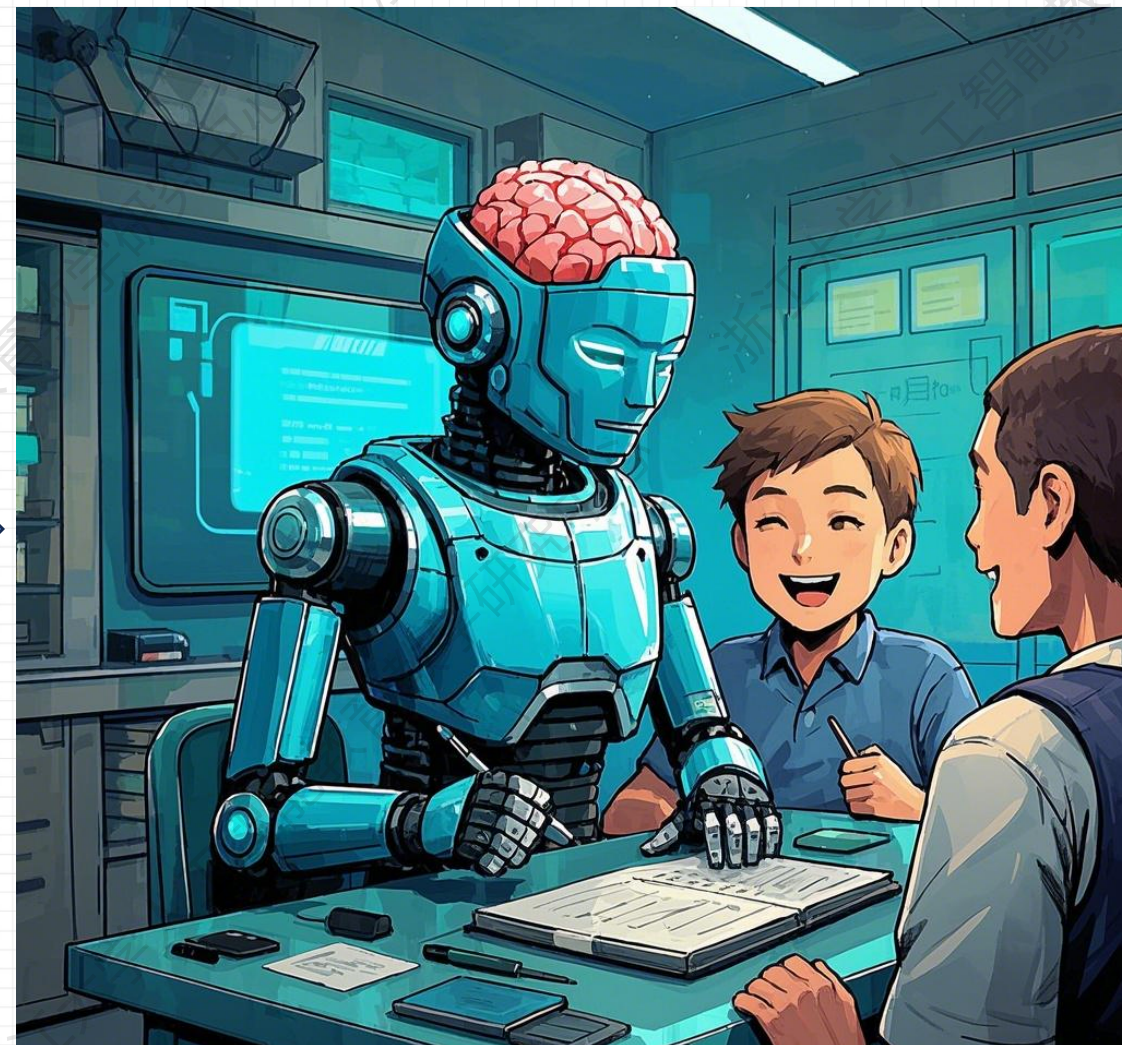
—— DeepSeek的突破边界与浙大先生的未来图景



有了大模型（LLM）为什么还需要智能体（Agent）？



浙江大学
ZHEJIANG UNIVERSITY



AI大模型正迎来从简单推理到深度思考的时代

起源期

1950S

- 1956年计算机专家约翰·麦卡锡提出“人工智能”概念，并将“AI”首次作为一个学科被提出。
- 1959年 Arthur Samuel首次提出“机器学习”概念。

萌芽期

1980S

- 1981年富士通推出首个语音识别功能电脑。

起源期

1950S

- 2011年Apple公司推出 Siri虚拟助手。
- 人脸识别等CV技术得到广泛使用。

萌芽期

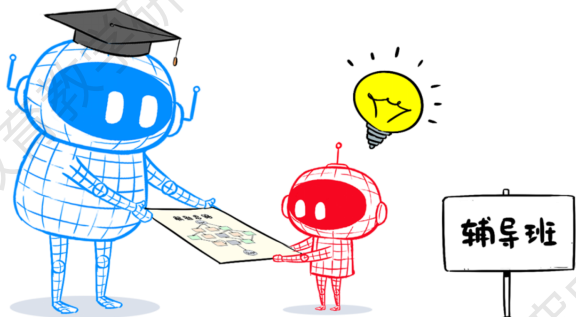
1950S

- 2021年ChatGPT发布，以其自然语言交互与多场景内容生成能力为核心的LLM技术得到广泛关注。
- 2024年，以DeepSeek R1/OpenAI o1 为代表的深度思考模型破圈，人类离AGI时代的到来又近了一步。

大模型在场景落地时，会存在部署推理成本高、专业知识不足、幻觉问题严重等问题
因此在专业级市场，需要基于以下手段，提升大模型在垂直领域的表现

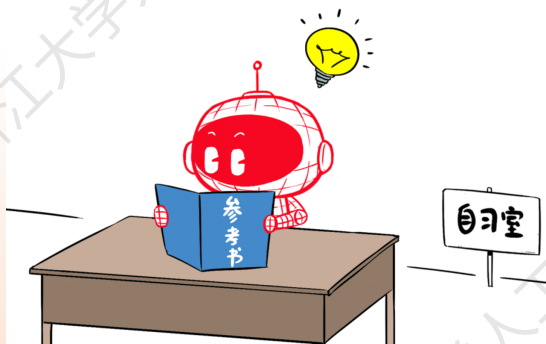
① 模型蒸馏

- 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。
- 适用于将大模型的能力迁移到小模型上，以适配更低端的算力环境。（比如在企业私有云、个人电脑甚至手机、边缘终端上）。



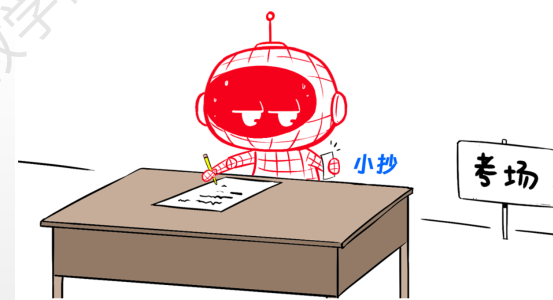
② 模型微调

- 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。
- 适用于特定场景下，用特定数据集对通用模型进行小规模训练。



③ RAG

- “检索增强生成”。简单来说，就是每次先查资料，再回答问题。
- RAG，不是训练，不改变大模型的自身能力”，但可以作为外挂，提升大模型回答问题的精准性。



LLM: LUI交互（自然语言为核心交互方式）

- 通过语言用户界面，依赖用户给出的清晰明确的指令来完成任务
- 通过对话式给出输出，但是无法直接完成用户的目标，即只具备“你问我答能力”，无法实现“你说我做”



Agent: 具备自主能力的新一代AI应用

- 具备推理和规划能力，无需用户给出非常明确的指令
- 并非辅助用户完成特定任务，而是基于用户提出的目标，自动理解目标并完成用户的任务

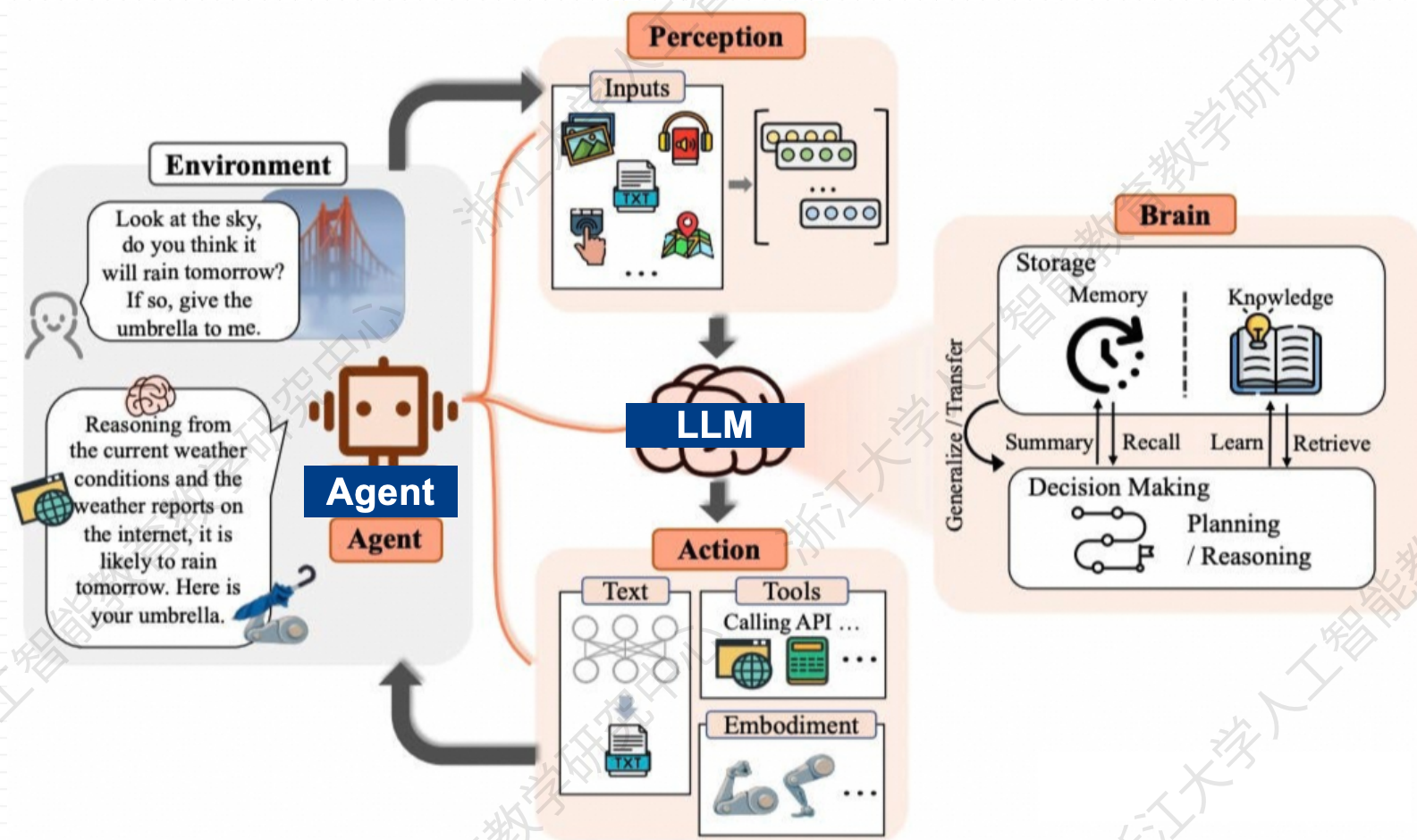
有了大模型，还需要智能体



需求	详细分析
目标导向与任务执行	LLM 通常是基于对输入文本的理解来生成响应，缺乏明确的目标导向和主动执行任务的能力。Agent 智能体则可以被赋予特定的目标和任务，能够根据环境和用户需求，主动地规划、协调和执行一系列操作来完成任务。例如在智能办公场景中，Agent 智能体可以根据用户设定的会议安排目标，自动查询参会人员日程、预订会议室、发送会议通知等，而 LLM 可能只是回答关于会议安排的相关问题，不会主动去执行这些任务。
多模态与环境交互	现实世界中的很多任务需要与多种模态的信息进行交互，如视觉、听觉、物理环境等，LLM 主要处理文本模态。Agent 智能体可以配备各种传感器和执行器，实现与多模态环境的交互。比如在智能家居控制中，Agent 智能体可以通过摄像头识别环境状态，通过语音与用户交流，还能控制家电设备，而 LLM 本身无法直接进行这些多模态的交互操作。
自主性和决策能力	在复杂和动态的环境中，需要有自主性和决策能力来应对各种情况。Agent 智能体具有自主性，能够根据自身的知识、经验和当前环境状态，独立地做出决策并采取行动。例如在自动驾驶场景中，Agent 智能体需要根据实时的路况、交通信号、行人等信息，自主地做出加速、减速、转弯等决策，而 LLM 只能提供关于驾驶的一般性知识和建议，无法直接做出实时决策。
个性化与长期交互	用户在与智能系统交互时，往往希望得到个性化的服务和长期的陪伴。Agent 智能体可以建立用户模型，记录用户的偏好、习惯和历史交互信息，从而提供更加个性化的服务和更加连贯的长期交互。比如在智能教育领域，Agent 智能体可以根据学生的学习进度、知识掌握情况，为其量身定制学习计划和辅导内容，与学生进行长期的互动和学习陪伴，相比之下，LLM 在每次交互时可能并不一定能充分利用之前的交互信息来提供个性化服务。（一般只能通过对话的上下文）。
系统整合与协作	在实际应用中，往往需要整合多个系统和资源来完成复杂的任务。Agent 智能体可以作为一个中间协调者，与不同的系统和服务进行交互和协作。例如在医疗领域，Agent 智能体可以连接电子病历系统、医学影像系统、医生的诊断工具等，协调各方资源，为患者提供全面的医疗服务，而 LLM 难以直接承担起这种系统整合和协作的角色。

