拓展课程

大模型时代

黑马程序员 www.itheima.com | 传智教育旗下 高端IT教育品牌

目录

Contents

大模型概述

- 1. 大模型概念解析
- 2. 大模型的发展历程

大模型概念解析

大模型的定义与特征 01

（1）定义：使用**大规模数据集**训练的深度学习模型，具有**非常高的参数数量**和计算能力。

（2）参数规模：**数十亿甚至数千亿**个参数

（3）特点：它们能够处理和分析大量的数据，从而生成复杂的输出，如自然语言文本、图像等。

（4）重要特征：它们具有强大的泛化能力，能够在多种任务中表现出色。

（5）**独立意义上的大模型：具备涌现能力的深度学习模型**

大模型与传统模型的区别 02

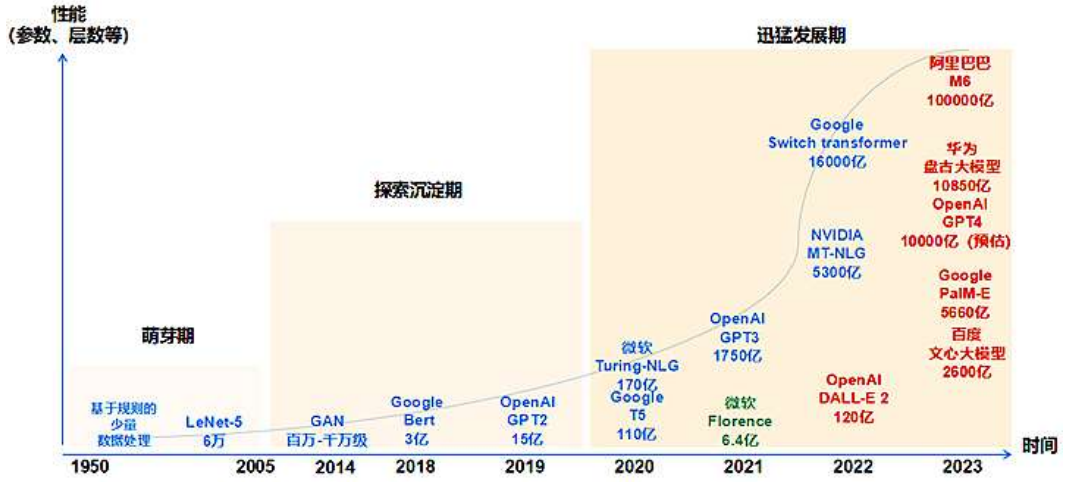
（1）**小模型**通常指参数较少、层数较浅的模型，它们具有轻量级、高效率、易于部署等优点，适用于数据量较小、计算资源有限的场景，例如移动端应用、嵌入式设备、物联网等。

（2）与传统模型相比，大模型拥有更多的参数和更复杂的结构，能够处理更复杂的任务和数据。

（3）与传统模型相比，大模型的训练需要更多的计算资源和时间，但一旦训练完成，它们可以用于解决多种问题。

（4）传统模型通常用于解决特定任务
大模型则更适用于**数据量较大、计算资源充足**的通用场景，例如云计算、高性能计算、人工智能等。

大模型的发展历程



（1）萌芽期（1950-2005）

以CNN为代表的传统神经网络模型阶段

- 1956年，约翰·麦卡锡提出“人工智能”概念
- 1980年，卷积神经网络的雏形CNN诞生
- 1998年，现代卷积神经网络的基本结构LeNet-5诞生

（2）探索沉淀期（2006-2019）

以Transformer为代表的全新神经网络模型阶段

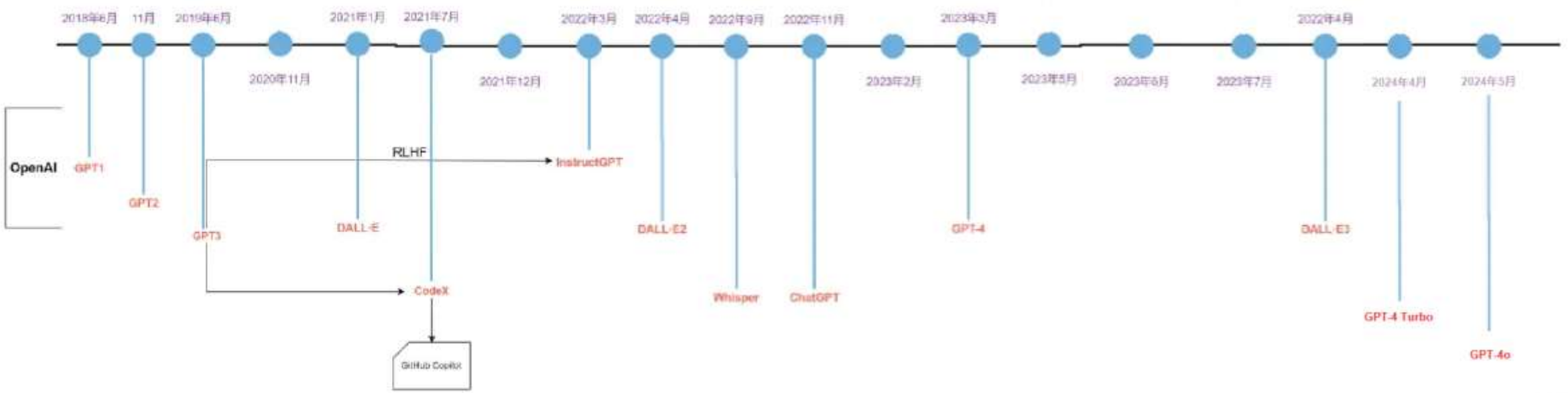
- 2013年，自然语言处理模型 Word2Vec诞生
- 2014年，被誉为21世纪最强大算法模型之一的GAN诞生
- 2017年，Google颠覆性地提出了基于自注意力机制的神经网络结构——Transformer架构
- 2018年，OpenAI和Google分别发布了GPT-1与BERT大模型

（3）迅猛发展期（2020-至今）

以GPT为代表的预训练大模型阶段

- 2020年，OpenAI公司推出了GPT-3
- 2022年11月，搭载了GPT3.5的ChatGPT横空出世
- 2023年3月，最新发布的超大规模多模态预训练大模型—GPT-4
- 2024年4月，最新升级超大规模多模态预训练大模型—GPT-4 Turbo
- 2024年4月19日,Meta正式发布开源大模型——Llama-3
- 2024年5月9日,阿里云发布通义千问2.5,并开源Qwen-110B(国产Llama3)
- 2024年5月14日，OpenAI发布最新多模态大模型 GPT-4o（o-omini全能）
- 2024年7月23日，Meta开源大模型—Llama3.1 405B,多项指标超越GPT-4o
- 2024年7月25日，Mistral Large 2(123B),多项指标超越GPT-4o和Llama3.1

