2025

通向AGI之路
全球人工智能展望报告

指导机构：天津市人工智能学会、深圳市人工智能行业协会

撰写机构：至顶智库

支持机构：中关村科学城公司、至顶科技、与非网

报告背景

当前，人工智能正朝着AGI的方向迈进。伴随技术突破与产业应用的深度融合，人工智能进入全新发展阶段。2025年被视为“智能体元年”，AI Agent已成为驱动产业变革的核心力量。智能体通过自主任务规划、动态决策与闭环执行，实现从被动响应指令到主动解决复杂问题的跨越。与此同时，各类AI硬件迎来新一轮迭代升级，从轻量化AI眼镜到便携AI录音卡片，硬件设备在便携性与场景适配性方面不断突破。“多模态模型”进一步打破技术边界，实现语言理解、视觉识别、音频处理等核心能力的深度融合。“世界模型”扮演更为重要的角色，基于内在模拟理解现实世界的物理与因果结构，通过预设未来场景指导决策。

在此背景下，天津市人工智能学会、中关村科学城公司、深圳市人工智能行业协会、至顶科技、至顶智库、与非网联合发布《通向AGI之路—2025年全球人工智能展望报告》。报告从AGI特征出发，全面总结全球人工智能产业的主要参与者、典型产品和应用，对AGI发展的关键领域、核心技术进行分析解读，并呈现当前关注度较高的智能体、AI硬件及应用进展。最后，报告对AGI的未来发展方向进行展望。报告为政府部门、行业从业者、教育工作者以及社会公众更好了解2025年全球人工智能的发展进程，以及探索通向AGI的前进路径提供参考。

报告支持专家

中国信通院人工智能研究所副总工程师—王蕴韬

达观数据董事长兼CEO—陈运文

至顶科技CEO兼总编辑—高飞

LangGPT社区创始人—云中江树

明略科技副总裁兼研发团队负责人—李梦林

明略科技高级产品总监/DeepMiner产品负责人—黄楠

Plug and Play璞跃中国基金合伙人—朱晓雯

Plug and Play璞跃中国投研负责人—杨钧

报告目录

1. AI演进路径与产业概况

人工智能演进路径　　何为通用人工智能

人工智能全景图谱　　人工智能发展路线图

1.1 通用人工智能成为AI演进路径的关键节点

弱人工智能
Artificial Narrow Intelligence

在特定领域具有感知能力的智能

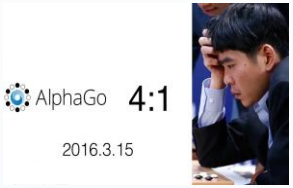
- 任务专用
- 不可迁移

深蓝计算机：

1997年5月，IBM“深蓝计算机”首次击败人类冠军卡斯帕罗夫。

AlphaGo：

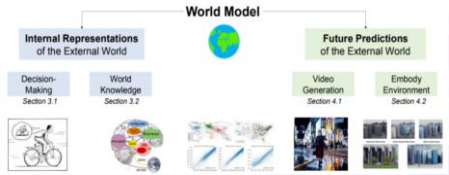
2016年3月，AlphaGo与围棋世界冠军、职业九段棋手李世石进行围棋人机大战，最终击败李世石。