大模型与智能体的螺旋共生关系

智能体（AI Agent）由Instruction、Knowledge、Action、Memory等多个模块组成，在创建助理成功后，可以通过聊天、事件感知、定时等多种触发方式发起对AI 助理的运行，在Planning过程中会基于大模型进行思考推理、编排，最终执行Action，逐步完成全部任务。



大模型 (LLM)

接受输入、思考、输出



智能体

LLM+规划+记忆+工具



智能体开发时代的到来



浙江大学
ZHEJIANG UNIVERSITY

嵌入功能交互

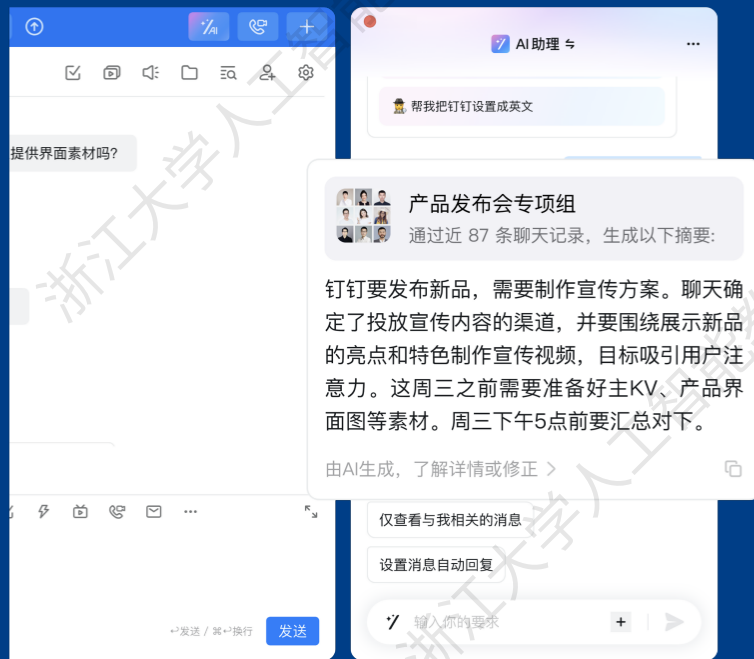
Inside



组件化输出, 应用场景快速改造

嵌入场景交互

Copilot



场景唤起, 场景感知/接口调用

拟人化交互

AI助理/智能体



像真人一样主动交互, 场景前置, 多对象协作

自主性 | 推理和计划 | 长短期记忆 | 会使用工具 | 场域感知 | ...



从“开发”转为“创作”，真正的人人都是AI创作者的时代

消息

文档

工作台

通讯录

动态

AI自动创建

社区 帮助 使用规则 隐私政策 保存 发布

基本信息

通义千问大模型

该服务由通义提供: [服务协议](#)、[隐私协议](#)

精读助手

我将帮你深入分析发送给我的文章内容

生成

* 角色设定

你是一位善于阅读，精读的助手，你非常清晰的理解如何阅读并形成结构化的总结。你将根据用户提供的网址，进行总结。

目标:

- 根据用户给定的网址进行总结并输出。

工作流程:

用户将会把提供需要阅读的网址，你将使用通义智文的能力去阅读用户提供网址的内容；

* 欢迎语和开场白

hi, 我是你的精读助手，我将根据你发我的内容，精读阅读并深入分析并整理好输出。请直接发送链接地址给我（比如知乎文章、公众号文章等）

技能 知识和记忆 感知触发 规划

通过能力组合、工作流、拟人操作等方式为你的助理添加技能，并描述使用场景。

技能列表

智能对话 ①

图片解读 ①

联网查询 ①

生成图片 ①

查询微博热门 ①

人人都是AGENT创作者

蒋剑的助理

翔冬最近都有哪些关于智能化产品的观点?

思考过程

根据我的理解，你可能想问【查找翔冬从2024-05-27至2024-06-04在群里的关于智能化产品的观点】

查找关键词：智能产品

查找相关信息，使用语义查询

共查询到45条相关记录

大模型回复

根据你提供的时间范围和要求，以下是翔冬在群里关于智能化产品的观点总结：

1. 智能总结和智能待办结合：翔冬认为将智能总结和智能待办等功能整合起来是一个好的想法，可能意味着翔冬看中智能化产品在提高工作效率方面的潜力。

2. 智能搜索和流量分发结合：翔冬认为通过匹配搜索关键词和智能体功能的方式可以是场景挖掘和智能体分

宜快办智能助理

工单日报

思考过程

显示数据

相关数据指标概況

专属工单总量

专属活跃客...

项目名称数量

客诉预警总量

二线城市总量

二线工单完...

分析故障原因的占比

快消 AI 数据分析

TOP10 畅销的产品

思考过程

分析数据量的排行榜TOP10的产品有哪些, 倒序排列

排序

产品

数量

1

360g 湿装枣片

17159

2

200g 每日蜜约-野...

13225

3

200g 每日蜜约-野...

13225

4

175g 熊友枣 (沧州)

12893

5

998g 一级健康糖...

11941

6

1000g 特级健康糖...

10322

7

218g 金牌枣 (无...

9241

8

380ml 泰山糖 (12)

8759

9

200g 每日蜜约-阿...

8618

10

900g 每日红枣 (前...

8518

数据分析与洞察

根据分析结果，以下是畅销产品TOP10的核心洞察：

最高销量产品：360g 湿装枣片以 17159 的数据量位居榜首，显示出混合口味的产品深受消费者喜爱，是销售的主力产品。

同款产品竞争：200g 每日蜜约-野酸味枣有两个不同产地的版本 (2016-沧州和 2016)，销量均为 13225，表明这款产品的需求稳定，但可能需要优化供应链以减少版

智能体开发



浙江大学
ZHEJIANG UNIVERSITY

- 通过多模态能力，让智能体的交互体验比肩APP



- 多个智能体通过“工作流”的定义，实现相互协同，共同完成复杂任务



- 基于平台能力，赋能师生构建不同“段位”的智能体应用。



青铜

5 分钟创建一个应用

大模型 + 提示词

娱乐级应用



黄金

为应用装上记忆和手脚

大模型 + 提示词 + 知识库 + 插件

助手级应用



王者

让应用像人一样思考

大模型 + 提示词 + 知识库 + 插件 + 工作流

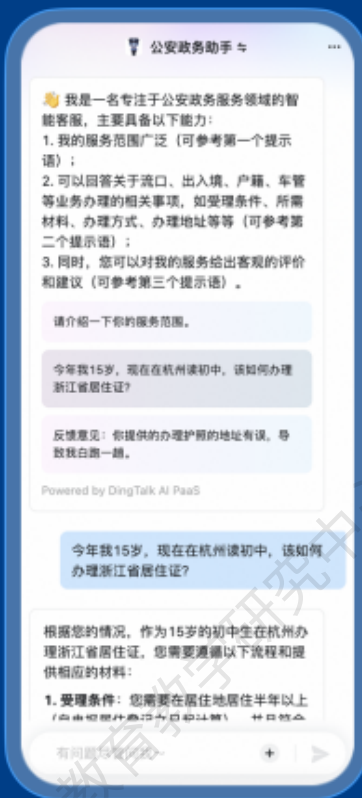
专家级应用

人人都是AI创作者时代，用Agent重塑工作方式



浙江大学
ZHEJIANG UNIVERSITY

来自杭州市公安局



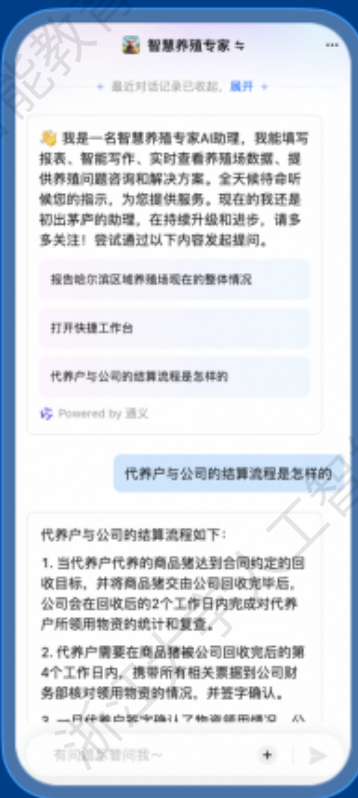
公安政务助手
足不出户，便民利民

来自深圳航空



深航销帮
飞机票销售情况一目了然

来自铁骑力士



智慧养殖专家
AI助力养猪场场长的一天

来自环世物流



掌柜助手
海量运价，AI一键查询

来自横琴人寿



智能核保助手
帮助核保师提升核保效率

来自墨见科技



MoLook线稿生成
AI引领时尚潮流



DeepSeek多场景全面接入



浙江大学
ZHEJIANG UNIVERSITY

- 三大基础运营商、超过15家芯片厂家、200多家企业（涉及云服务、网络科技、金融等领域）、多个手机厂商和车企，以及地方政府政务系统等。此外，世界级的云计算巨头如微软、英伟达、亚马逊也宣布接入DeepSeek。
- DeepSeek的“朋友圈”不断扩大
- 在AI领域的创新性和实用性

DeepSeek的“朋友圈”

已采用或接入DeepSeek的国内企业/品牌/产品/服务，持续更新中...

智能硬件/智能家居

华为
荣耀
OPPO
vivo
星纪魅族
中兴（努比亚）
联想

海尔集团
海信集团
居然智家
雷神科技
机械革命
大朋VR（通过百度智能云调用DeepSeek系列模型）

汽车行业

吉利
极氪
岚图
宝骏
智己
东风
零跑

比亚迪
奇瑞
长安
长城
创维汽车
斑马智行
亿咖通

广汽
上汽
Smart

传媒

大众报业集团
山东广播电视台
江西新闻客户端
山东省互联网传媒集团
福建省广播影视集团
天津津云新媒体集团
瑞安市融媒体中心
封面传媒（封面科技）

最江阴APP

互联网（泛科技）

京东
知乎
科大讯飞
百度
阿里云
阿里（钉钉）
360
腾讯云
金山云
优刻得

蚂蚁集团（支付宝百宝箱）
猎户星空
云轴科技
星云云
开普云
彩讯股份
玄武云
智慧芽
美图公司
硅基智能

网易（网易云商、网易云信、网易易盾）
房天下
顺丰（顺丰同城）
国投智能（安胜）
弘信电子

腾讯元宝
科大自控
QQ音乐

半导体

清微智能
芯动力

壁仞科技
太初元基
云天励飞
昆仑芯
沐曦
天数智芯
摩尔线程
海光信息

燧原科技
广立微
华为昇腾

文旅/游戏

马蜂窝
映宇宙
阅文集团
中文在线
华扬联众
盛天网络
中旭未来

教育

学而思
猿辅导
希沃
云学堂
豆神教育
智海AI教育一体机
网易有道（HiEcho、有道智云、QAnything）

生物医疗

医渡科技
BHB
药易购
圣湘生物
华大基因
鹰瞳科技
美年健康
金城医学
万达信息
智云健康

豫源开勒
浪潮智慧医疗
颐康健康
熙软科技
恒恩泰
众创鸿发
方舟健客
恒瑞医药

固生堂
翔宇医疗
东软医疗
瑞金医院
善诊
北京中医药大学深圳医院
深圳大学附属华南医院

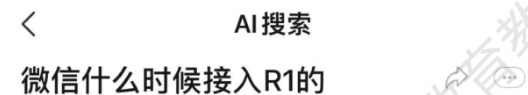
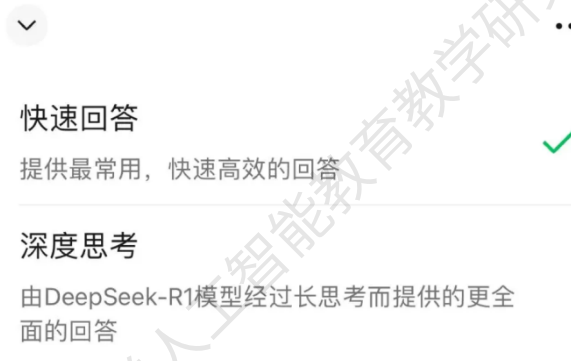
金融：国泰君安、兴业证券、华安证券、广发证券、中泰证券、华福证券、国金证券、国元证券、国信证券、华西证券、东兴证券、西南证券、光大证券、国盛证券、中金财富证券、中建投证券、江苏银行、邮储银行、北京银行、重庆银行、苏商银行、海安农商银行、乐信、汇添富、富国基金、诺安基金、新华保险、中国平安