大模型发展现状-海外-OpenAI



大模型发展现状-海外-OpenAI

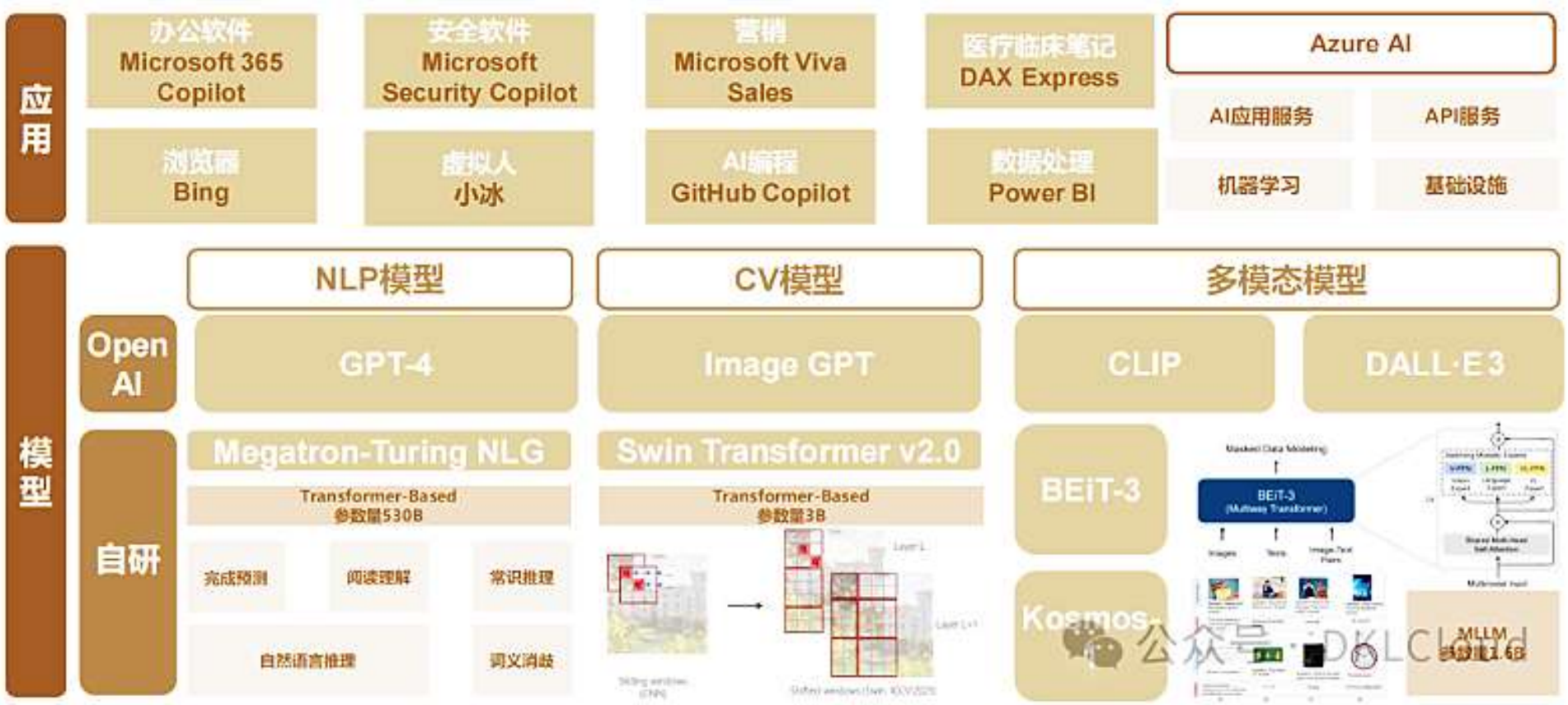
中文数学多步推理基准 SuperCLUE-Math6

模型名称	机构	使用方式	推理等级	综合分数	推理步数加权得分	准确率综合得分
GPT-4o	OpenAI	POE	5	91.77	92.94	90.60
GPT-4-Turbo-1106	OpenAI	API	5	90.71	91.65	89.77
Claude3-Opus	Anthropic	API	5	90.36	91.26	89.46
GPT-4	OpenAI	API	5	88.40	89.10	87.71
通义千问2.5	阿里云	API	5	86.53	87.72	85.33
DeepSeek-V2	深度求索	API	5	86.39	87.81	84.97
文心一言4.0	百度	API	5	85.60	86.82	84.38
GLM-4	智谱AI	API	5	84.24	85.72	82.77
Llama-3-70B-instruct	Meta	模型	5	83.77	85.01	82.53
讯飞星火V3.5	科大讯飞	API	5	83.73	85.37	82.09
ChatGLM-Turbo	智谱AI	API	4	57.70	60.32	55.09
GPT3.5-Turbo	OpenAI	API	4	57.05	59.61	54.50
Qwen-14B-Chat	阿里云	API	4	53.12	55.99	50.26
讯飞星火V3.0	科大讯飞	API	3	40.08	45.27	34.89
ChatGLM3-6B	智谱AI	模型	3	40.90	44.20	37.60
文心一言3.5	百度	API	2	25.19	27.70	22.67
Chinese_Alpaca2_13B	Yiming Cui	模型	2	20.55	22.52	18.58

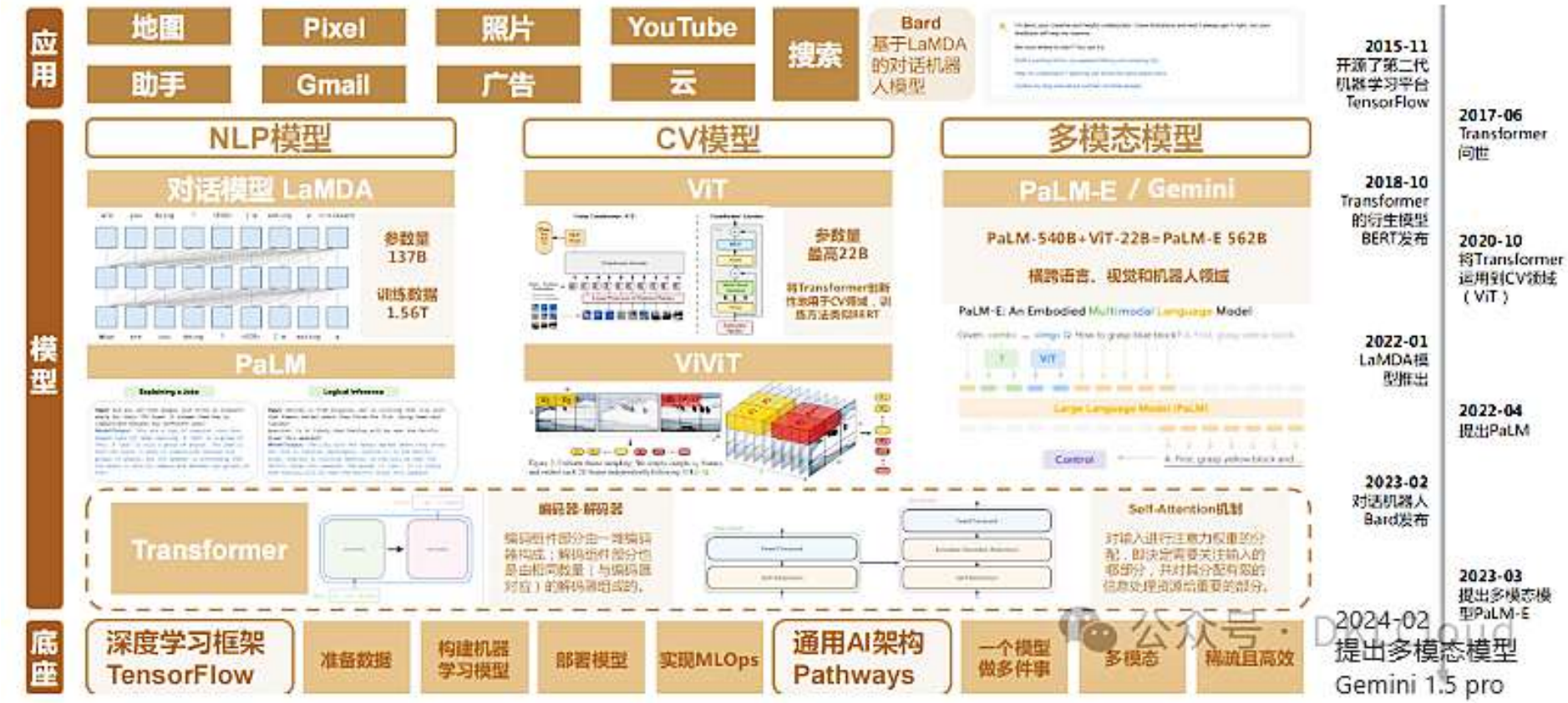
中文原生等级化代码测评基准 SuperCLUE-Code3

模型	SC-Code3总分	初级分数	中级分数	高级分数
GPT-4o	71.68	92.22	82.09	57.89
GPT-4-Turbo-1106	69.57	85.56	79.10	57.89
GPT-4-Turbo-0125	68.00	88.89	80.60	52.63
GPT-4	63.74	90.00	79.10	44.74
通义千问2.5	63.32	85.56	76.12	47.37
Llama-3-70B-instruct	62.57	90.00	71.64	47.37
DeepSeek-V2	62.52	87.78	68.66	50.00
GPT-35-Turbo-0125	55.51	82.22	70.15	36.84
deepseek_coder-6.7b	47.78	67.78	46.27	42.11
Gemini-Pro	46.50	68.89	53.73	34.21
XVERSE-13B-Chat	30.53	63.33	28.36	21.05
qwen-14b-chat	24.67	57.78	25.37	13.16
Code-Llama-13-instruct	21.11	52.22	25.37	7.89
ChatGLM3-6B	15.29	32.22	17.91	7.89