通用人工智能
Artificial General Intelligence

在跨领域达到人类认知能力的智能

- 自主学习
- 跨域推理



世界模型：

实现AGI的关键路径之一，世界模型融合物理理解、推理、规划与持久记忆能力

具身智能：

实现AGI的关键路径之一，具身智能的本质是让AI从虚拟的信息处理走向真实的物理实践。

超级人工智能
Artificial Super Intelligence

在全领域拥有超越人类能力的智能

- 自主创新
- 自主实现

量子神经网络（QNN）：

基于量子力学原理设计的新型计算模型，将传统神经网络架构与量子计算技术相结合。核心机制通过引入量子门避免量子比特的过早坍缩，提升计算效率。

全脑模拟（WBE）：

通过逆向工程复制生物大脑的认知功能。包含神经符号系统融合和意识连续性测试等技术体系。

递归自我改进（RSI）：

在无人工干预下，自主增强自身能力。包含元学习控制器、代码自主修改和目标函数进化等技术路径。

资料来源：至顶智库结合公开资料整理绘制。

1.2 何为通用人工智能（AGI）

OpenAI

OpenAI联合创始人兼首席执行官Sam Altman表示，AGI是一个能够在人类水平上解决许多领域日益复杂问题的系统。

Meta首席AI科学家Yann LeCun提出，“先进机器智能”（Advanced Machine Intelligence，AMI），AMI不追求通用性，而是一种能够理解物理世界、具备推理规划能力、拥有持久记忆并服从目标导向的智能形式。

Meta

ANTHROP\C

Anthropic创始人兼首席执行官Dario Amodei认为，AGI拥有完整的数字接口、可以自主规划并长期执行任务、没有物理实体但可以控制与其连接的任何机器人，训练资源可以重新部署，以运行数百万个示例并且每个示例可独立运行。