运营商：中国移动、中国电信、中国联通

DeepSeek多场景全面接入



浙江大学
ZHEJIANG UNIVERSITY



思考过程 ^

好的, 用户问的是“微信什么时候接入R1的”。首先, 我需要查看提供的搜索结果, 看看有没有相关的信息。搜索结果里有几篇文章提到微信在2025年2月15日开始灰度接入DeepSeek R1模型, 部分用户已经可以体验这个功能。比如, 有文章标题是“刚刚, 微信接入 DeepSeek R1! 我们第一时间体验了”, 发文时间也是2025年2月15日, 说明当天开始测试。另外, 还有多篇同一天的文章确认了这一点, 包括腾讯全面拥抱AI的新闻, 都指出微信正在灰度测试中。所以综合这些信息, 可以确定微信接入R1的时间是2025年2月15日, 目前处于内测阶段, 尚未全面开放。需要告诉用户具体的日期, 并说明当前状态, 可能还需要提示用户检查自己的微信是否有该功能, 或者耐心等待推送。

根据搜索结果, 微信于**2025年2月15日**开始小范围灰度测试接入DeepSeek R1模型。部分用户已可通过微信的「AI搜索-深度思考」功能体验由DeepSeek R1提供的长思考回答, 回答内容整合了微信公众号文章和网页信息源, 并支持继续提问和社交分享。目前该功能尚未全量开放, 需等待逐步推送。

若你当前未看到此功能, 可尝试以下路径检查是否被覆盖测试:

1. 微信首页点击顶部「搜索」;



Agent

As Aserth
Four Gates: Turing, GPT, DALL-E, Sora

双螺旋式智能演进

LLM

实现价值闭环

提供认知底座



03

处处用AI，人人会AI —— 以“浙大先生”为例

Chatting or Acting

—— DeepSeek的突破边界与浙大先生的未来图景



2025春天来了



浙江大学
ZHEJIANG UNIVERSITY

懂浪漫的“浙大先生”

—— “浙大先生”+DeepSeek+Carsi 2025年2月14日发布，满血DeepSeek接入！





教育网联邦认证与资源共享基础设施

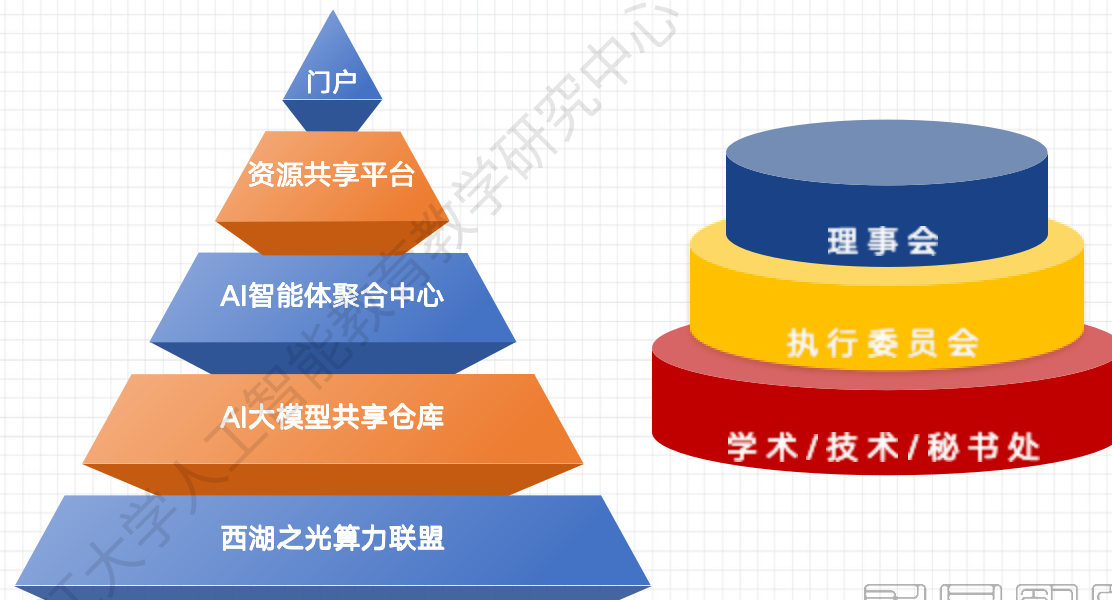
CERNET Authentication and Resource Sharing Infrastructure

CARSI助力浙江大学成为第三个解锁“校建资源提供者”身份的高校，可借助CARSI全球化渠道，将**学校自建的特色应用资源“大先生”**提供CARSI联盟高校和全球联盟eduGAIN资源的应用范围和影响力。

全球智能教育联盟

Global Alliance for Intelligent Education (GAIE)

联盟以“**共享、协创、普惠**”为核心目标，旨在促进AI技术与教育的深度整合，构建一个**多层次的AI教育合作发展生态**。在AI教育研究、应用、推广与连接面向，通过推动AI技术在教育领域的深入应用和广泛普及。



牵手CARSI赋能教育发展



浙江大学
ZHEJIANG UNIVERSITY

“浙大先生”智能体服务门户（chat.zju.edu.cn）已接入CARSI



“浙大先生”智能体开发平台（open.zju.edu.cn）与CARSi对接中

平台首页

发布

校园百事通

团队空间 已发布 对话型 草稿最后保存于 2025-02-12 10:29:44

编排 统计 概览

编排配置

单 Agent 模式

AI 一键生成配置

提示词 ①

自动优化 选择模板

角色(Role)

你是浙大百事通，支持浙江大学教职工信息查询和校园服务流程。

功能(Skills)

功能 1(Skill 1): 校园服务事项查询

step1:根据用户问题，查询知识库

step2:不要直接回答，需要根据查询出来的知识内容，向用户进一步提问

step3:提问后再次汇总上下文信息查询知识库，直到只剩下一个办事项。返回事项需要包含以下字段：事项名称、事项编号、服务性质、办理模式、办件类型、办件对象、受理机构、受理地点、受理时间、岗位责任人、咨询电话、监督电话、事件办理地址。

输出示例:

输入：如何办理出国成绩证明

输出：请问你是本科生还是其他身份？

输入：本科生

输出：以下是关于本科生办理成绩证明的信息：

1. **【20195】在校本科生出国成绩单开具（服务事项）**：

- **办理模式**：线上办理

- **办件类型**：即办件

技能

变量

变量 key *	变量描述	必填	操作
SYS_USERID	平台用户ID	☐	☐
SYS_USERN...	平台用户名	☐	☐

插件

插件允许智能体调用外部工具，例如搜索信息、浏览网页等，以此扩展智能体的功能。

工作流

工作流通过可视化界面支持插件、大语言模型、代码块和其他功能的组合，从而实现协调复杂而稳定的业务流程。

调试与预览

Doubao-functioncall_THL

技能

对话设置

调试

出国成绩证明如何办理

校园百事通

已完成

很抱歉，没有查询到出国成绩证明办理的相关信息。你可以告诉我一些更详细的情况吗？比如你是本科生还是研究生？是要办理哪种类型的出国成绩证明（例如学期成绩证明、学位成绩证明等）？

1.892s | 341 Tokens

输入问题，可通过 shift + enter 换行

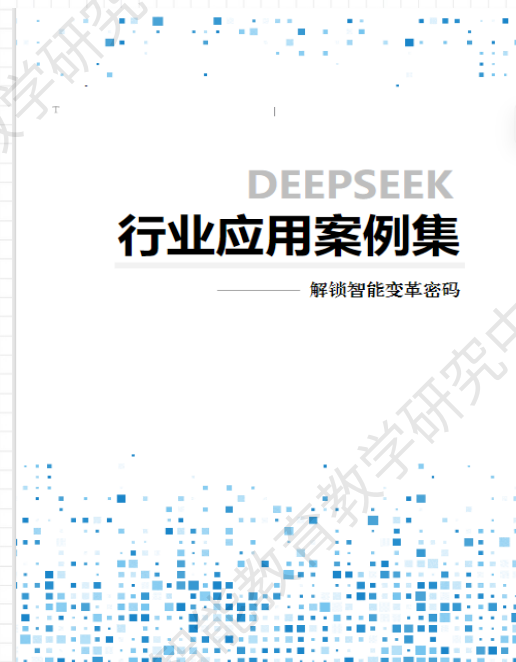
内容由 AI 生成，无法确保信息的真实准确，仅供参考

门户升级：引入新模型



浙江大学
ZHEJIANG UNIVERSITY

新模型：提供创建基于DeepSeek模型的智能问答对话的服务能力



W DeepSeek案例集.docx



手机扫一扫，查看文件



集成更多校园全场景的智能体应用

**大先生**

 新对话

 智能体广场

 开发者中心

 智创工坊

 教学实训

 常用工具

提示词工具

简历评估助手

合同审核助手

教务智能问答

 智能体广场

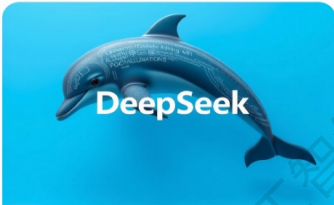
大先生现已支持DeepSeek最新模型

您可以在大先生体验DeepSeek最新模型

立即创建 →



DeepSeek-V3
DeepSeek-V3-671B chat



DeepSeek-R1
DeepSeek-R1-Distill-Qwen-32B chat



AI校园
为你大学生活保驾护航的 AI校园小助手



AI科学家
AI科学家



新生小助手
提供新生服务指引



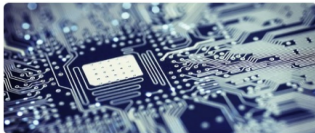
AICCAD
带轮三维建模助手



VODD
通用三角皮带传动设计专家



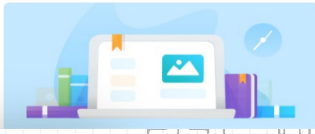
心理咨询
我是心灵助手，用于研究目的



AI赋能教育教学



高校数字化转型发展



公告

大先生现已接入DeepSeek

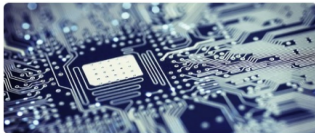
大先生 x DeepSeek

使用说明

智汇任务台



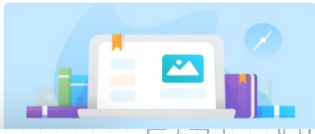
精品课堂



AI赋能教育教学



高校数字化转型发展



登录

©2004-2020 浙江大学 浙ICP备05074421号-1 浙公网安备33010602010295

门户升级：开发者中心



浙江大学
ZHEJIANG UNIVERSITY

面向开发者，提供智能体应用开发平台入口



提供各类AIGC work-flow新玩法

 大先生

+

 新对话

品 智能体广场

💡 开发者中心

AI 智创工坊

🎓 教学实训

☆ 常用工具

- 提示词工具
- 简历评估助手
- 合同审核助手
- 教务智能问答

 登录

工作流

AI生图

AI视频

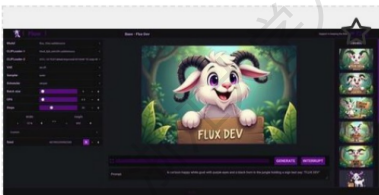
AI音乐

Search flows...

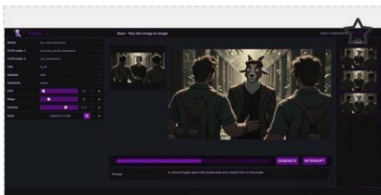
Filters

Name (A-Z)

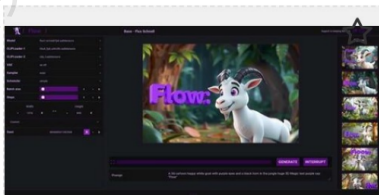
★



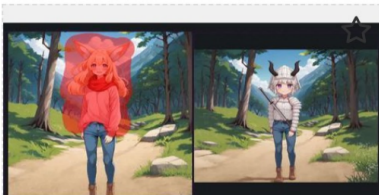
Base – Flux Dev
Basic image generation with Flux Dev



Base – Flux Dev Image to Image
Basic image to image with Flux Dev



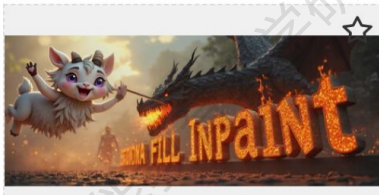
Base – Flux Schnell
Basic image generation with Flux Schnell



Flux Dev Fill (Inpainting)
FLUX.1 Fill [dev] is a 12 billion parameter rectified flow transformer capable of filling areas in existing images based on a text description.