大模型发展现状-海外-微软

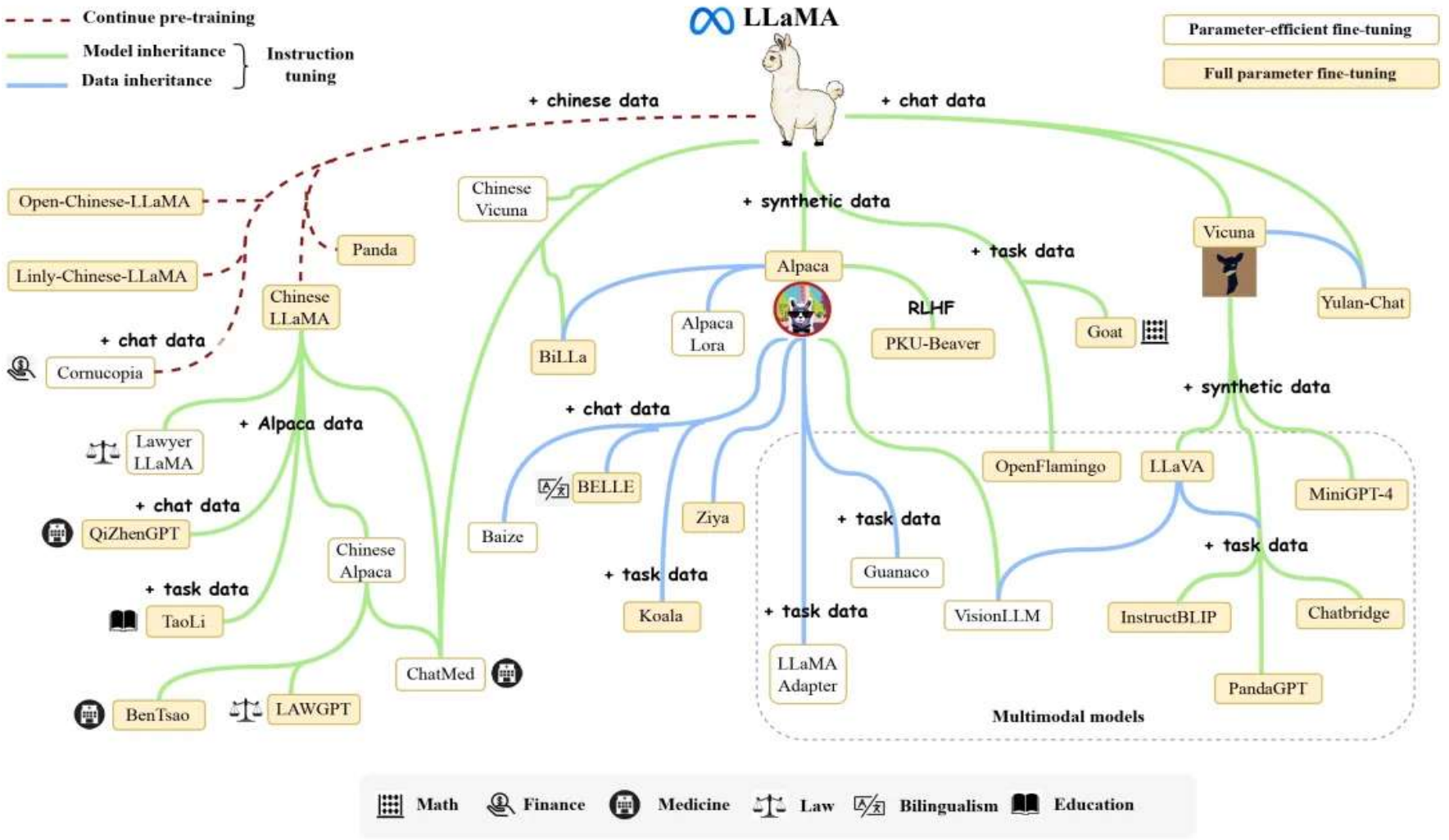


大模型发展现状-海外-谷歌



大模型发展现状-海外-Meta

大模型发展现状-海外-Meta



大模型发展现状-国内-阿里巴巴

大模型发展现状-国内-百度

大模型发展现状-国内-腾讯

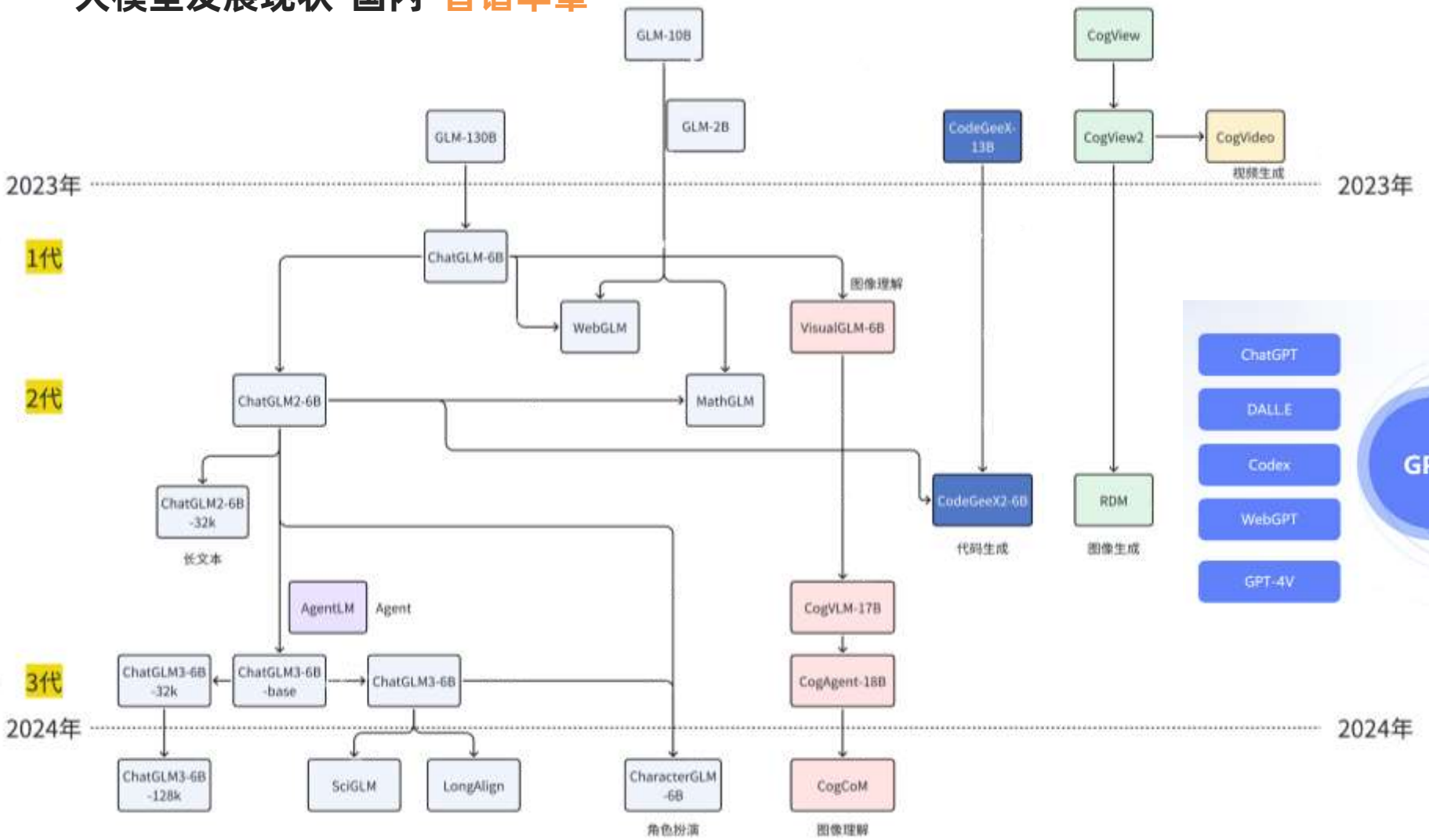
大模型发展现状-国内-华为

大模型发展现状-国内-智源研究院

2018年11月14日，北京智源行动计划正式发布，北京智源人工智能研究院揭牌成立。在科技部和北京市委市政府的指导和支持下，依托北京大学、清华大学、中国科学院、百度、小米、字节跳动、美团点评、旷视科技等北京人工智能领域优势单位共建的新型研究机构。

大模型发展现状-国内-智源研究院

大模型发展现状-国内-智谱华章



大模型发展现状-国内-科大讯飞

大模型发展现状-国内-商汤科技

02

大模型方向市场分析

- 1. 产业规模、政策引导、人才需求
- 2. 工作年限、年薪分析、地域及匹配薪资
- 3. 紧缺人才、核心竞争力、研发方向

大模型市场分析—产业规模

人工智能产业规模快速增长，为人才市场带来新机遇

- 信通院指出，2023年全球人工智能市场收入预计达5132亿美元，同比增长20.7%。截至2023年三季度，全球人工智能企业达到29,542家，中国企业数量仅低于美国，占全球总数的15%；
- 彭博行业研究数据显示，随着企业改变经营方式并对产品和服务进行强化，未来10年，生成式AI有望在硬件、软件、服务、广告、游戏等众多领域创造1.3万亿美元收入，占科技领域总支出的10%–12%，复合年增长率预计达到约42%。

大模型市场分析—政策引导

政策指引生成式人工智能应用创新，鼓励企业汇聚人才

- 生成式人工智能逐渐进入政策红利期，从完善基础设施布局到核心领域应用，政策密集出台，多部门协同发力，以组合拳促发展。

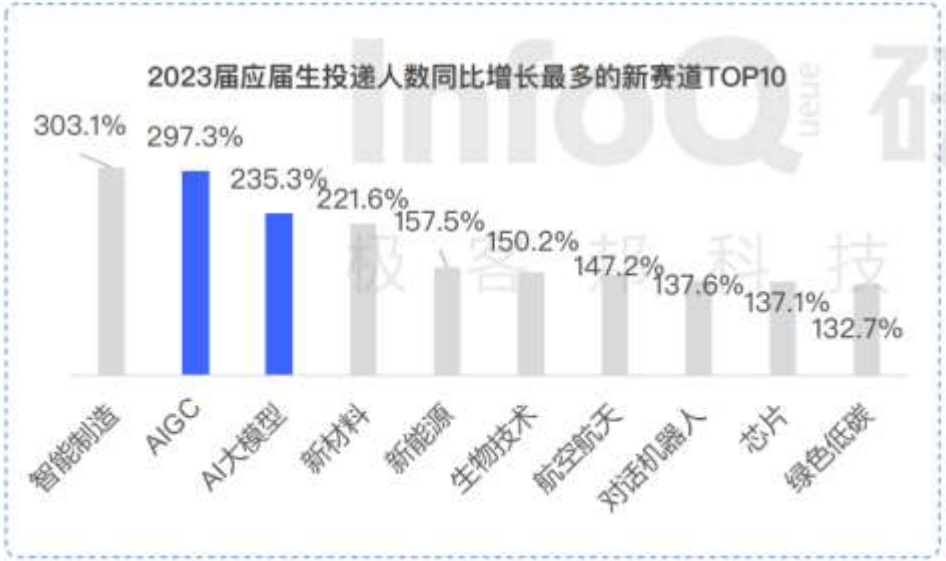
时间	政策	发布主要部门	内容
2023年2月	《数字中国建设整体布局规划》	国务院	✓ 系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局； ✓ 统筹布局一批数字领域学科专业点，培养创新型、应用型、复合型人才。
2023年4月	《生成式人工智能服务管理办法（征求意见稿）》	国家互联网信息办公室	✓ 首次明确了生成式人工智能“提供者”内容生产、数据保护、隐私安全等方面的法定责任及法律依据，确立了人工智能产品的安全评估规定及管理办法。
2023年4月	《关于推进IPv6技术演进和应用创新发展的实施意见》	工业和信息化部等	✓ 推动IPv6与5G、人工智能、云计算等技术的融合创新，支持企业加快应用感知网络、新型IPv6测量等“IPv6+”创新技术在各类网络环境和业务场景中的应用； ✓ 培养IPv6创新人才，丰富人才挖掘和选拔渠道，强化复合型领军人才培养。
2023年7月	《生成式人工智能服务管理暂行办法》	国家互联网信息办公室等	✓ 促进生成式人工智能健康发展和规范应用，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管。
2023年9月	《关于实施专精特新中小企业就业创业扬帆计划的通知》	工业和信息化部等	✓ 按照国家有关规定，动态调整职称专业设置，根据当地产业发展和专精特新中小企业需要，增设人工智能、大数据、工业互联网等新专业。
2023年12月	《关于加快推进视听电子产业高质量发展的指导意见》	工业和信息化部等	✓ 支持骨干企业做大做强，支持人工智能企业研发视听应用大模型。
2023年12月	《“数据要素x”三年行动计划（2024-2026年）》	国家数据局	✓ 以科学数据支持大模型开发，建设高质量语料库和基础科学数据集，支持开展通过人工智能大模型和垂直领域人工智能大模型训练。

大模型市场分析—人才需求

行业的蓬勃发展吸引了大量人才为其创新续航

- 生成式AI的蓬勃发展产生了大量人才需求，在2023届应届生投递人数增长最多的TOP10赛道中，生成式AI和AI大模型分别排名第二和第三，其中生成式AI增长率大幅超越其他赛道。AI大模型、生成式AI、芯片对硕博应届生的需求增长可观，其中AI大模型对博士生的招聘需求更是增长了430.0%。