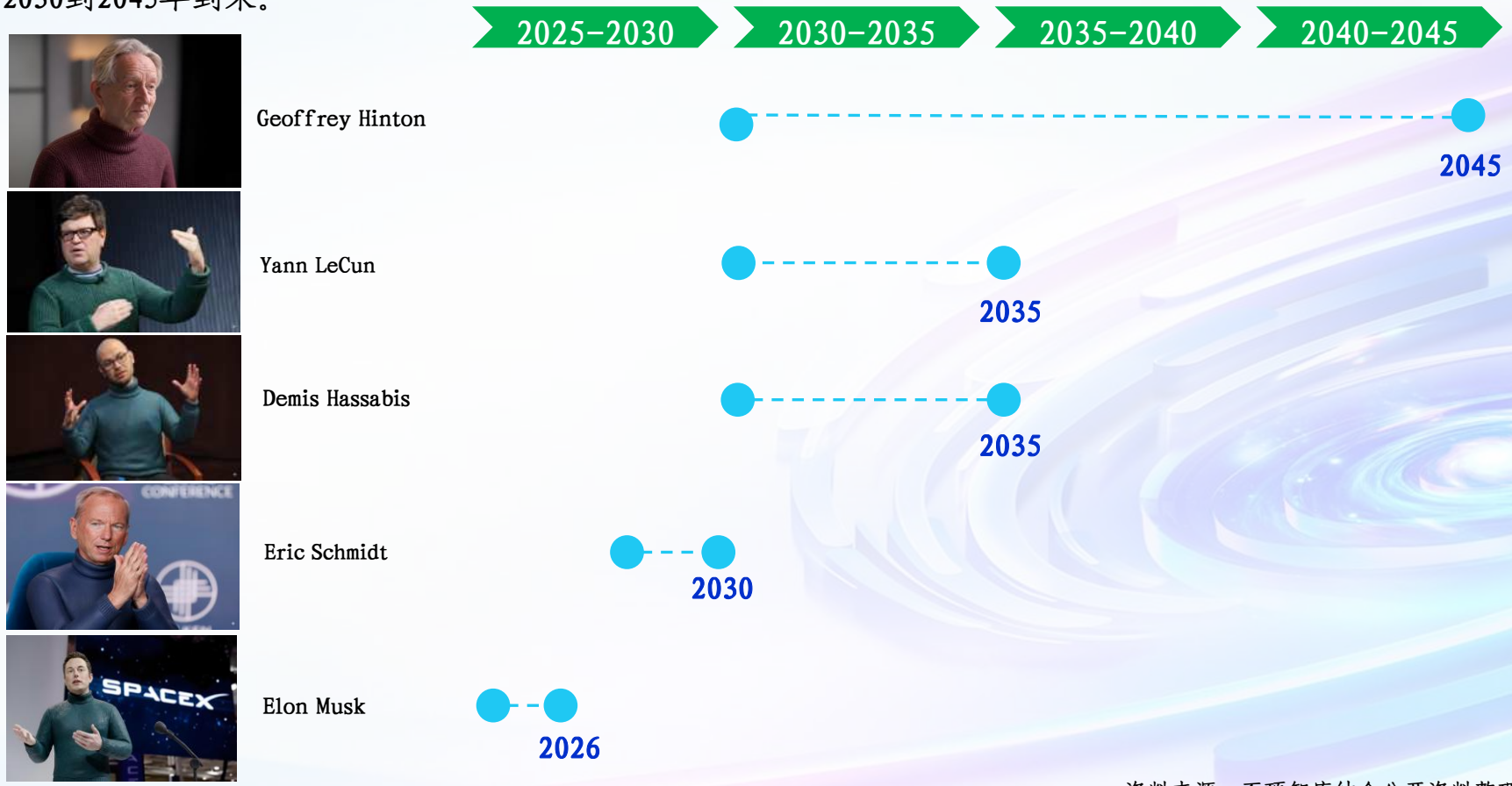
Google DeepMind

Google DeepMind联合创始人兼首席执行官Demis Hassabis提出，真正的AGI需要证明“该系统能做到历史上最优秀的人类用相同大脑架构做到的事情。

资料来源：至顶智库结合公开资料整理绘制。

1.3 通用人工智能何时到来

关于AGI落地时间，全球人工智能领域的业内代表纷纷做出预测。xAI创始人Elon Musk较为乐观，他认为AGI已初具雏形，有望在2026年到来。Google DeepMind联合创始人兼首席执行官Demis Hassabis与Meta首席AI科学家Yann LeCun均认为AGI会在2030到2035年到来。图灵奖得主Geoffrey Hinton则持相对保守的立场，预测AGI将在2030到2045年到来。



资料来源：至顶智库结合公开资料整理绘制。

1.4 2025全球人工智能全景图谱

2025年全球人工智能全景图谱包含应用硬件层、模型层以及基础设施层。其中，应用硬件层广泛覆盖各细分领域，涉及智能体、智能助手、智能硬件、具身智能、C端/B端各类应用；**模型层**包含多模态基础模型、图像模型、视频模型、语音模型、推理模型、开源模型；**基础设施层**涵盖AI芯片、AI服务器、AI计算集群、开发平台、数据服务。图谱中涉及各领域全球具有代表性的AI应用、AI硬件、AI模型以及典型企业，为读者提供更为详实的参考信息。

2025全球人工智能全景图谱概览

应用硬件层								
通用智能体	编程智能体	办公智能体	营销智能体	金融智能体	医疗智能体	客服智能体	HR智能体	工业智能体
19	10	26	10	6	6	8	7	6
智能体开发平台	智能助手	AI手机	AI眼镜&录音	AI PC	智能汽车	具身智能	AI搜索	AI办公
44	31	11	13	7	7	10	14	26
AI写作	AI图像	AI视频	AI音乐	AI音频	AI + 营销	AI + 医疗	AI + 金融	AI + 教育
14	31	20	9	11	22	15	16	14

模型层					
多模态基础模型	图像模型	视频模型	语音模型	推理模型	开源模型
17	7	9	8	25	23

基础设施层				
AI芯片	AI服务器	AI计算集群	开发平台	数据服务
13	5	9	10	10

注：数字代表图谱所涉及的企业、产品应用或模型数量。

资料来源：至顶智库整理绘制。

2025全球人工智能全景图谱—智能体(AI Agent)

AI Agent应用

通用智能体

ChatGPT agent · Genspark Super Agent · Pokee AI · HyperWrite AI Agent · Agent Maven · Otto · Genspark Super Agent · superhuman AI agents

Manus · Flowith · MiniMax Agent · AutoGLM 沉思 · 百度心响 · 万智 Agent · 夸克 AI超级框 · 天禧个人 超级智能体 · 天工超级 智能体 · YOYO智能体 · 纳米AI

编程智能体

Codex agents · AlphaEvolve Gemini CLI Jules · Claude Code · OpenHands · Anysphere-Cursor Composer Agent