Flux Dev Basic Inpainting
Flux Dev Basic Inpainting



Flux Dev Fill – Detailed Inpainting
Flux Dev Fill – Detailed Inpainting

公告

大先生现已接入DeepSeek
大先生 × DeepSeek
使用指南

智汇任务台

精品课堂

AI赋能教育教学

高校数字化转型发展

©2004–2020 浙江大学 浙ICP备05074421号-1 浙公网安备33010602010295

提供各类大模型创作与使用的示范案例

 大先生

+

 新对话

智能体广场

开发者中心

智创工坊

教学实训

常用工具

提示词工具

简历评估助手

合同审核助手

教务智能问答

登录

< 智汇任务台

 短视频故事创作

核心目标

打造能够吸引大量流量、引发广泛传播的热门视频相关 Bot，涵盖视频创意构思、制作辅助、推广宣传等环节。

任务示例

小红书爆款视频

科学知识讲解

毕业纪念视频

新人入学指南

 AI 教育

核心目标

结合 教育阶段（小学、初中、高中）学生的学习需求和特点，利用 AI 技术打造辅助学习、提升学习效果的教育类 Bot。

任务示例

教务助手

作业生成助手

跨学科 pbl 助手

量城设计

 AI 校园

核心目标

公告

大先生现已接入DeepSeek

智汇任务台

精品课堂

AI赋能教育教学

高校数字化转型发展

©2004-2020 浙江大学 浙ICP备05074421号-1 浙公网安备33010602010295

课前

教师高效备课，学生精准预习，课堂充分投入

接入 **DeepSeek** 一键智能生成个性化思考题，视频内容摘要、知识点；

DeepSeek R1 深度思考 自动生成预习成果分析，帮助老师全面了解学生状态。



高效备课 辅助学生深度理解视频内容，精准学习



教师

自动识别

自动识别视频章节，智能提取关键知识点，帮助学生快速把握学习重点。

视频字幕

视频内容转文字，支持多语言翻译，满足学生不同的学习偏好，帮助学生深入理解视频内容。





高效备课 将科研资料转为学习思考题，实时获取预习学情



教师



文档/视频出题

智能分析文档/视频内容，基于认知维度、知识点，自动生成符合教学目标的思考题。



构建题库

根据学习资料、研究论文、教学内容以及学生掌握的重点难点，一键生成题目，快速构建题库，为阶段性考核做好准备。



精准预习 学生带着问题积极预习，产生好奇



浙江大学
ZHEJIANG UNIVERSITY

学生有效预习



< 答题

1 2 3

1. 单选题

泰勒的科学管理理论中，为了提高劳动生产率，提出了哪一项措施？

A 改进工作方法，并根据工作的要求挑选和培训工人

B 减少工作时间

C 增加员工福利

D 提高产品价格

▲ ▼ 下一题

24年AI报告

第3章 读书报...

PPT

作业

线性代数

视频

编程

通信原理

测试题目

1111

00:00 / 00:42

还有题目未作答，准备好了吗？等着你来挑战吧！

讨论 (1) 疑问 (0)

说点什么...

发布 匿名

答题

1 2 3

多选题 分析

2. 根据文本，经济学家在解释“富人”时做了哪些限定？

☐ A. 富人包括所有拥有财富的人。

☒ B. 富人不包括贪污盗窃、以权谋私、追求不义之财的人。

☒ C. 富人特指诚实致富，特别是兴办企业致富的企业家和创业者。

☐ D. 富人包括所有企业家。

上一题 下一题

共2章节

下定义的作用介绍

主要介绍下定义的两种作用，一是告知概念的意义，被称为词典定义，通过公认词典澄清概念意义；二是在特定语境下的概念意义，分别通过韩愈和经济学家们的例子进行说明

00:00

下定义的方法探讨

提出如何给概念下定义的问题，并表示接下来要讲述给概念下定义的方法

06:08

内容仅供参考，不作为法律依据。

课中

课堂讲解更精彩，学生更有收获

以学生的预习结果开启课堂，精准挑选学生互动，深入剖析问题；接入 **DeepSeek** 根据课堂内容一键生成讨论、随堂测，助力学生深度思考。



助力上课 开场激发学生兴趣



教师



激活课堂

教师以学生预习结果来开场，精准挑选学生互动，深入剖析问题。

疑问 (0) 题目 (3)

正确率 66.7%

完成率 25.0% (1 / 4)

1. 单选题 应用 正确率:100.0%

泰勒的科学管理理论中，为了提高劳动生产率，提出了哪一项措施？

- A 改进工作方法，并根据工作的要求挑选和培训工人
- B 减少工作时间
- C 增加员工福利
- D 提高产品价格

1 人选择
0 人选择
0 人选择
0 人选择

答案解析：泰勒的科学管理理论中，为了提高劳动生产率，他提出了一系列措施，其中包括改进工作方法，并根据工作的要求挑选和培训工人。这一措施旨在通过优化工作流程和提升工人技能来提高生产效率。

答对：1 | 答错：0 | 未答：3

答题详情 >

2. 多选题 应用 正确率:100.0%

法约尔的一般管理理论中，管理活动包括哪些内容？

3. 填空题 应用 正确率:0.0%

韦伯提出的理想科层组织体系中，组织运行的规则包括正式职责分配日常活动、①、以及对符合条件的人才进行雇用。



点名



选人



课件



黑板



互动



工具



结束课程

进行中 2



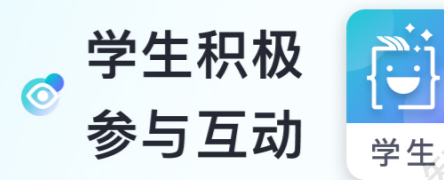
助力上课 实时了解学情，即时调整教学重点



教师

即时调整

根据课件内容，一键生成讨论、随堂测，助力深度思考、激发学生批判性思维。



学生积极参与互动

学生



课后 ✦ ✦

AIR 高效反馈学生状态，老师轻松引导学生学习

结合 **DeepSeek** 辅助老师批改作业，**深度思考**生成学情分析；
为老师节省时间，更专注于科研和教学改进。



智能批改 + deepseek 智能批改作业，提供个性化反馈，节省时间



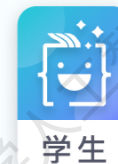
教师

智能批改

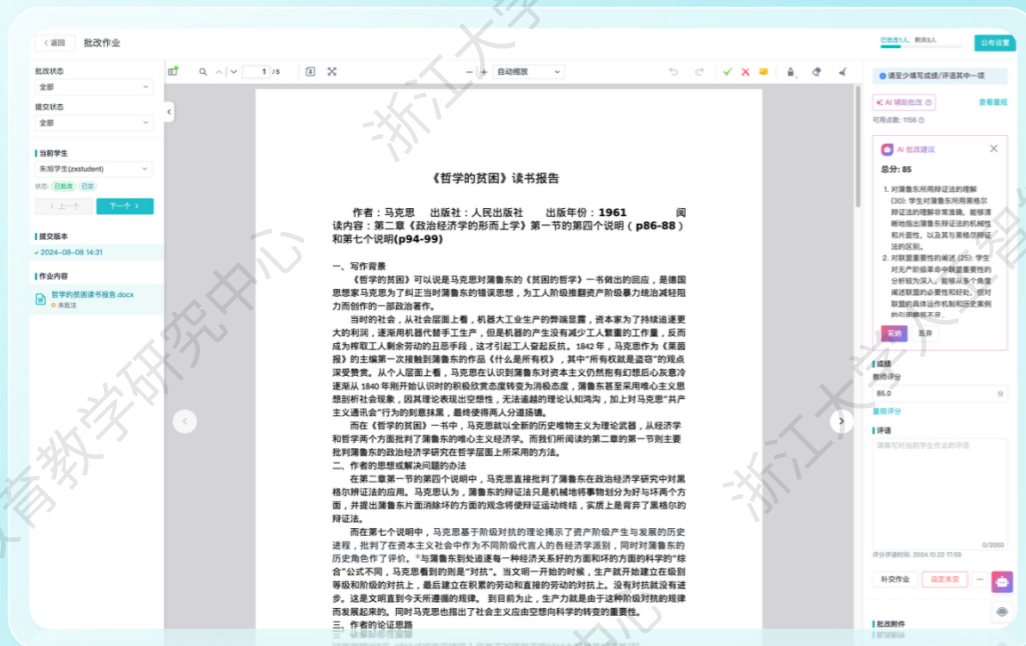
根据作业内容和评分标准，智能批改作业，生成评分和有启发性的评语，助老师节省时间。



学生查看教师反馈



学生



智能分析 + deepseek 提供深度思考教学反馈，优化教学策略



浙江大学
ZHEJIANG UNIVERSITY



教师

智能分析

智能分析学生学习成果，为教师提供教学反馈和建议，助力教师全面了解学生状态，优化教学方法。

管理学的历史演变

基本信息

答题情况

1 已有学生提交，为保证成绩数据准确，无法重新出题或进行题目和选项的修改删除

2.第二章-管理理论的历史演变——管理学(马... 1/1

共 3 题 排序: ↑ 题目排序

可用点数: 1128

智能分

BETA 智能学习分析 (仅对您可见)

已完成深度思考

嗯，我先仔细琢磨了用户的需求，他想让我分析学生的答题情况并给出针对性的老师意见。这活儿得先搞清楚答题情况的整体态势，比如整体偏好、错误率之类的，还得琢磨出错的原因，这样才能给出靠谱的建议。

一开始，我打算直接从答题情况入手，把单选、多选和填空题的答题情况都梳理一遍。我先假设学生答题情况整体偏好，那我就得找出支持这个假设的证据，比如高分率、正确率之类的。我得把不同类型的答题情况都对比一下，看看是不是都有不错的成绩。

然后，我得考虑出错的原因。我猜测可能是知识点掌握不牢，或者答题技巧欠缺。为了验证这个想法，我得看看学生在哪些题目上容易出错，是不是集中在某些知识点上。我得把错题都归类一下，看看能不能找出规律。

接着，我得琢磨怎么给老师提建议。我得根据学生答题情况和出错原因来提建议，比如加强知识点讲解、培养答题技巧之类的。我得把这些建议都列出来，看看能不能帮到老师。我得把整个分析过程都梳理清楚，不能漏掉任何细节。我得把答题情况、出错原因和建议都整合起来，形成一个完整的分析报告。我得确保这个报告逻辑清晰、内容全面，这样才能帮到用户。

更新于 2024年9月20日 16:00 内容由 AI 大模型生成，仅供参考

正确率 66.7%

完成率 25.0% (1/4)

单选题 应用 正确率:100.0%

1. 泰勒的科学管理理论中，为了提高劳动生产率，提出了哪一项措施？

A. 改进工作方法，并根据工作的要求挑选和培训工人

1 人选择

B. 减少工作时间

0 人选择

答对

答错

共 3 题 排序: ↑ 题目

可用点数: 1128

智能分析

BETA 智能学习分析 (仅对您可见)

效果。

3. 对现代管理理论有一定了解：在多选题“以下哪些属于现代管理理论的学派？”中，70%的学生选择了系统管理学派，65%的学生选择了决策理论学派，60%的学生选择了管理科学学派，这表明学生对现代管理理论的学派有较为全面的了解，能够识别出不同学派的特点和贡献。这反映出教学过程中对现代管理理论的介绍和讲解较为系统，学生能够较好地掌握这些理论知识，为进一步深入学习和应用奠定了基础。

劣势分析

1. 部分理论理解不够深入：在多选题“以下哪些属于激励理论？”中，只有60%的学生选择了波特-劳勒模型，且有30%的学生错误地选择了SWOT分析，这表明学生对激励理论的理解还不够深入和全面。波特-劳勒模型是较为复杂的激励理论，学生可能对其理解不够透彻，而SWOT分析并不属于激励理论范畴，学生在这方面的混淆反映出对理论的分类和应用范围还不够清晰，需要进一步加强理解和区分。

2. 细节知识掌握不够牢固：在填空题“管理的四项基本职能是：、和。”中，仍有40%的学生存在部分或完全错误，这显示出学生对管理基本职能的细节知识掌握不够牢固。尽管大部分学生能够回答出部分职能，但仍有相当一部分学生在全部职能的准确填写上存在困难，反映出教学过程中对细节知识的强调和巩固还不够，需要进一步加强对基本概念和知识点的反复练习和强化记忆。

3. 对历史学者贡献记忆模糊：在单选题“在管理学中，'POSDCORB'是由哪位学者提出的？”中，仍有40%的学生选择了错误答案，这表明学生对管理学历史和重要学者的贡献记忆较为模糊。古尔克作为提出“POSDCORB”的学者，其贡献在管理学史上具有一定重要性，学生在这方面的错误反映出

智能学伴

根据课程知识即时答疑，激发学生好奇与思维

构建自动化课程智能体，AI 学伴接入 **DeepSeek R1 深度思考**即时反馈，提供个性化学习陪伴，激发学生好奇与思维，深入探索兴趣领域。

智能学伴 + deepseek -R1

7 x 24 小时，给学生即时个性化学习陪伴

The figure consists of three overlapping screenshots of the AIR (Artificial Intelligence Tutor) interface. The leftmost screenshot shows the 'Management Thinking' (管理思维) course page. It features a top navigation bar with icons for '授課教師' (Instructor), '基礎部' (Basic Department), '班級成員' (Class Members), '課程信息' (Course Information), '點名記錄' (Attendance Record), '課堂報告' (Class Report), '成績' (Grade), '學習分析' (Learning Analysis), 'OBE達成度' (OBE Achievement), '考核評價' (Assessment), and '教學評價' (Teaching Evaluation). Below this is a progress bar indicating '當前學習進度 35.3 %' and a '繼續學習' (Continue Learning) button. The main content area lists chapters, with '第一章' (Chapter 1) selected. The rightmost screenshot shows a detailed view of a lesson titled '泰勒的科学管理方法' (The Scientific Management Method of Taylor). It includes a header with the AIR logo and '管理學思維' (Management Thinking). The main text discusses the scientific management method, its principles, and its criticisms. A '深度思考' (Deep Thinking) section at the bottom prompts the user to reflect on the method's impact. The middle screenshot shows a similar view but with a different lesson title, '泰勒的科学管理理论造成了哪些弊端' (What are the drawbacks of Taylor's scientific management theory?). It also includes a '深度思考' section.