成为对应届生最具有吸引力的行业之一



吸引大量高学历人才进入行业

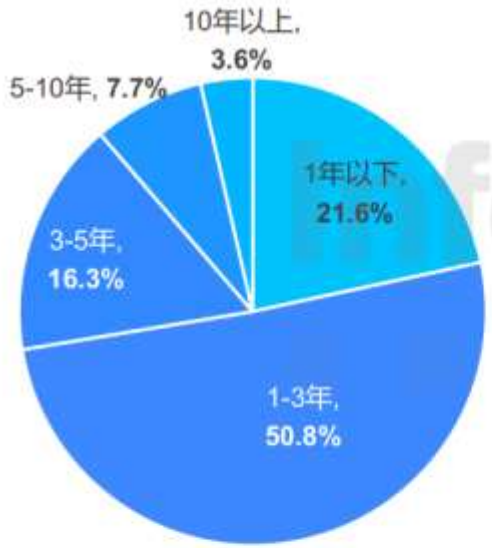
数据来源：猎聘大数据

https://www.infoq.cn

大模型市场分析—工作年限

作为新兴行业从业者，生成式AI开发者普遍相关工作年限较短

生成式AI开发者工作年限分布



各职位平均工作年限（年）

生成式AI开发者职位分布

数据说明："其他"包括个人/自由职业、毕业但尚未工作的毕业生、实习生、数据分析、营销、咨询顾问等类型

数据来源：InfoQ 2023年12月发起的《中国生成式AI开发者画像调研》

https://www.infoq.cn

大模型市场分析—年薪分析

生成式AI开发者50万以上年薪占比高达23.3%

数据来源：InfoQ 2023年12月发起的《中国生成式AI开发者画像调研》，猎聘大数据

https://www.infoq.cn

- InfoQ调研统计，2023年生成式AI开发者人均年收入为36.7万，相关工作经验在3年以上生成式AI开发者的年收入超越均值，近4成生成式AI开发者年收入处于20-50万区间，远超2023年上半年北京招聘平均薪资（18976元/月）。由于AI应用范围广、技术含量高、供需两旺等因素，互联网企业、科技企业、初创企业展现出强大的招聘势头，即使是工作年限较短的生成式AI开发者，薪资水平也超越北京平均招聘薪资水平。

大模型市场分析—薪资情况分析

大模型市场分析—地域及匹配薪资

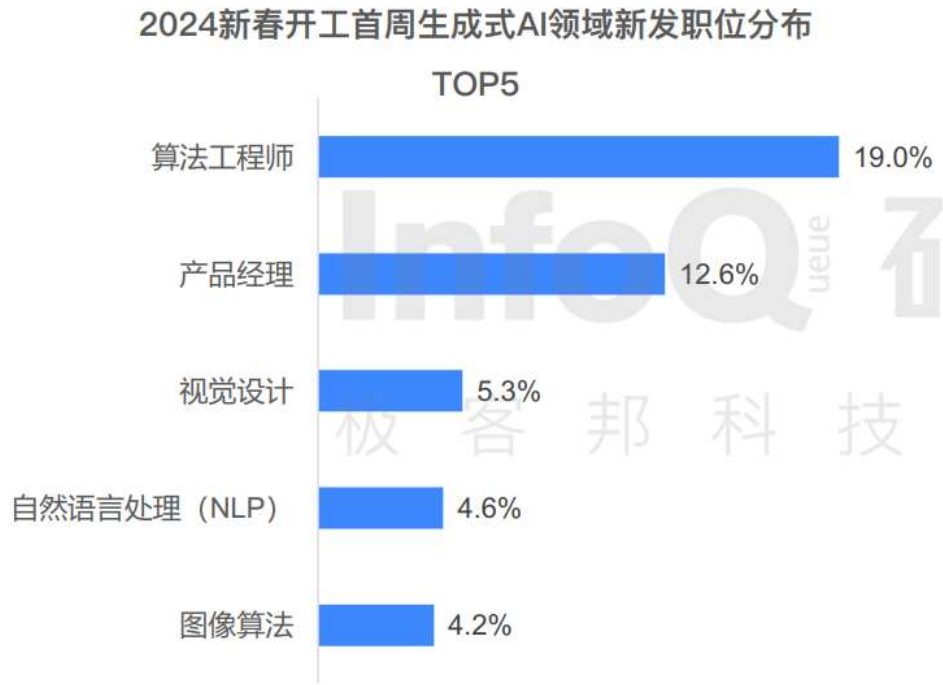
北京生成式AI开发者规模最大，但上海资深生成式AI开发者更多且人均薪资更高

数据来源：InfoQ 2023年12月发起的《中国生成式AI开发者画像调研》 https://www.infoq.cn

- 生成式AI开发者主要集中在北京，广东省和上海属于第二梯队的相关人才聚集地；
- 北京的生成式AI开发者中，超过三成为资深研发人员，人均年薪为44.2万，人均工作年限为2.6年，近六成就职于信息传输、软件和信息技术服务业企业；
- 广东省的生成式AI开发者中，近四成为资深研发人员，人均年薪为39.9万，人均工作年限为2.7年，六成就职于信息传输、软件和信息技术服务业企业；
- 上海的生成式AI开发者中，超过四成为资深研发人员，人均年薪为50.4万，人均工作年限为3.3年，近七成就职于信息传输、软件和信息技术服务业企业。

大模型市场分析—紧缺人才

算法工程师、产品经理是目前市场最为紧缺的人才类型



数据来源：猎聘大数据

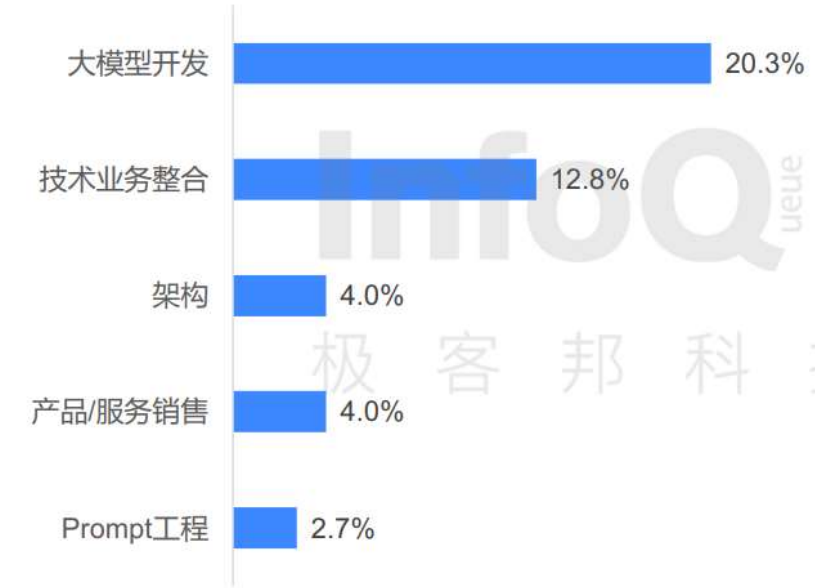
- 在新发职位最多的TOP5职能中，除了算法工程师，产品经理占比最多：
  - 在偏向技术类的岗位职责中，熟悉常见的机器学习和深度学习算法、熟悉常见的生成式模型、熟悉Python/Java/C++等至少一种以上后端语言、熟悉SQL编程成为硬性要求；有相关方向顶级会议/期刊论文或竞赛经验是就职加分项。
  - 在偏向应用类的岗位职责中，有AI基础能力（了解NLP、机器学习、深度学习的大致原理、熟悉主流算法、熟悉至少一种编程语言）、能够与技术无障碍沟通、了解AI在特定行业的应用、熟悉至少一种大语言模型的能力边界及应用场景成为硬性要求，有对接算法和工程经验或有海外互联网App产品经验是就职加分项。