通义灵码 · 文心快码 Comate Zulu · CodeBuddy Agent · 美团NoCode · aiXcoder Agent

办公智能体

Joule Agents · Researcher Agent Analyst Agent · AI Agent · Search Agent · Bardeen · Agent Assist · AI Agent · Meeting Agent · Breeze Agents · lutra · Basis agents · Rox Agent · Workist agent

乐享企业 超级智能体 · 京东云JoyAgent · 讯飞智文 讯飞文书 · Shadow AI · WPS灵犀 · 校对通多模态 内容校对智能体 · 达观Agent · 智能分析 SwiftAgent · 差旅智能体 · boardmix AI 智能体 · 市场分析智能体 合同审核智能体 · 自动化魔术师 · CoMi Agent

营销智能体

WordLift Agent · AI Sales Agents · Piper the AI SDR · Jasper Agents · Ava sales agent

Auto-Pilot Agents · Xaver AI agent · 企点营销云Agent · DeepMiner 智能体 · 有赞智能体

金融智能体

AI Banker Agent · Roots' AI Agents · Concourse's AI Agents

Alice Agent · AI合规助手 · 容犀Agent & Copilot

医疗智能体

Voice AI agent · Healthcare Agent · AI Agents

AI Agents · AI Contact Center Agent · 医疗智能体

客服智能体

Voice AI Agents · Agent Echo · Clerk Agent · avientic AI agents

ChatBot · AI Voice Agent · 蚂蚁数科 客服智能体 · 七陌客服智能体

HR智能体

AI HR Agent · Borderless AI · Max · Moonhub's AI Recruiter

人事智能助理 · 候选人 筛选智能体 · iBuilder

工业智能体

Industrial Copilot · Industrial AI agents · 设备维护智能体

工业装备节能 智能体 · 设备管理 智能体 · 赛意制造业 智能体

资料来源：企业/产品/应用排序不分先后，至顶智库结合公开资料整理绘制。

2025全球人工智能全景图谱—智能体开发平台&智能助手

AI Agent开发平台

国外

- Amazon Bedrock AgentCore
- NVIDIA NeMo
- Azure AI Foundry
- Agentforce 3
- Dify
- Sana Agents
- UiPath Agentic Automation
- AI Agent Studio
- GTM Agent Platform
- Wizr Enterprise AI Platform
- Agent OS
- AI Studio
- Engineering Multi-Agent AI Systems
- Dust Developer Platform
- Conversational AI
- Agent Studio
- Mindpal Platform
- Agent Operating Procedures (AOPs)
- Bricklayer AI
- n8n
- AI HQ
- Workato ONE
- AgentX
- Beam Platform

国内

- 字节扣子开发平台
- 百度文心智能体平台
- 阿里云百炼应用开发平台
- 腾讯元器
- 腾讯云智能体开发平台
- JoyAgent智能体平台
- 小艺开放平台
- “百宝箱”智能体开发平台
- 讯飞星辰智能体平台
- 智谱清言
- SkyAgents开放平台
- OPPO 智能体平台
- 汇智智能Gnomic智能体平台
- 极简未来AI智能体搭建平台
- 星辰智能体平台
- BetterYeah AI Agent开发平台
- 智能体平台
- 句子秒懂
- 意言科技TyrionCAIP智能体开发平台
- OmAgent Studio智能体构建平台

智能助手

国外

- Copilot
- Gemini
- Meta AI
- Siri
- Alexa
- Bixby
- ChatGPT
- Claude
- Grok

国内

- Kimi
- 文小言
- 豆包
- 通义
- 腾讯元宝
- 讯飞星火
- DeepSeek
- MiniMax
- 智谱清言
- 阶跃AI
- 百小应
- 天工
- 商量
- 问小白
- 心流AI助手
- 华知
- WPS灵犀
- 灵犀语音助手
- 小艺
- 超级小爱
- YOYO 助理
- 小布助手