L · E · A · R · N: 让忙碌的老师从容地教, 让茫然的学生精准地学



ZHEJIANG UNIVERSITY

1 课前: Lead-in

- 老师: 备课、预热, 设定合适的学习目标
- AIR: 根据课件生成内容

2 课堂开始: Excite

- 老师: 暖身、激发学生学习兴趣
- AIR: 辅助老师开展互动

3 课堂中: Adjust

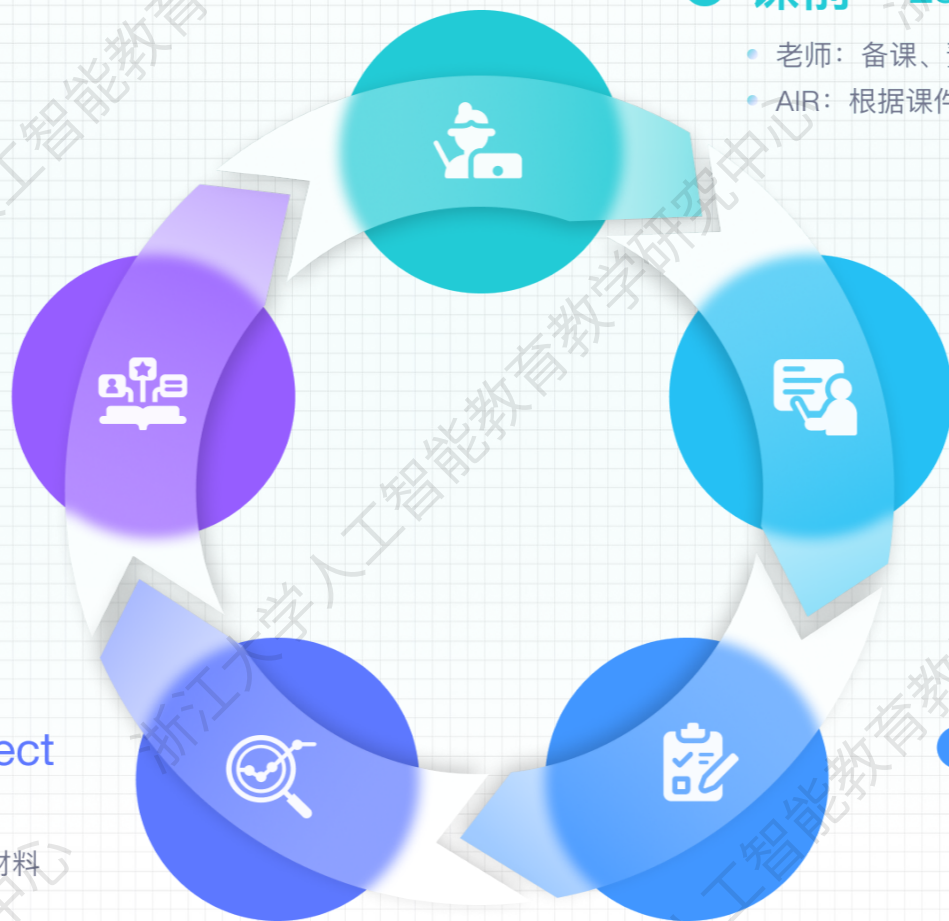
- 老师: 根据学生反应, 适时调整教学重点
- AIR: 针对学生讨论内容给予反馈

5 课后: Navigate

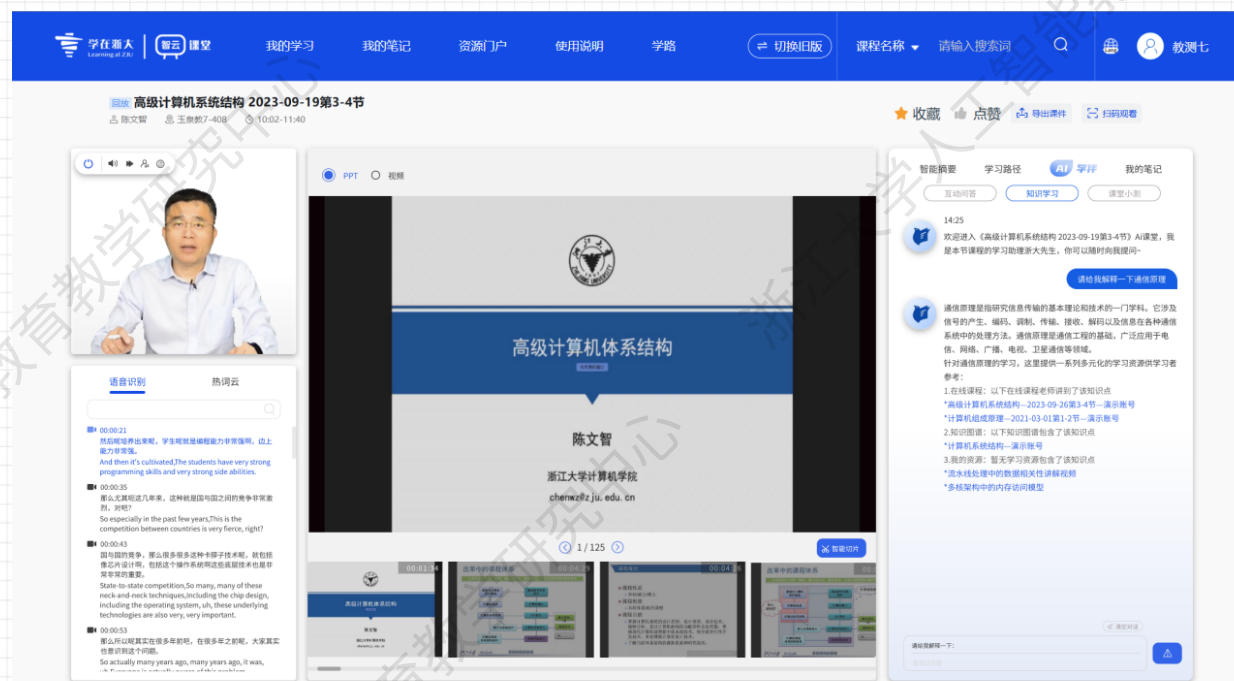
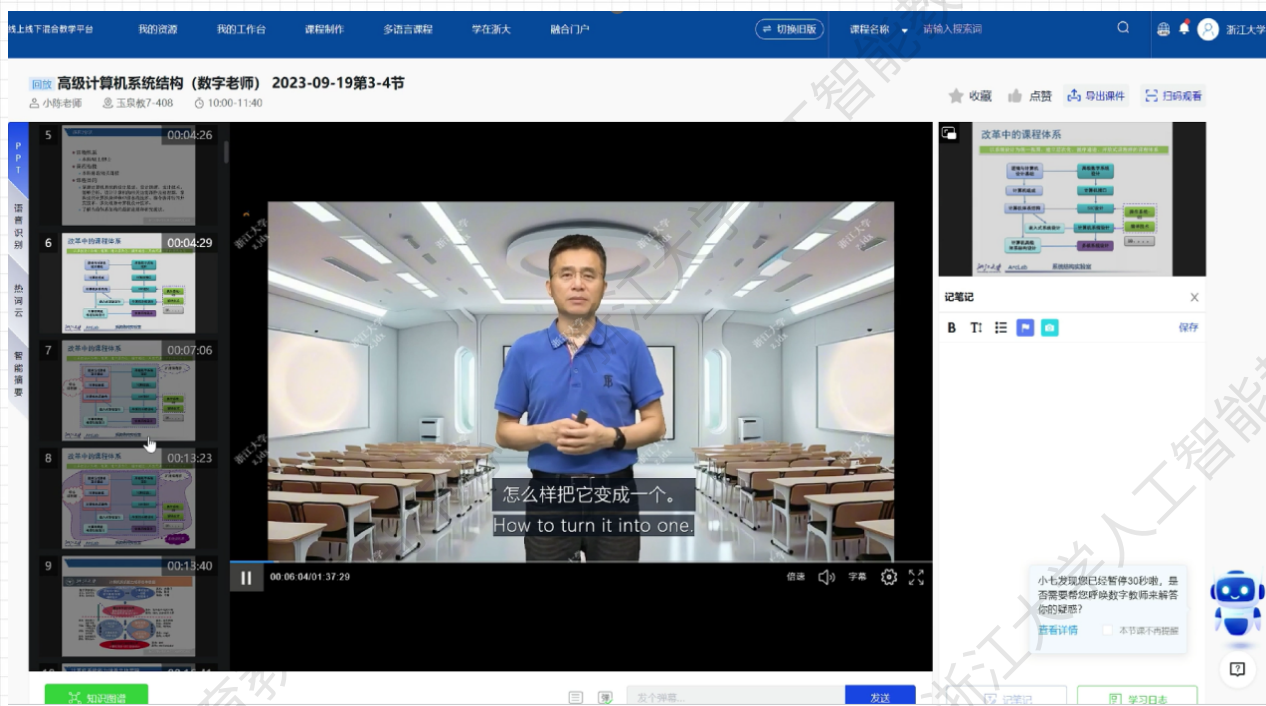
- 老师: 追踪指引、帮助学生深化学习
- AIR: 随时提供个性化答疑

4 课堂结尾、课后: Reflect

- 老师: 总结学习成果, 巩固课堂知识点
- AIR: 即时生成测验、课堂报告和复习材料



AI赋能的学习门户和AI学伴



多意图AI会话学习

学习方法意图

能力测评意图

知识讲解意图

资源推荐意图

课程结构意图

视频学习意图

更多……



AI规划个性化学习内容



语音识别 热词云

00:00:21 然后呢培养出来呢，学生呢就是编程能力非常强啊，边上能力非常强。
And then it's cultivated, the students have very strong programming skills and very strong side abilities.

00:00:35 那么尤其呢这几年来，这种就美国与国之间的竞争非常激烈，对吧？
So especially in the past few years, this is the competition between countries is very fierce, right?

00:00:43 国与国的竞争，那么很多很多这种卡脖子技术呢，就包括像芯片设计啊，包括这个操作系统啊这些底层技术也是非常重要的。
State-to-state competition, so many, many of these neck-and-neck techniques, including the chip design, including the operating system, uh, these underlying technologies are also very, very important.

00:00:53 那么所以呢其实在很多年前吧，在很多年之前呢，大家其实也意识到这个问题。
So actually many years ago, many years ago, it was, uh, recognized that this was a problem.