https://www.infoq.cn

大模型市场分析—核心竞争力

大模型研发和业务复合型能力是市场核心竞争力

2024年第一季度生成式AI相关岗位能力要求TOP5占比



数据来源：InfoQ 2024 年 3 月从百度数据、百川智能、月之暗面、智谱收集的405条招聘信息统计获得

https://www.infoq.cn

01 基础研发能力：
算法工程师或数据科学家

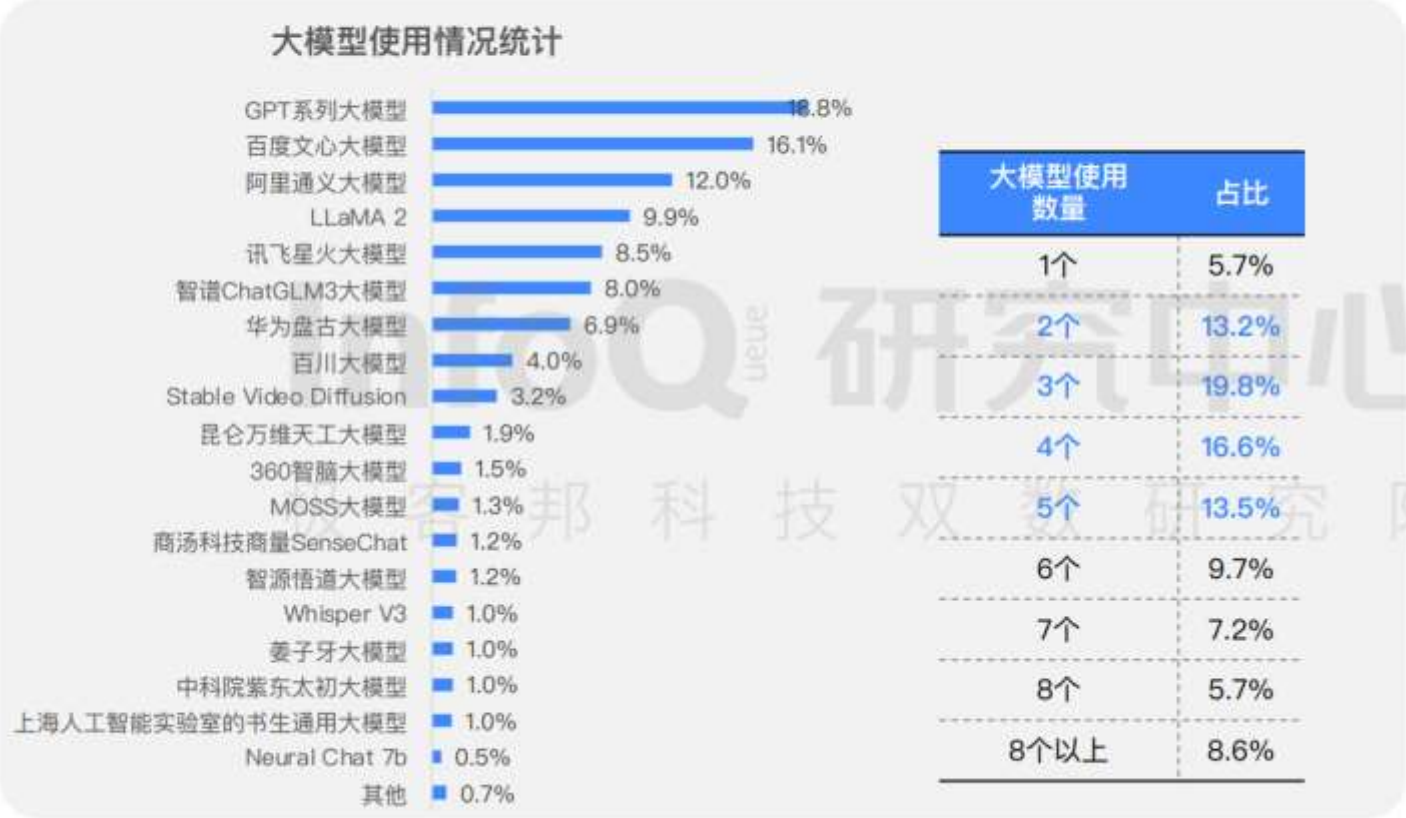
- 第一类紧缺人才为具备基础研发能力的专业技术人员，以**算法工程师或数据科学家**为主，需要熟悉生成式AI技术原理，了解如何去做大模型架构搭建、模型推理及训练，同时需要**关注业务上下游环节**，能够与团队齐头并进。

02 复合能力：
产品经理或跨领域人才

- 第二类紧缺人才为掌握基础技术且了解某个行业的复合型人才。需要知道大模型能够实现的需求和实现程度，既熟悉生成式AI基础原理，也了解某个行业某类业务的解决方案，知道AI在特定行业的应用。能够**将大模型融入到企业的整个生产流程中**，成为企业产品的一部分。技术能力结合行业能力帮助企业快速将产品AI化。

大模型市场分析—大模型使用率

GPT、文心、通义大模型是生成式AI开发者使用率最高的大模型



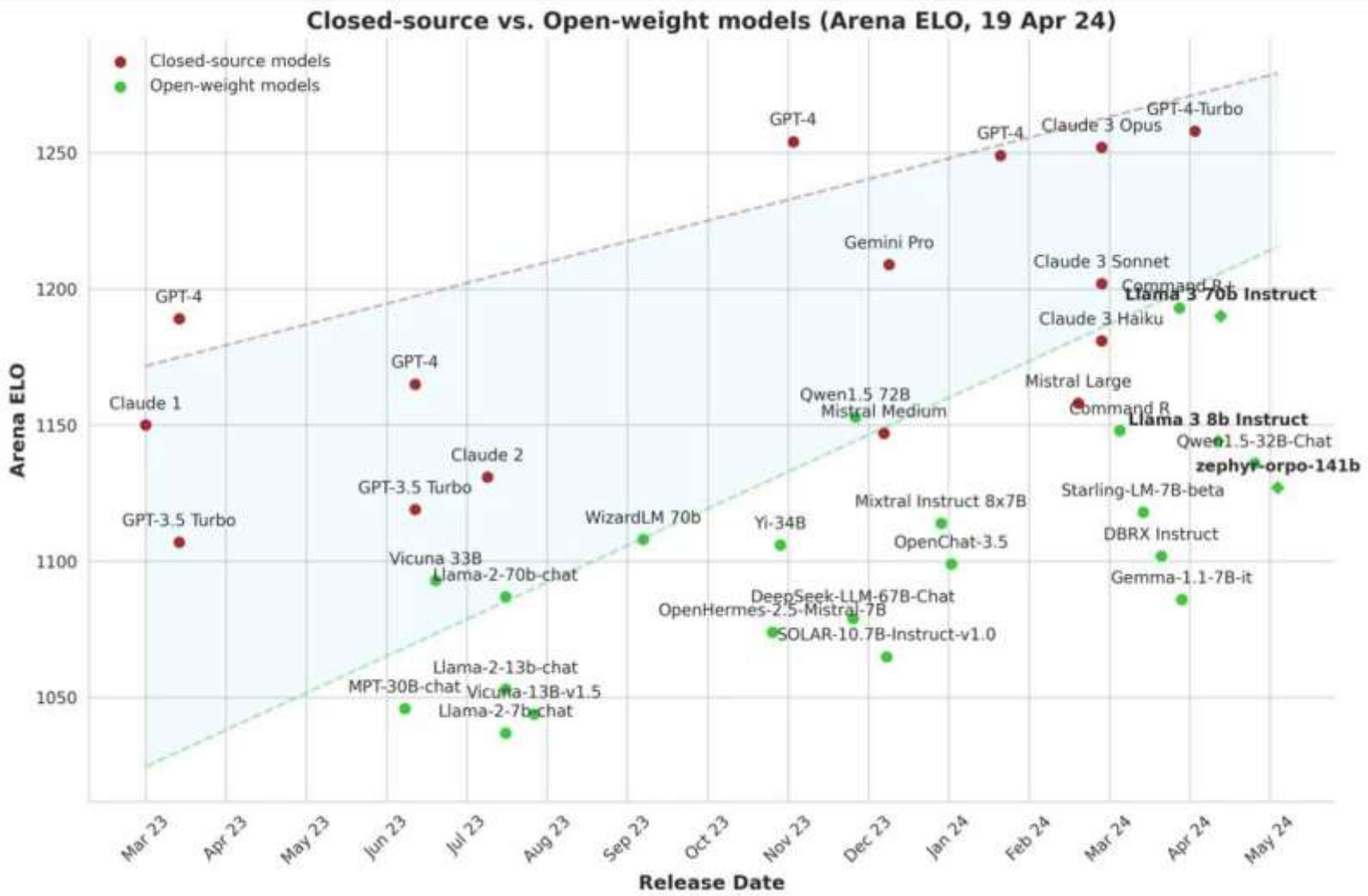
大模型使用数量	占比
1个	5.7%
2个	13.2%
3个	19.8%
4个	16.6%
5个	13.5%
6个	9.7%
7个	7.2%
8个	5.7%
8个以上	8.6%

数据来源：InfoQ 2023年12月发起的《中国生成式AI开发者画像调研》

https://www.infoq.cn

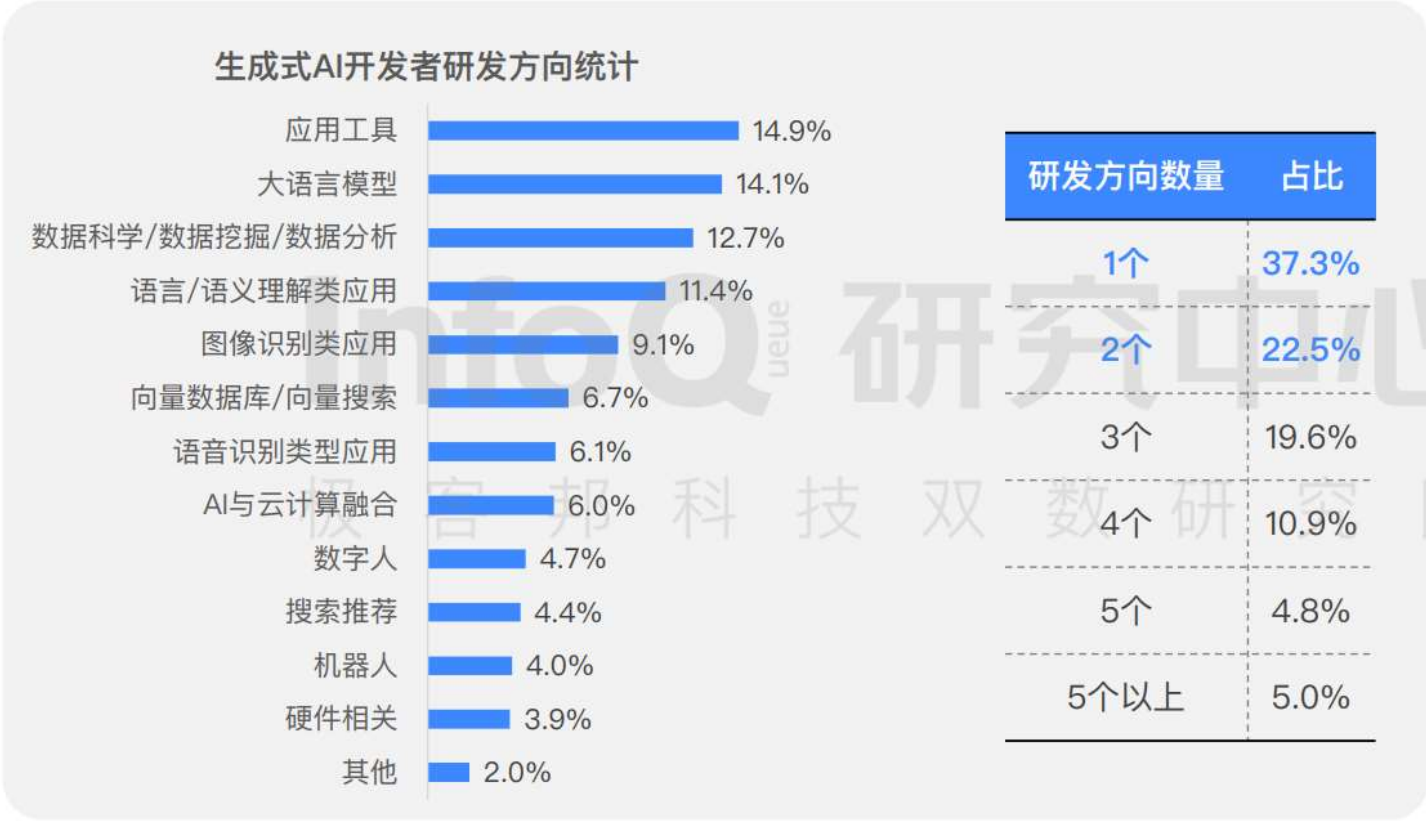
- 生成式AI开发者或企业对大模型的要求不仅是实现通用功能，还需要能够在特定领域、特定场景具备应用价值，真正解决业务痛点。因此，生成式AI开发者和企业逐渐产生更多自建模型的需求，或者通过使用多个大模型综合解决业务难点；
- 对比国内外大模型，在某些细分领域，国内大模型能够更好理解使用者指令，生成式AI开发者会比较输出结果选择与需求更契合的大模型。