资料来源：企业/产品/应用排序不分先后，至顶智库结合公开资料整理绘制。

2025全球人工智能全景图谱—硬件&具身&AI搜索&AI办公

智能硬件

AI手机

Google SAMSUNG HUAWEI vivo
oppo HONOR Lenovo 联想 ZTE中兴 MEIZU

AI眼镜

Meta RayNeo Rokid 夸克
HUAWEI MEIZU Lawaken 李未可 INMO

AI录音设备

PLAUD.AI 钉钉
出门问问 科大讯飞 IFLYTEK

AI PC

DELL Technologies ASUS Lenovo 联想 HUAWEI

智能汽车

TESLA XPENG LI AUTO BYD AITO

具身智能

Boston Dynamics | Tesla | Figure | 1X Technologies | 宇树科技 | 银河通用 | 星动纪元 | 逐际动力 | 优必选 | 傅利叶智能

AI 搜索

AI Mode | Bing | Perplexity | 秘塔AI搜索 | 知乎直答 | 百度智能框 | AMiner沉思
开搜AI | C知道 | Suppr超能文献 | SeekAll | Dia浏览器 | 博查 | Miku AI搜索

AI 办公

HelpLook | Lavender | Beautiful.ai | Figma Slides | 钉钉 | 腾讯文档 | 百度文库 | WPS AI | 360AI办公
万知 | 讯飞智文 | ChatExcel | Notion | VERSE | xmind | 亿图脑图MindMaster | GitMind
AiPPT | iSlide | 轻竹办公 | 笔格PPT | 歌者PPT | 博思AIPPT | 麦当秀 | 会译 | 灵办AI

资料来源：企业/产品/应用排序不分先后，至顶智库结合公开资料整理绘制。

2025全球人工智能全景图谱—AI写作&图像&音视频&音乐

AI 写作

HyperWrite
Sudowrite
LongShot
ProWritingAid
讯飞写作
火山写作
彩云小梦
新华妙笔AI写作
树熊AI写作
光速写作
笔灵AI
Giiso写作机器人
火龙果写作
Effie写作

AI 图像

Midjourney
figma
Picsart
ZMO.AI
Artbreeder
Freepik
Canva可画
稿定AI
墨刀
美图设计室
MasterGo
懒设计
万相营造
象寄
妙多AI
百度网盘AI修图
标小智
Pixso AI
数画
创客贴
豆绘AI
美间
JsDesign
易可图
千鹿AI
造物云AI
Seede AI
造梦日记
360鸿图
像素蛋糕
咻图AI

AI 视频

runway
Pika
Modify Video
HeyGen
synthesia
descript
VEED
loom
Clipfly
Boolvideo
快手可灵
即梦AI
腾讯智影
Vidu
海螺AI
度加创作工具
智谱清影
智小象
模力视频
MOKI

AI 音乐

X studio
ACE studio
TME studio
Mureka
海绵音乐
反谱
网易天音
歌歌AI写歌
和弦派

AI 音频

ElevenLabs
RESEMBLE.AI
WellSaid
MURF.AI
PlayAI
有道文档FM
米可智能
魔音工坊
蓝藻AI
悦音配音
音剪AI

资料来源：企业/产品/应用排序不分先后，至顶智库结合公开资料整理绘制。

2025全球人工智能全景图谱—AI+营销&医疗&金融&教育

AI+营销

OpenAI, salesforce, copy.ai, klaviyo, HubSpot, jasper, albert, PERSADO, Baidu 百度, Tencent 腾讯, 京东, 明略科技, 科大讯飞 iFLYTEK, DATASTORY 数说故事, donson 东信, 小酷科技 www.ixiaoku.com, 原圈科技 Circles Technology, 深演智能 Deep Zero, 帷幄, 观远数据, eclicktech 易点天下, 数云 shuyun

AI+医疗

Google, Meta, OpenAI, VARIATIONAL AI, GenHealth.ai, Baidu 百度, 腾讯觅影, 4Paradigm 第四范式, 百川智能 BAICHUAN AI, 左手医生, 云知声 Unisound, 商汤 sensetime, 推想医疗, 微脉 WeiMai, 医联

AI+金融

Google, Meta, OpenAI, ANTHROPIC, MISTRAL AI_, cohere, Baidu 百度, 蚂蚁集团 ANT GROUP, 华为云 HUAWEI, 拓尔思 TRS, 东方财富, 达观数据 DATA GRAND, 容联云 CLOOPEN, MEMECT 文因互联, HUNDSUN, TRANSWARP 星环科技