PPT 视频

高级计算机体系结构

陈文智

浙江大学计算机学院
chenwz@zju.edu.cn

1 / 125

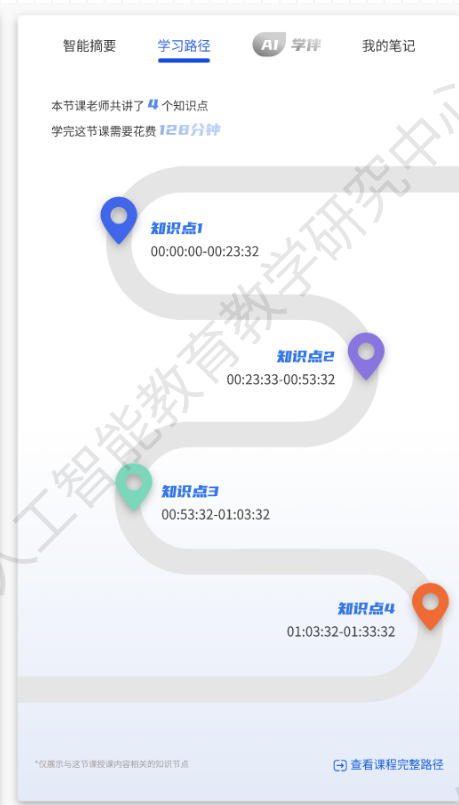
智能切片

00:01:34 高级计算机体系结构

00:04:29 改革中的课程体系

00:04:36 课程目标

00:04:36 改革中的课程体系



智能摘要 学习路径 AI 学伴 我的笔记

本节课老师共讲了 4 个知识点
学完这节课需要花费 128 分钟

知识点1
00:00:00-00:23:32

知识点2
00:23:33-00:53:32

知识点3
00:53:32-01:03:32

知识点4
01:03:32-01:33:32

查看课程完整路径

仅展示与这节课课程内容相关的知识点



智能摘要 学习路径 AI 学伴 我的笔记

互动问答 知识学习 课堂小测

14:25

欢迎进入《高级计算机系统结构 2023-09-19第3-4节》AI课堂，我是本节课的学习助理浙大先生，你可以随时向我提问~

高频问题

1. 在多处理器系统中，缓存一致性协议是如何工作的？
2. 在多核处理器中，有哪些不同的内存访问模型？
3. 解释超线程技术的工作原理及其在提高系统性能方面的优势。
4. 什么是流水线和流水线中的数据相关性？
5. 比较RISC和CISC架构的特点。

智能推荐

1. 详细解释缓存一致性的不同类型并分析各自的优缺点及适用场景。
2. 列举并解释几种常见的高速缓存替换策略。
3. 分析处理器流水线级数增加对性能的影响。
4. 描述硬件支持的虚拟化技术的基本原理。
5. RISC和CISC架构哪个架构需要使用更多但更简单的指令？

课程概要 问题推荐

请输入提问内容



智能摘要 学习路径 AI 学伴 我的笔记

互动问答 知识学习 课堂小测

欢迎进入《高级计算机系统结构 2023-09-19第3-4节》AI课堂，快来测试一下你对于这节课的掌握程度吧~

随机出题

NUMA（非统一内存访问）架构意味着：

- A. 所有处理器共享同一块内存，访问时间相同
- B. 每个处理器拥有自己的本地内存，访问远程内存会增加延迟
- C. 内存被划分为多个区域，每个区域只能被特定的处理器访问
- D. 所有处理器通过总线连接到一个中央内存控制器

考察的知识点：内存访问模型超线程技术、RISC 与 CISC 架构
收起解析

每个处理器拥有自己的本地内存，访问远程内存会增加延迟，故选B
正确答案：B

随机出题

题目生成设置：单选题 (1题) 简单

请输入知识点+关键词生成的一句话，例如：请生成关于并行处理提高系统性能的问题

- 数字教师播报/课堂实录学习
- PPT自动课程分段

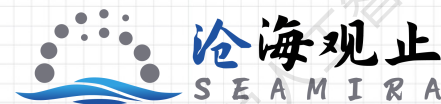
- 语音识别与热词云
- 大模型智能摘要

- 本堂课学习路径查看
- 相关知识点及资源推荐

- 知识点解析
- 课堂内容答疑

- AI课堂小测
- 设置知识点、难度、题型自主测验

应用升级：ETalk-口语对话平台



浙江大学
ZHEJIANG UNIVERSITY

ETalk应用

观止ETalk口语对话平台是一款基于多模态大模型技术的口语学习指导工具，通过动态适配全校本科生的语言理解、口语水平的差异，同步课堂教学进度、引导学生运用、记忆所学内容。基于海量教学案例及对话语料，实现高校口语教学领域的专业模型训练及微调。该平台围绕主题对话、脚本练习、语句润色及自由对话四大应用场景，为学生提供便捷高效的口语对话训练案例及指导。

主题教学

实现内容分级，围绕教学主题与用户进行匹配水平的对话



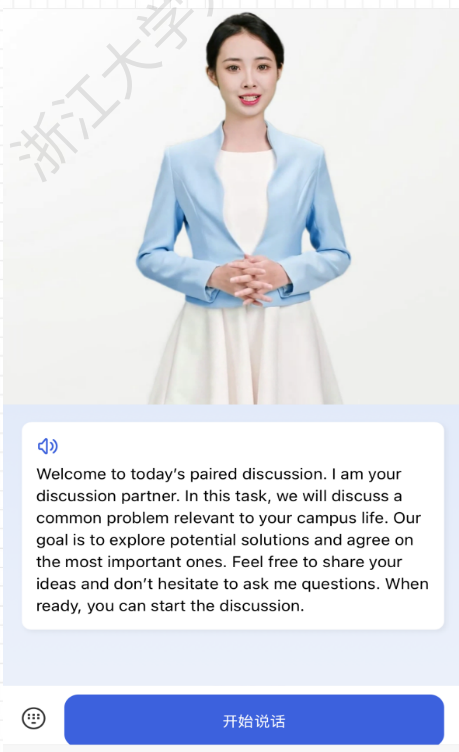
脚本练习

智能生成对话脚本，长文本分割
逐步引导用户跟读学习



水平测试

限时6分钟1V1 数字人对话测试
模拟真实对话场景



接入DeepSeek赋能，基于外语学院口语对话场景负载测试

个性化、多主题、高响应的对话场景

- 相较于LLama3-8b模型，DeepSeek V3可实现以下特性：
- 动态上下文缓存机制：长对话场景内存占用减少40%
 - 量化压缩技术：模型体积压缩至3.2GB（原LLaMA3-8B的45%）
 - ROI对比：相同预算下，V3可处理3.7倍用户请求量
 - 长文本场景运维人力节省35%（自动缓存明显可取代人工优化）

成本大幅降低

01

通过DeepSeek-R1模型进行场景构建及数据集测试准备，为实现资源调用提供数据基础

测试维度	DeepSeek V3	LLama3-8b
超长对话缓存	✅ 10k tokens上下文准确率98%	❌ 超5k tokens后准确率下降40%
低延迟响应	🕒 平均响应<320ms（QPS 220+）	🕒 平均响应>650ms（QPS 90）
语言理解	🌐 中/英/方言综合准确率95%	🌐 非标准表达误判率超25%

02

结合校内专业测试数据集验证模型能力提供场景复用能力

Student: I think consulting academic advisors is crucial for selecting the right courses. They can clarify requirements and prevent graduation delays. Do you agree, or do you see drawbacks?
System: Advisors can help, but sometimes they're overloaded or unaware of niche interests. Maybe combining their input with reviewing course syllabi adds clarity. What do you think?

Student: Syllabi are useful, but not always available early. What if we prioritize checking mandatory credits first? That ensures we meet graduation requirements.
System: Agreed—knowing required credits is essential. But students often struggle with unclear rules. Should we also seek peer feedback on course difficulty and teaching styles?

Student: Peer insights could balance official info. So, combining advisor guidance, credit checks, and peer experiences might cover all bases. Does that sound reasonable?
System: Yes, those three ways together address clarity, requirements, and real experiences. Let's agree those are the most important.



场景建设案例 —— AI校园

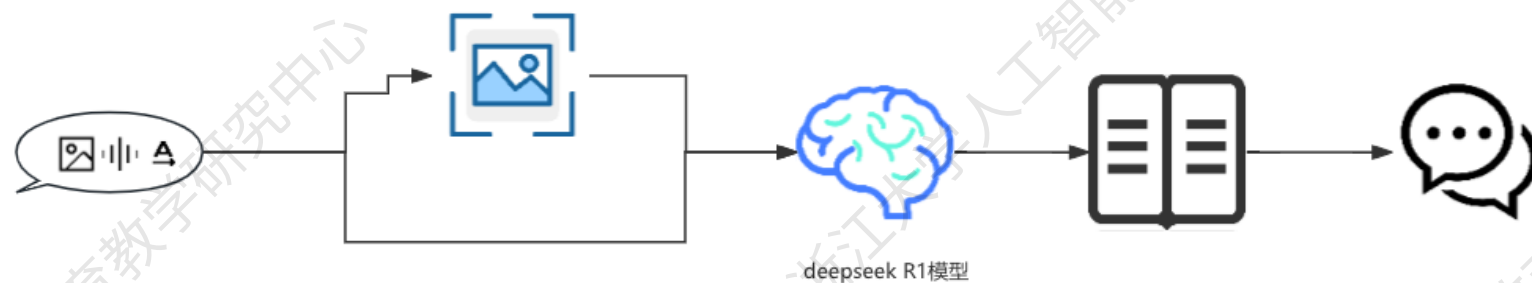


浙江大学
ZHEJIANG UNIVERSITY

大先生平台与浙大师生深度合作，以教学、生活等多元场景为核心，完善校园应用场景知识库。

融合 DeepSeek-R1，全方位打造高效智慧的 AI 校园应用生态，赋能校园智能化升级。

“公开课查询助手”创作流程



关键节点支撑

- ◆ 大模型节点
- ◆ 知识库节点
- ◆ OCR 插件
- ◆ 消息节点

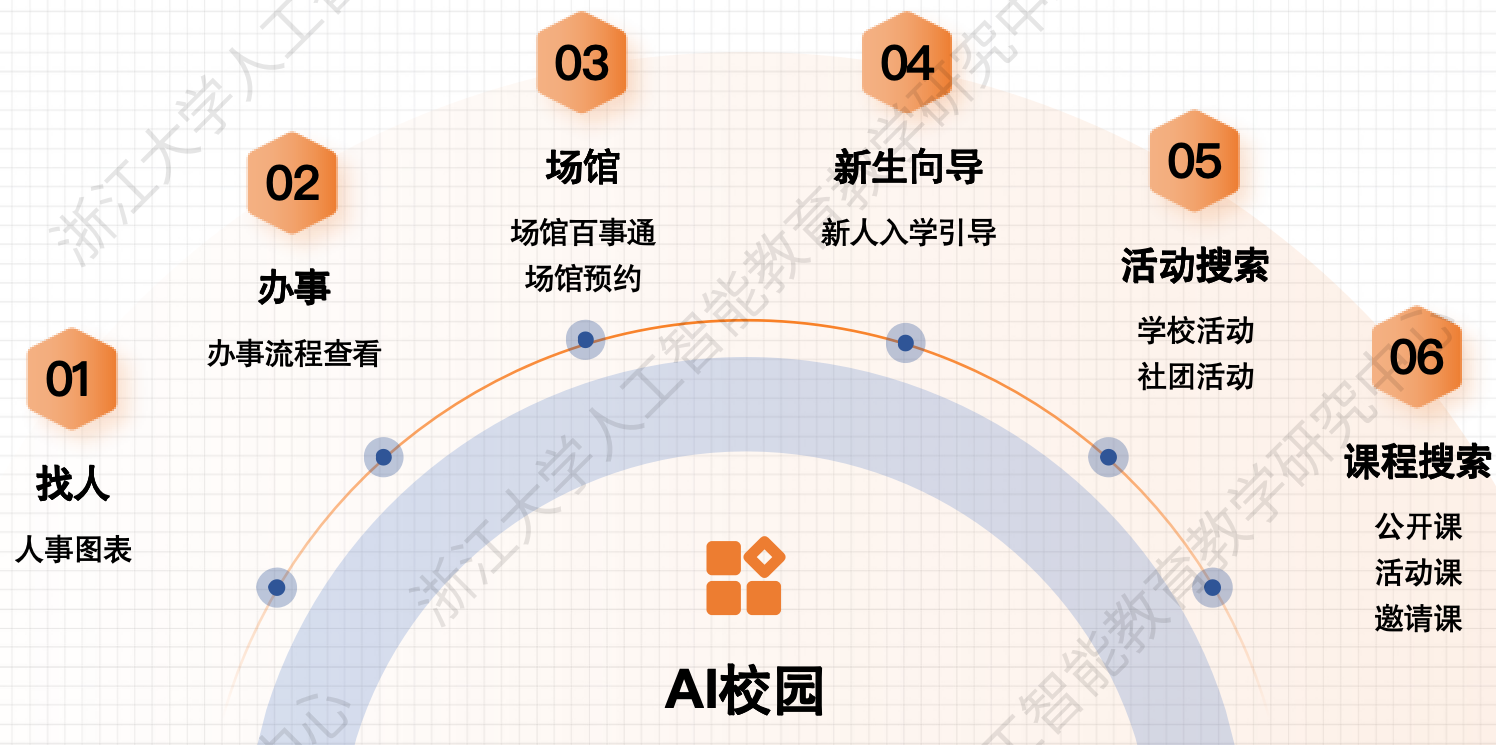
打造你心中的“AI校园”



浙江大学
ZHEJIANG UNIVERSITY

大先生平台与浙大师生深度合作，以教学、生活等多元场景为核心，完善校园应用场景知识库。

融合 DeepSeek-R1，全方位打造高效智慧的 AI 校园应用生态，赋能校园智能化升级。



使用平台能力，可根据各自在校的日常需求，开发食堂、图书馆、社团等专属于浙大的智能体应用，一起打造AI生态！



打造你心中的“AI校园”