大模型市场分析—开源与闭源大模型性能差距



大模型市场分析—研发方向

近6成生成式AI开发者研发方向超过2个，整体人才呈现短缺状态



研发方向数量	占比
1个	37.3%
2个	22.5%
3个	19.6%
4个	10.9%
5个	4.8%
5个以上	5.0%

- 应用工具（如智能编码工具）、大语言模型、数据科学/数据挖掘/数据分析、语言/语义理解类应用（如对话机器人）和图像识别类应用（如拍照搜图）是最主要的五个生成式AI开发者研发方向；
- 近四成生成式AI开发者工作内容集中在特定方向，其中应用工具和语言/语义理解类应用是主要聚焦方向。

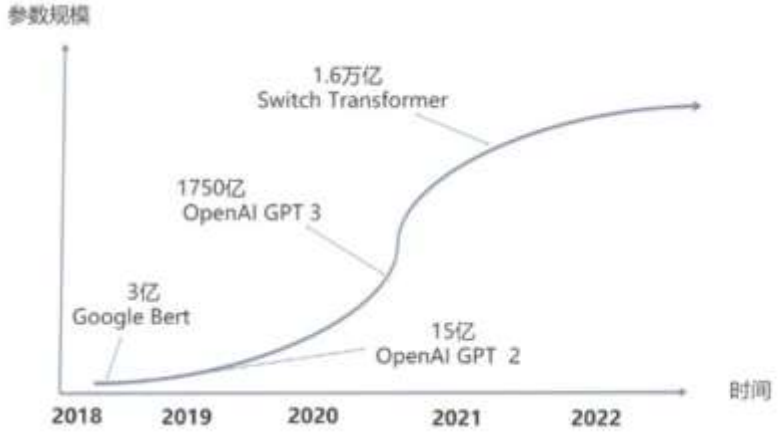
数据来源：InfoQ 2023年12月发起的《中国生成式AI开发者画像调研》

https://www.infoq.cn

大模型时代前沿技术

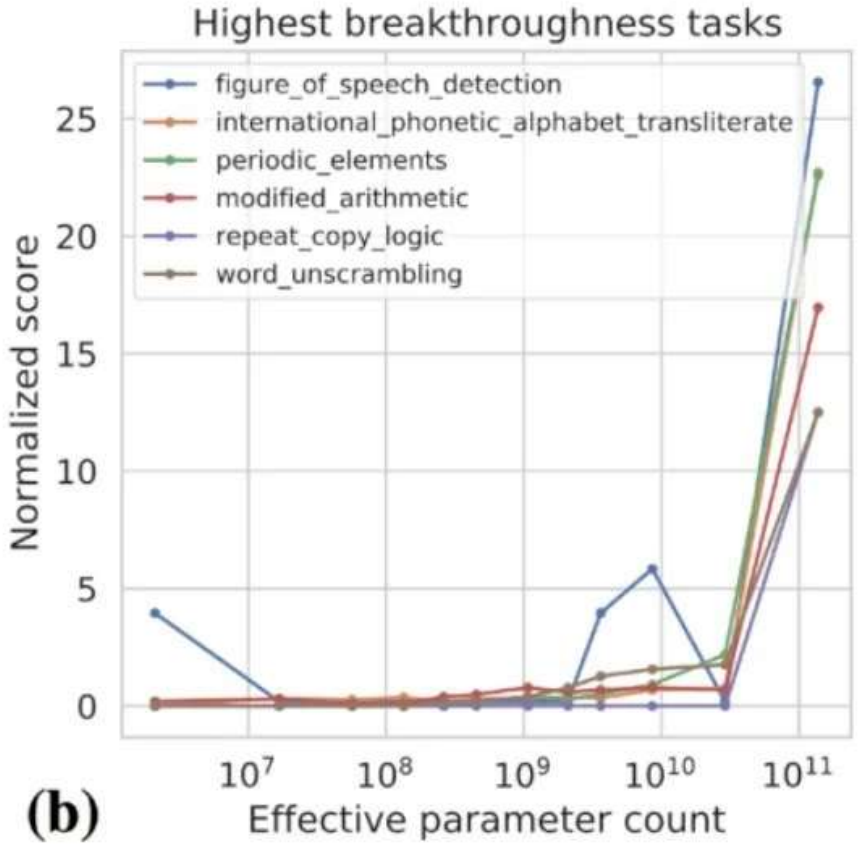
- 1. 基础概念（智能涌现），显存占用计算
- 2. 基础架构（ Transformer ）
- 3. Agent-大模型改变世界的“钥匙”
- 4. 具身智能-人工智能的下一个浪潮

智能涌现



规模大的Large Language Model：

- ✓ GPT 3.0 :175B
- ✓ GPT 3.5:175B
- ✓ LaMDA:130B
- ✓ Gopher:280B
- ✓ PaLM:540B
- ✓ PaLM-E:566B



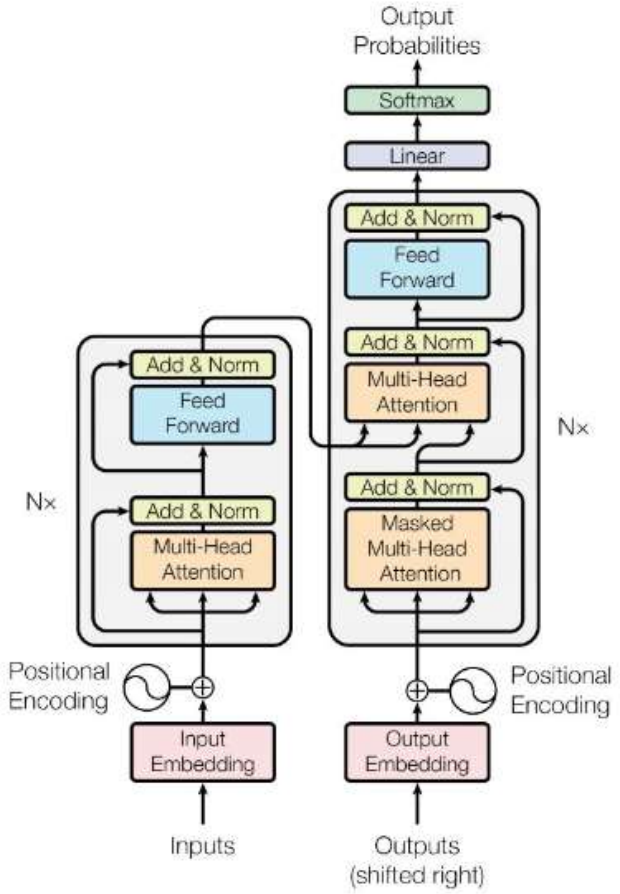
涌现能力：多步骤构成的任务

智能涌现

- 人的大脑一般有**120到140亿**个神经元。
- 所谓“涌现”,在大模型领域指的是当模型突破某个规模时，性能显著提升，表现出让人惊艳、意想不到的能力。比如语言理解能力、生成能力、逻辑推理能力等。一般来说，模型在**100亿到1000亿参数**区间，可能产生能力涌现。
- 强大的逻辑推理是大语言模型“智能涌现”出的核心能力之一，好像AI有了人的意识一样。而推理能力的关键，在于一个技术——**思维链（Chain of Thought，CoT）**。
- 百亿参数是模型具备涌现能力的门槛，千亿参数的模型具备较好的涌现能力。但这并不意味着模型规模就要上升到万亿规模级别的竞争，因为现有大模型并没有得到充分训练。

  如 GPT-3 的每个参数基本上只训练了 1-2 个Token
  DeepMind 的研究表明，如果把一个大模型训练充分，需要把每个参数量训练 **20 个 Token**。