AI+教育

Google, Meta, duolingo, grammarly, BYJU'S The Learning App, Quizlet, KHANMIGO by Khan Academy, CENTURY ACTIONABLE INTELLIGENCE, Baidu 百度, TAL 好未来, 有道 youdao, 猿辅导, 作业帮, 松鼠Ai

资料来源：企业/产品/应用排序不分先后，至顶智库结合公开资料整理绘制。

2025全球人工智能全景图谱—多模态模型

多模态基础模型

- GPT-5
- Gemini 2.5 Pro
- Llama 4
- Grok 4
- Mistral Medium 3
  Pixtral Large
  Mistral Small 3.1
- 文心4.5系列
- 豆包1.6
- Hunyuan-Large-Vision
- CoGenAV
- MiniMax-VL-01
- MiniCPM-V 4.5
- SenseNova V6.5
- Skywork UniPic 2.0
- GLM-4V-Flash
- 阶跃AI Step 3

图像模型

- DALL·E 3
- Gemini 2.5 Flash Image
  (Nano Banana)
- Stable Diffusion 3.5
- Aurora
- Seedream 4.0
- HunyuanImage 2.1
- CogView-4

视频模型

- Sora
- Veo 3
- Gen-4
- Wan2.2
- 豆包·视频生成模型
  Seedance 1.0 pro
- 混元视频生成模型
- 可灵2.0视频生成模型
- Hailuo 02
- Vidu 2.0

语音模型

- MuseNet
- Lyria 2
- Stable Audio 2.0
- 豆包·实时语音模型
- MiniMax Speech 2.5
- GLM-ASR
- Mureka O1
  Mureka V7.5

资料来源：主要体现2024年10月以来推出的各类模型，至顶智库结合公开资料整理绘制。

2025全球人工智能全景图谱—推理模型&开源模型

推理模型

GPT-5 Thinking	Gemini 2.5 Pro	Claude Opus 4.1 Claude Sonnet 4	Llama 4
Grok 4	Magistral	Phi-4-reasoning	LFM2 LFM-7B
文心X1.1 ERNIE-4.5-21B-A3B-Thinking	Qwen3 QwQ-32B	Hunyuan-A13B	doubao-seed-1.6-thinking
Deepseek-V3.1 Deepseek-R1	讯飞星火X1	日日新SenseNova V6	Kimi K2
MiniMax-M1	GLM-4.5	Skywork-OR1	阶跃AI Step 3

开源模型

GPT-OSS	Gemma 3	Llama 4	Mistral Small 3.1 Pixtral Large
文心4.5系列 ERNIE-4.5-21B-A3B-Thinking	Hunyuan-A13B Hunyuan-MT-7B	Qwen3 QwQ-32B Qwen2.5-Omni-7B	Deepseek-V3.1 Deepseek-R1
MiniMax-M1	GLM-4.5V CogVideoX v1.5 CogAgent-9B	Skywork-OR1 Skywork UniPic 2.0 Matrix-Game 2.0 Matrix-3D	Kimi K2

资料来源：主要体现2024年10月以来推出的各类模型，至顶智库结合公开资料整理绘制。

2025全球人工智能全景图谱—AI基础设施

基础设施层

AI芯片

- NVIDIA
- AMD
- Google Cloud
- aws
- Meta
- intel
- Qualcomm
- HUAWEI
- 昆仑芯 KUNLUNXIN
- Cambricon 寒武纪
- METAX 沐曦
- 壁仞科技 BIREN TECHNOLOGY
- Enflame 燧原科技

AI服务器

- DELL Technologies
- Hewlett Packard Enterprise
- inspur 浪潮
- H3C
- Lenovo 联想

AI计算集群

- Google Cloud
- aws
- Microsoft Azure
- Meta
- 阿里云
- 百度智能云
- 腾讯云
- 华为云 HUAWEI
- 商汤 sensetime