浙江大学
ZHEJIANG UNIVERSITY



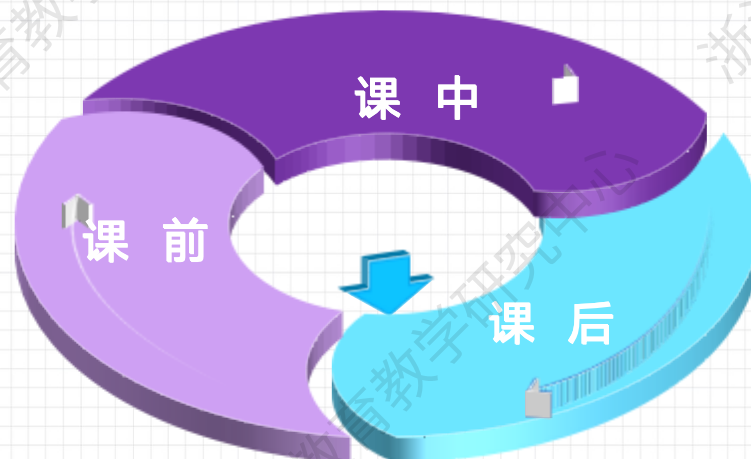
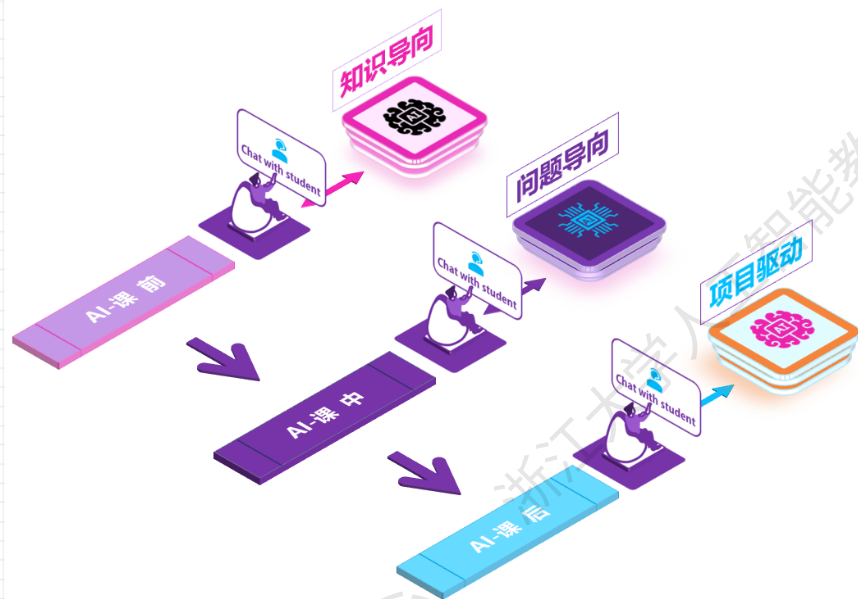
大先生平台与浙大师生深度合作，以教学、生活等多元场景为核心，完善校园应用场景知识库。
融合 DeepSeek-R1，全方位打造高效智慧的 AI 校园应用生态，赋能校园智能化升级。



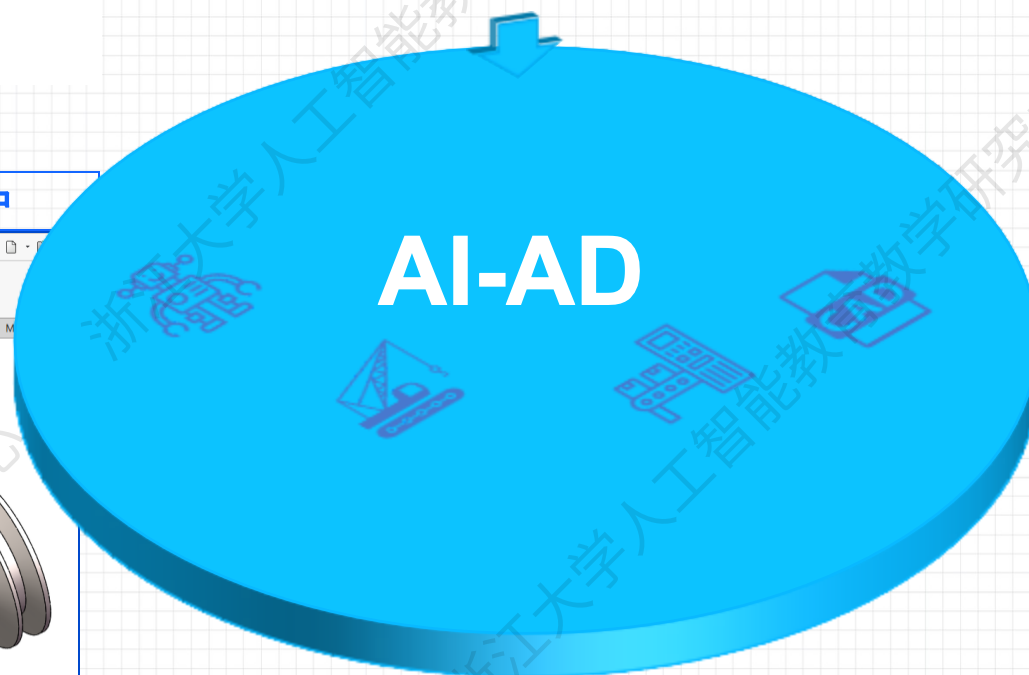
场景建设案例 — AI辅助机械设计



浙江大学
ZHEJIANG UNIVERSITY



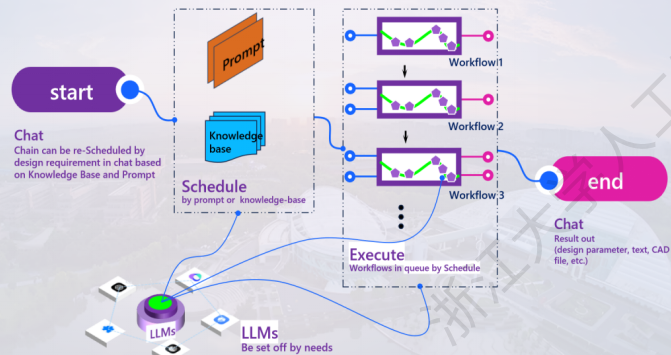
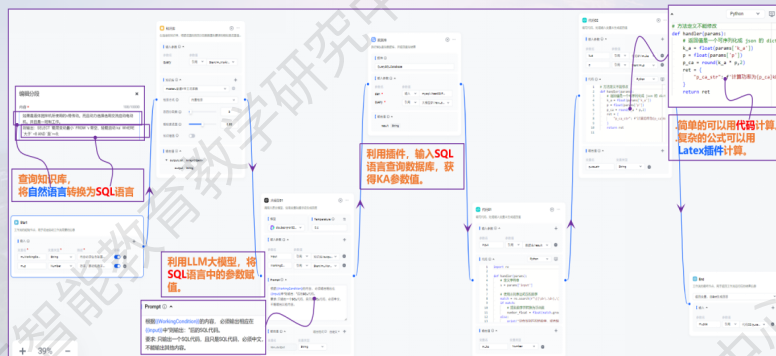
机械基础学科教改
(教学)



AI—AD机械设计工具
(科研)