Transformer



1.输入处理：Transformer 首先将输入数据编码为模型可以理解的格式，通常使用嵌入来合并序列中每个元素的位置。

2.注意力机制：注意力机制的核心是计算一个分数，用来表示在理解当前元素时对输入序列的其他部分的关注程度。

3.编码器-解码器架构：Transformer模型由处理输入的编码器和生成输出的解码器组成。每个层都包含多个层，可细化模型对输入的理解。

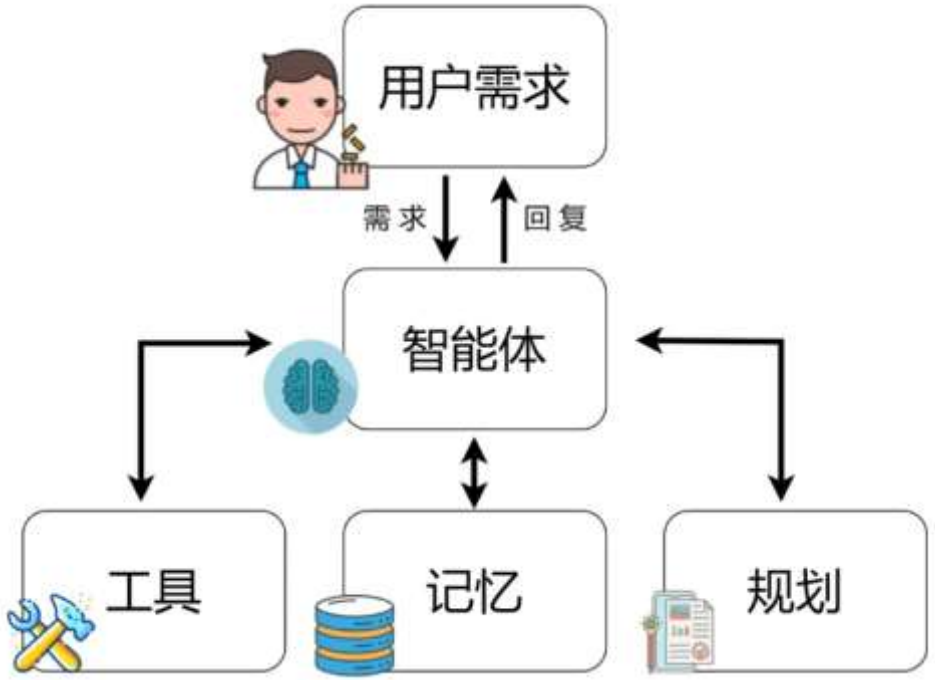
4.多头注意力：在编码器和解码器中，多头注意力允许模型同时关注来自不同表示空间的序列的不同部分，从而提高其从不同上下文中学习的能力。

5.位置前馈网络：在注意之后，一个简单的神经网络会单独且相同地处理每个位置的输出。这通过残差连接与输入相结合，然后进行层归一化。

6.输出生成：然后，解码器预测输出序列，该序列受到编码器上下文及其迄今为止生成的内容的影响。

好书推荐：《深入浅出Embedding》

Agent-大模型改变世界的“钥匙”



[AI Agent--大模型时代重要落地方向]

- **画像模块：**主要描述 Agent 的背景信息
- **记忆模块：**主要目的是记录 Agent 行为，并为未来 Agent 决策提供支撑
- **规划模块：**主要目的通过分解为必要的步骤或子任务来回应用户请求
  - **任务分解技术：**思维链（COT）、思维树（TOT）
  - **反思与批评机制方法：**ReAct 、Reflexion
- **动作模块：**主要作用是通过外部环境（例如Wikipedia搜索API、代码解释器和数学引擎）来获取信息或完成子任务

Agent-大模型改变世界的“钥匙”——AI小镇

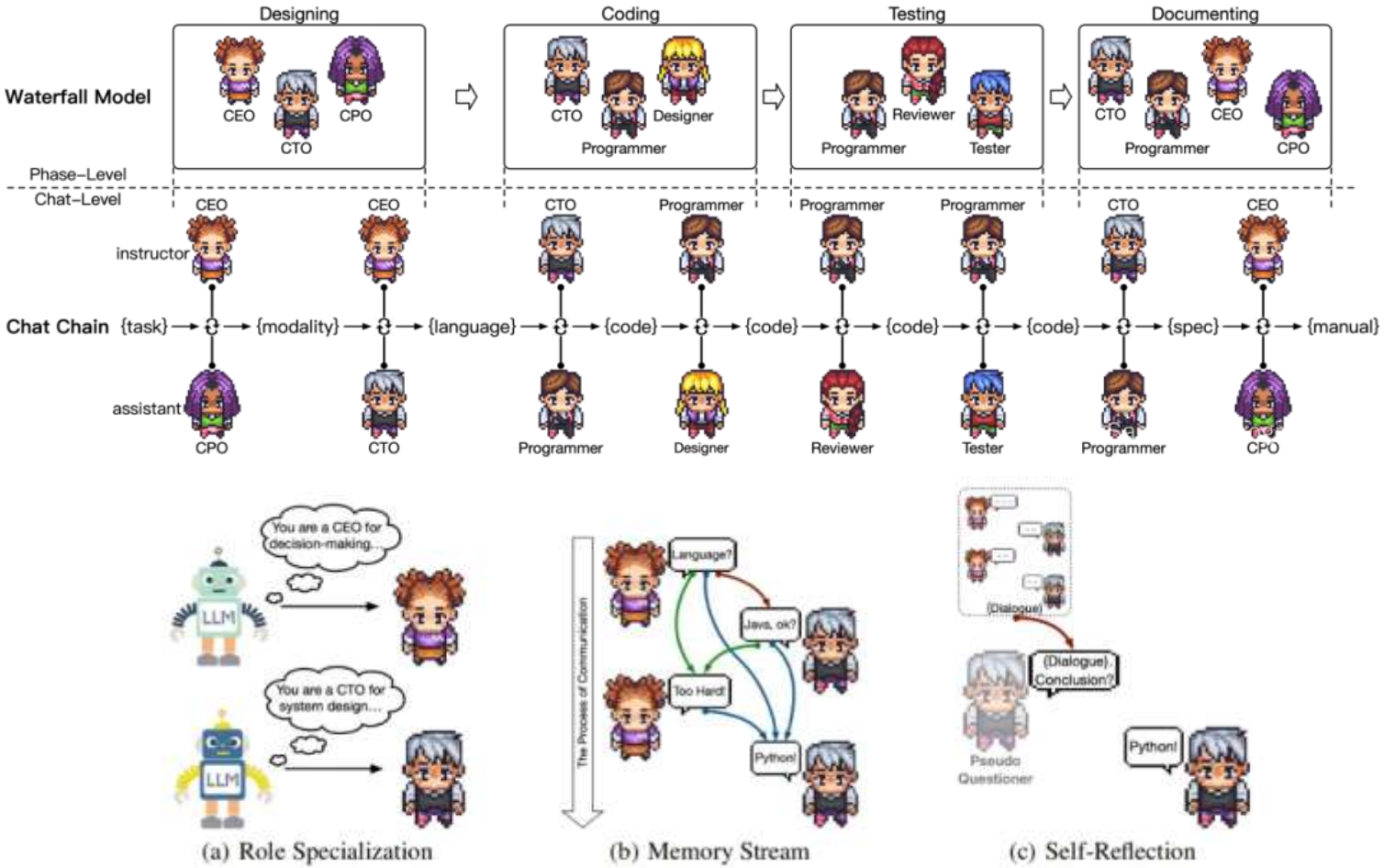
- 8个智能体，复刻西部世界

Alex：一身黑色西装，满头金发的男生，喜欢绘画、编程和阅读科幻书籍。
Alice：一位杰出的科学家，凭借自己的智慧和洞察力，发现了宇宙中无人能解的奥秘。
Peter：一个虔诚的教徒，倾向于从宗教角度解释世界万物。
Bob：头发花白的老爷爷，性格有些孤僻，所以喜欢园艺这项独处的活动。
……

- 主要功能：生活、交友、探索

AI Town

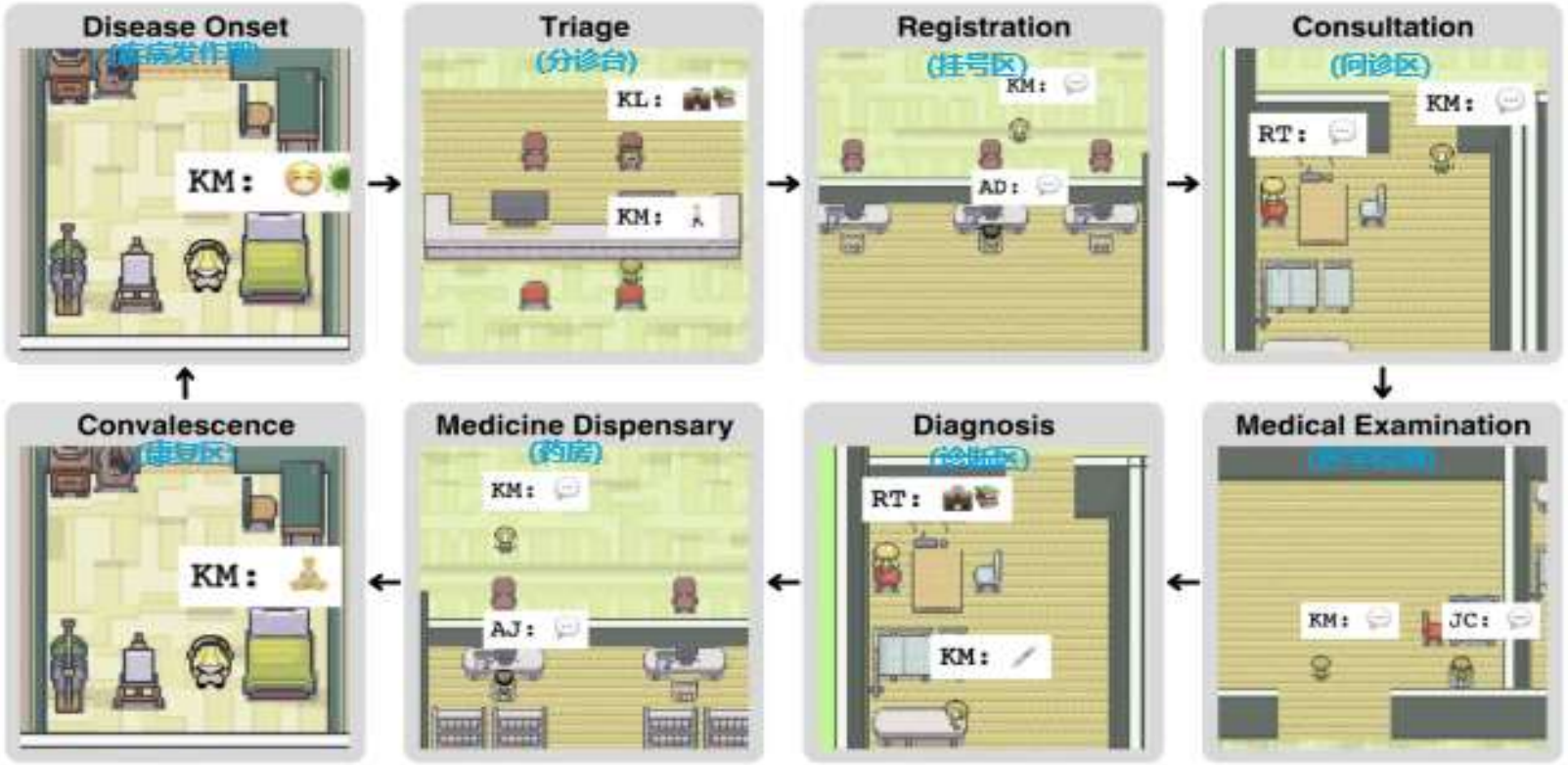
Agent-大模型改变世界的“钥匙”— ChatDEV(面壁智能+清华大学NLP实验室等)