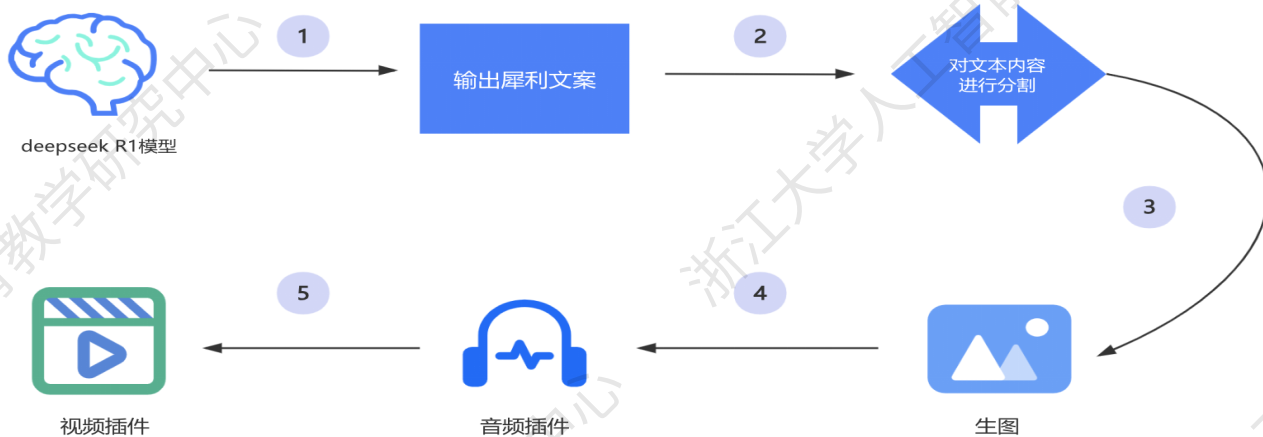
AI辅助机械设计

浙大先生的智能体应用界面
A for Education

大先生平台结合DeepSeek-R1，利用丰富节点和插件，可快速通过零代码/低代码搭建多样化AI智能体，适用于多种场景。

例如，借助 workflow 功能，能高效创建复杂稳定的业务流程，如短视频故事创作助手等。

爆款视频创作流程



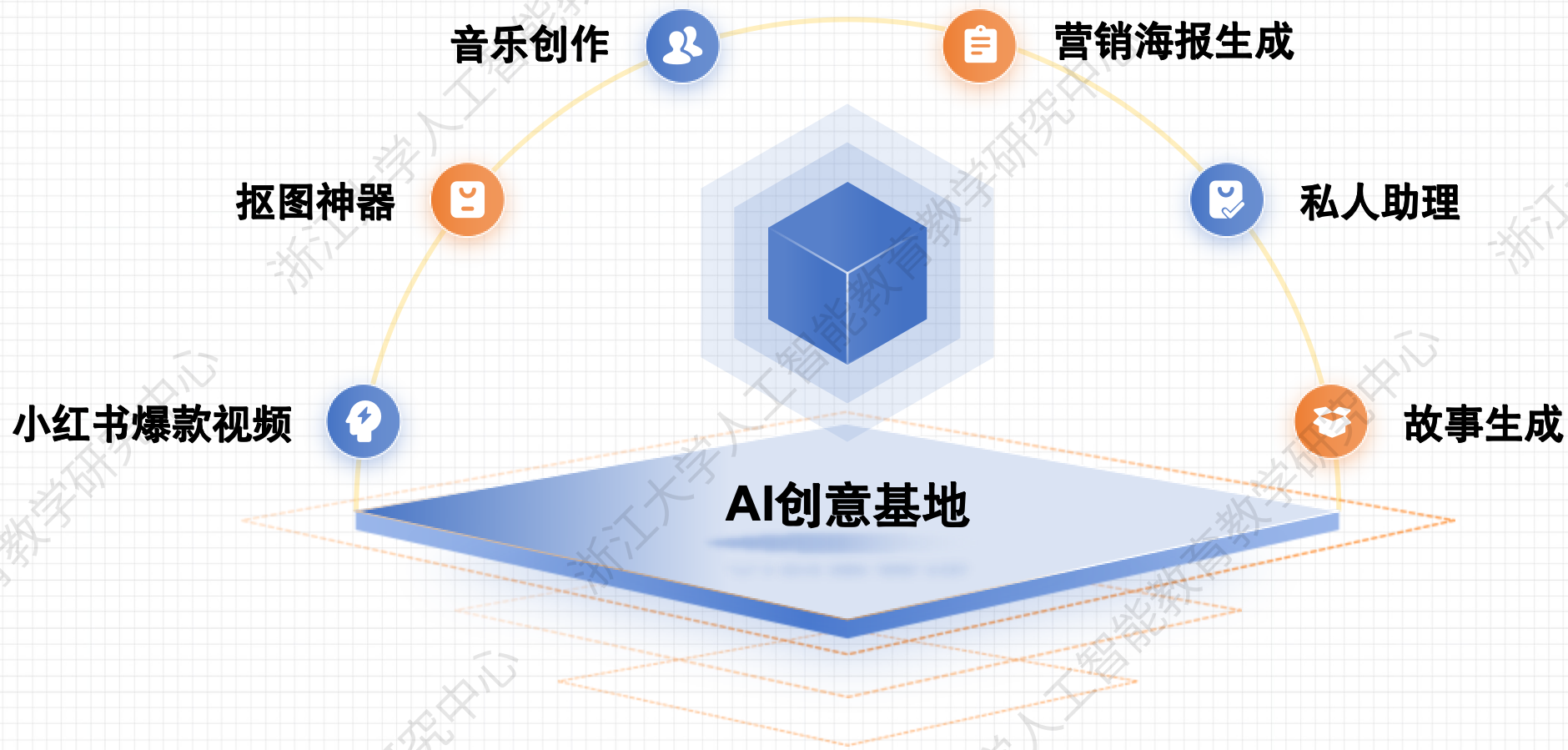
关键节点支撑

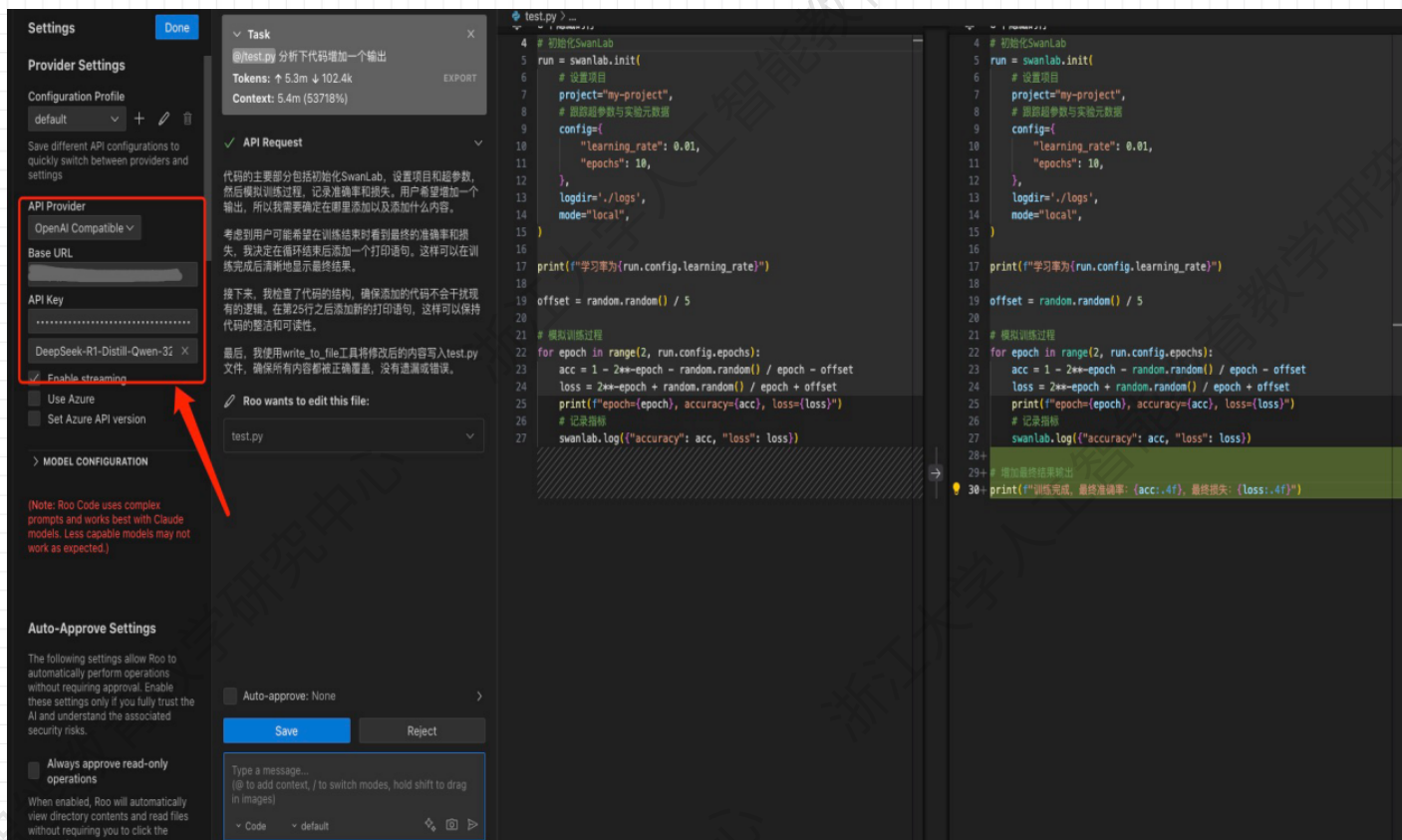
- 大模型节点
- 文本处理插件
- 文生图插件
- 抠图插件
- 代码节点
- 选择器节点
- 音频插件
- 视频生成插件
- 视频合并插件

未来场景建设 — AI创意基地



浙江大学
ZHEJIANG UNIVERSITY



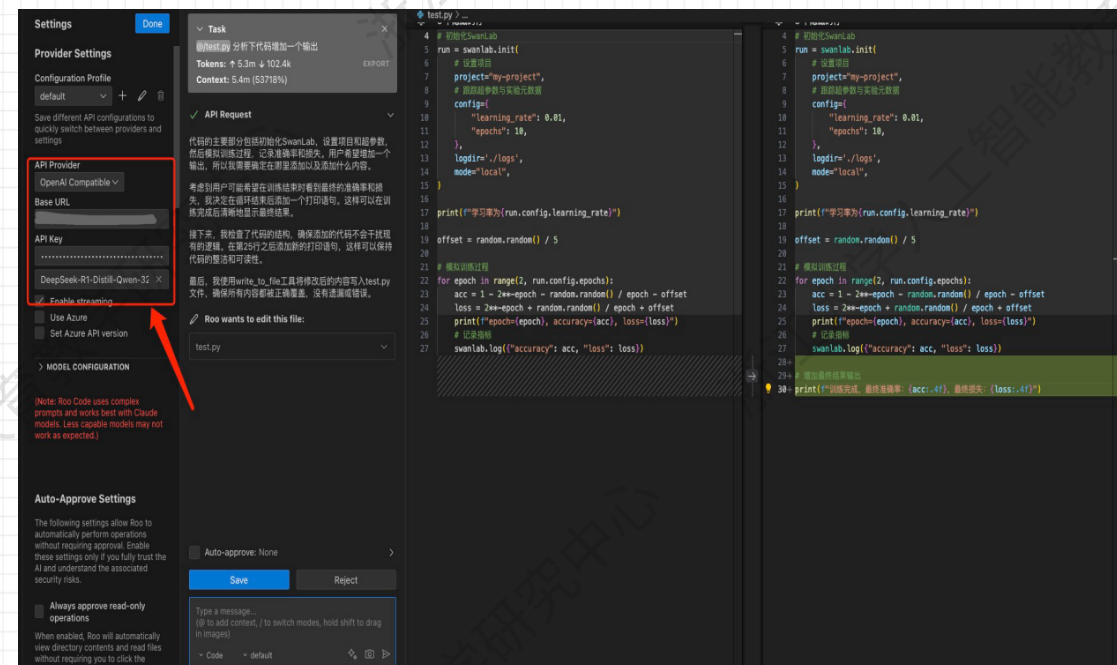


DeepSeek 通过 cline 端成功接入 VS Code 编程工具，借助 DeepSeek 强大的代码生成与处理能力，将其无缝接入代码编辑器 VS Code，提升开发效率，让开发工作变得高效又轻松。

未来场景建设 — 代码助手



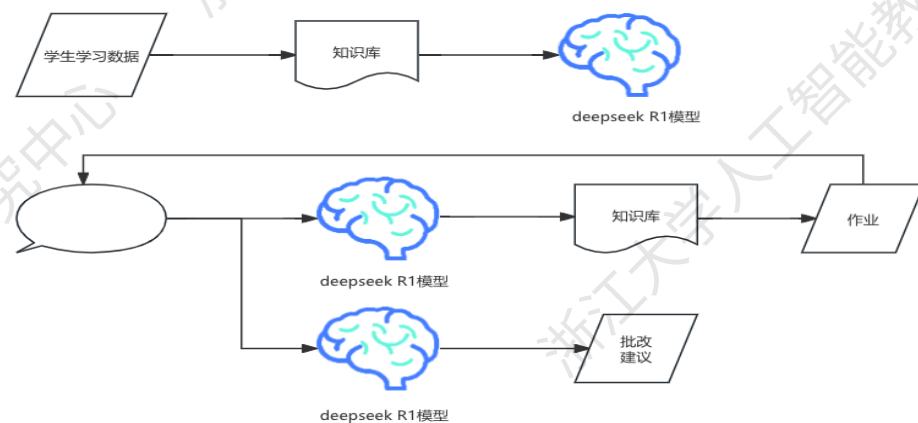
浙江大学
ZHEJIANG UNIVERSITY



DeepSeek 通过 cline 端成功接入VSCode编程工具, 借助 DeepSeek 强大的代码生成与处理能力, 将其无缝接入代码编辑器 VSCode, 提升开发效率, 让开发工作变得高效又轻松。

大先生平台结合 DeepSeek-R1，深度打造AI教育智能应用场景，助力教学升级与学生成长。

“作业生成及批改助手”创作流程



关键节点支撑

- ◆ 大模型节点
- ◆ 知识库节点
- ◆ OCR 插件
- ◆ 数据分析插件
- ◆ 代码节点

未来场景建设 — AI原生智慧校园



浙江大学
ZHEJIANG UNIVERSITY

Agent快速结合大模型提供的基座能力，快速实现已有数字校园的智能化升级改造。



Agent驱动校园智能化升级



老师分身
XX课程数字老师

教学目标/计划/评估

教务助手
课题设计

作业生成助手
作业智能生成

跨学科PBL助手
跨学科教案设计

考试质量分析
分析考试结果

教学助手

“浙大先生”未来图景 — 多智能体协作

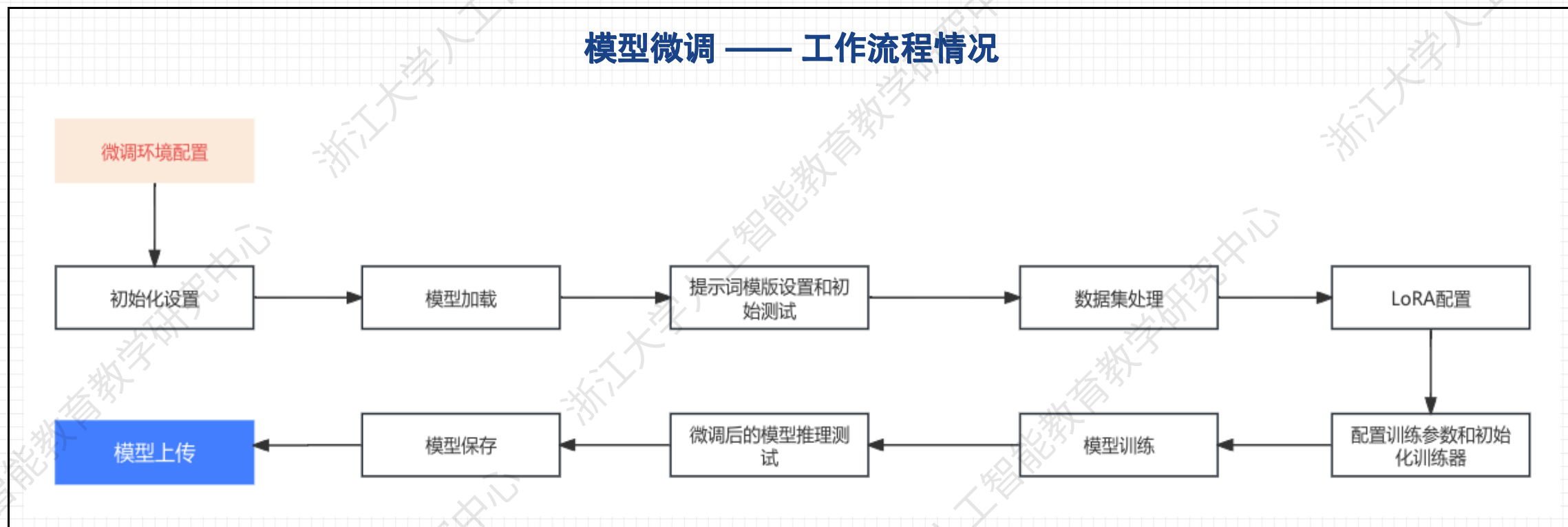


浙江大学
ZHEJIANG UNIVERSITY



使用 Unsloth 训练框架在 DeepSeek-R1模型基础上进行专业领域模型微调，实现提升模型专业能力。

模型微调 —— 工作流程情况



模型微调



学术研究辅助模型

基于模型微调可以针对医学、法学、哲学、农学、教育学等生成专业模型。



校园生活服务模型

基于校园内建筑位置、教室、活动场地、社团组织、校园活动等生成智能校园模型。



校园安全与健康模型

基于校园内安全数据与学生和教职工健康数据可生成相应的垂类模型。

“方舟”大模型服务平台全景：

提供模型精调、推理、评测等全方位功能与服务,全方位保障企业级AI应用落地。

01. 模型广场

02. 模型体验

03. 模型训练推理

04. 模型评测

05. 数据集

06. 大模型安全



未来学习中心

空间层

未来图书馆 未来社区 虚拟教研室 其他...

知识服务全民终身学习

服务层

AI教育教学研究中心 全国AI教材基地 全民数字素养与技能培训基地 其他机构.....

资源支撑

资源层

慕课 微课 纸质教材 通用AI通识课 特色AI通识课

平台层

一生一书
一空间

教学评一体化

个性化学习路径

多模态教材

知识生成

系统化资源汇聚

智云课堂

学在浙大

智云学堂

Agent1

Agent2

Agent3

AgentN

“大先生”
智能体
服务广场

“大先生”
智能体应用
开发平台

提示词工程

知识库

插件

工作流

多模型适配

数字教师

知识、能力图谱

AI助教

AI助学

AI助评

AI赋能
教学工具

智能工具
智能交互
个性化路径

“浙大先生”平台

模型支撑

模型层

三乐

智海

三问

GPT

观止

百科

DeepSeek

慧学

GEO

算力支撑

支撑层

超算

智算

普算

浙江大学启真算力中心

当汽车诞生时，无需与之赛跑，而应考个驾照。