[CHATDEV-软件开发的交流代理，这是让ChatGPT开软件公司么？]

Agent-大模型改变世界的“钥匙”— Agent Hospital(清华大学)

- 虚拟世界中，所有的医生、护士、患者都是由LLM驱动的智能体，可以自主交互，并能够实现自主进化。
- 模拟整个诊病看病的过程，包括分诊、挂号、咨询、检查、诊断、治疗、随访等环节。
- 进化后的医生智能体，在涵盖主要呼吸道疾病的MedQA数据集子集上，实现高达93.06%的最新准确率。

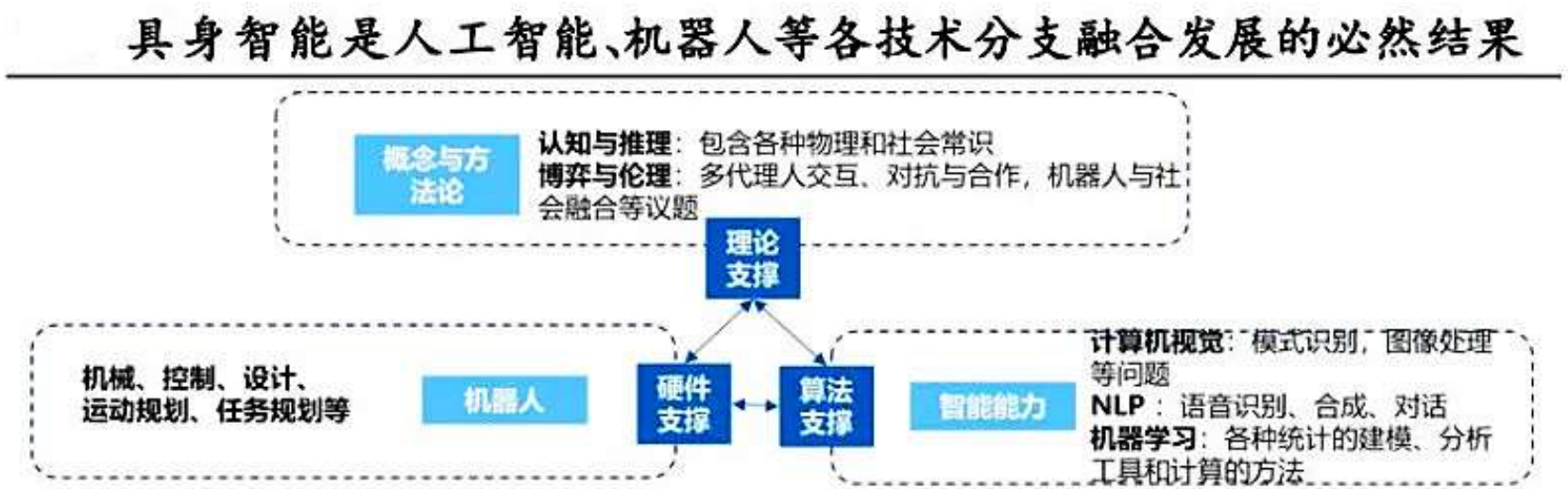


清华首个AI医院小镇来了！AI医生自进化击败人类专家，数天诊完1万名患者

具身智能-人工智能的下一个浪潮

定义：Embodied AI = Embodied Intelligence = 具象AI = 具身智能

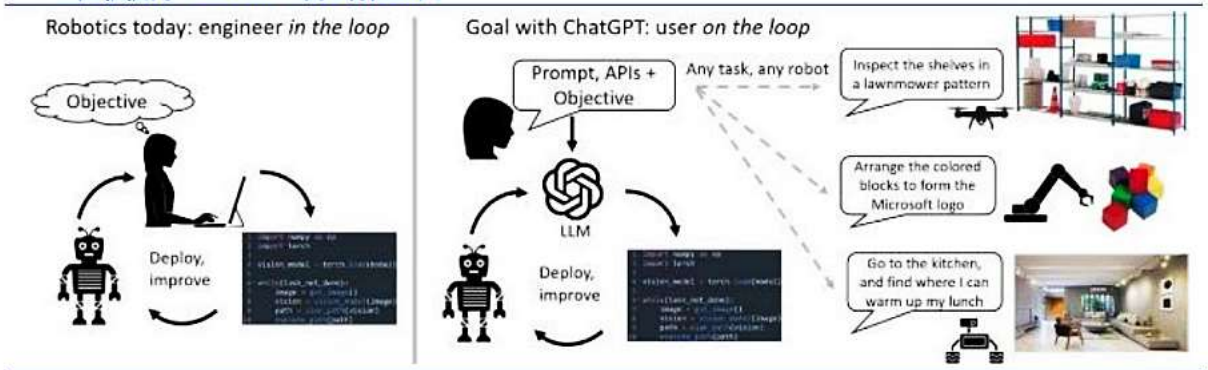
有身体并支持物理交互的智能体，如家用服务机器人、无人车等。——“身体力行”



具身智能-人工智能的下一个浪潮

PIE 方案：具身智能的解决方案之一

微软使用 ChatGPT 来控制机器人

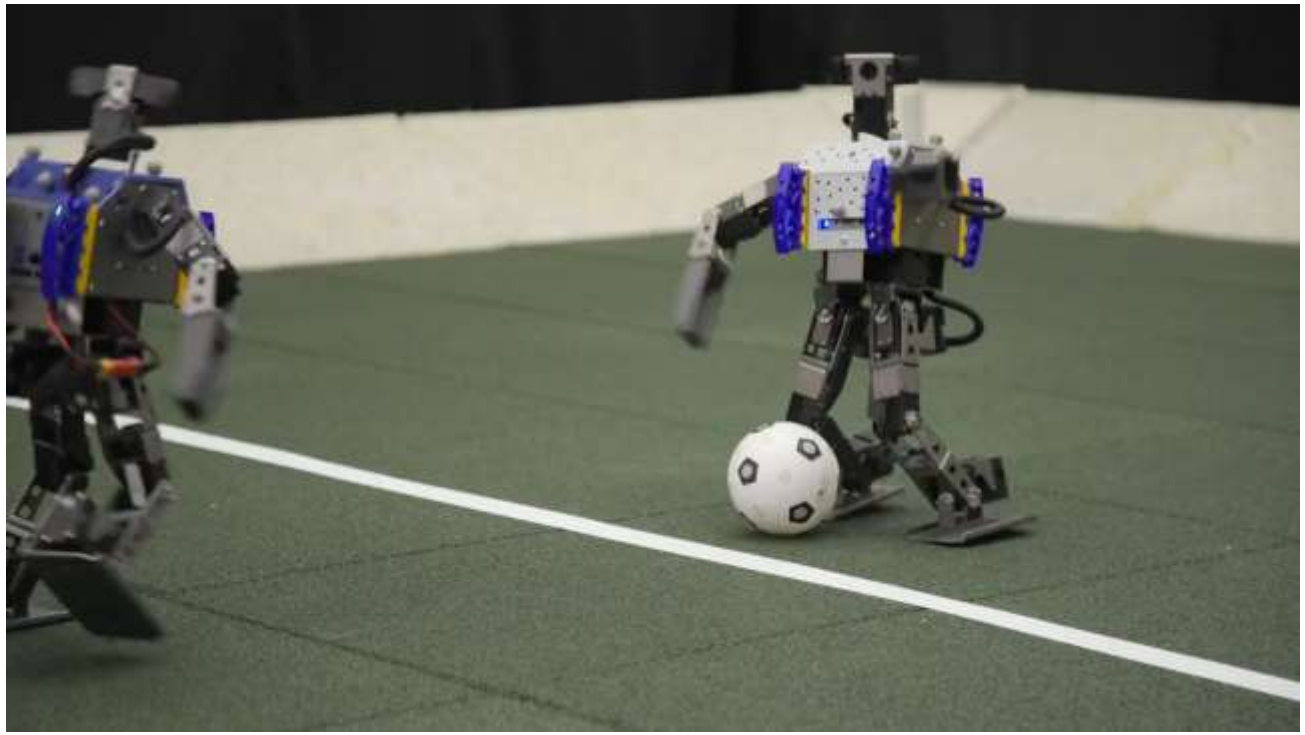


具身智能-最新成果

- **Google DeepMind 研发的具身智能体（Agent）**
  一个微型人形机器人，**不仅可以快速“奔跑”“过人”“进攻”，还可以阅读比赛，正确预测足球移动方向，以及阻挡对手射门等。**

- 在实验中，与对比基线相比，该机器人奔跑速度快了 **181%**，转身速度快了 **302%**，（跌倒后）起身时间缩短了 **63%**，踢球速度快了 **34%**，同时也能有效地将各种技能结合起来，远远超出了人们此前对机器人的固有认知。

- 相关研究论文以“Learning agile soccer skills for a bipedal robot with deep reinforcement learning”为题，以封面文章的形式已发表在 Science 子刊 Science Robotics 上。



DeepMind推出具身智能“足球运动员”，过人、射门、防守样样精通

推荐阅读

弱智吧：大模型变聪明，有我一份贡献 (myzaker.com)

音乐ChatGPT时刻来临！Suno V3秒生爆款歌曲，12人团队创现象级AI

谷歌更新Transformer架构，更节省计算资源！50%性能提升

AIGC启元2024


黑马程序员
www.itheima.com
传智教育旗下高端IT教育品牌