开发平台

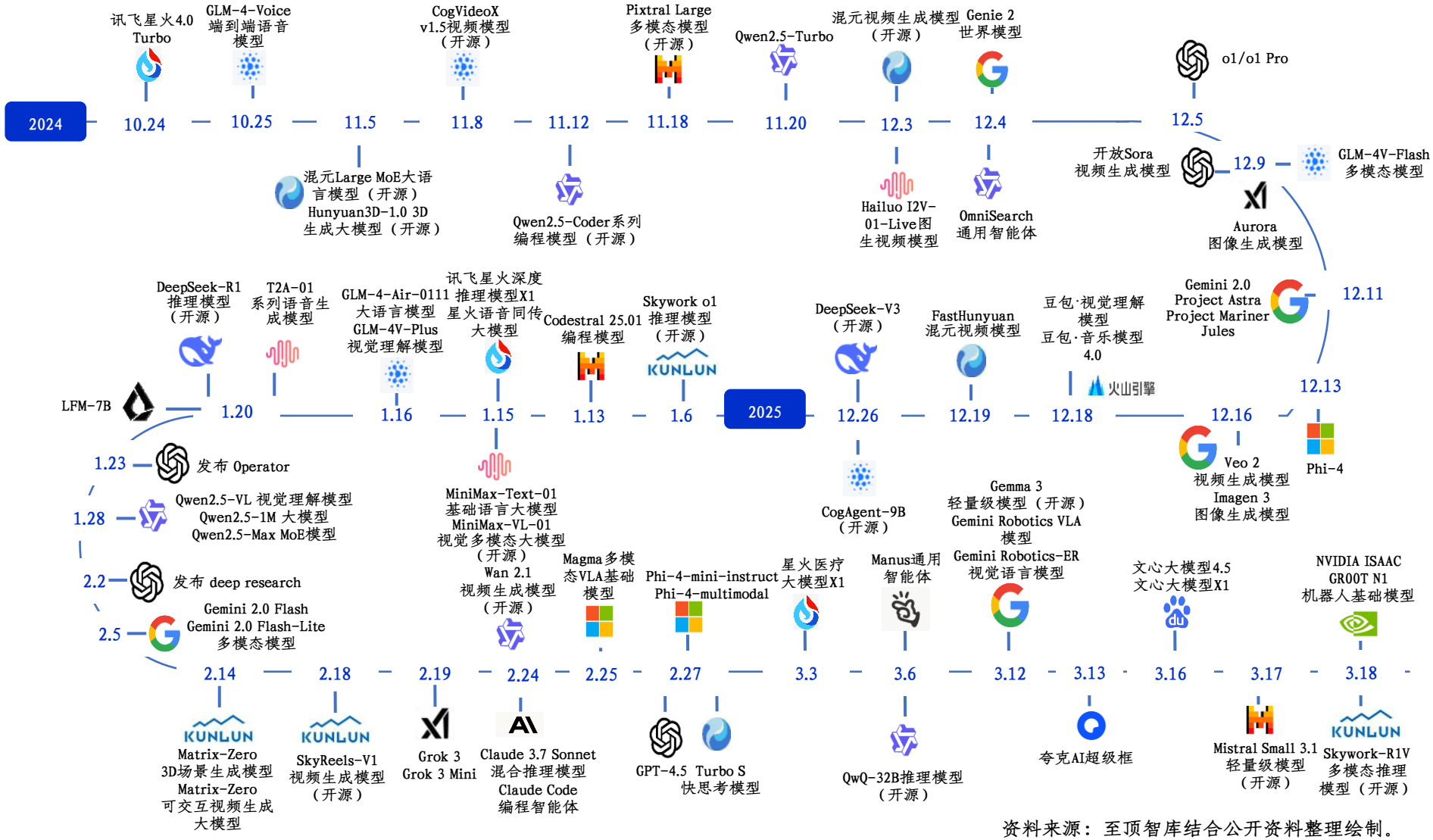
- Google Cloud
- aws
- Microsoft Azure
- 阿里云
- 百度智能云
- 腾讯云
- 火山引擎
- 华为云 HUAWEI
- 商汤 sensetime
- 智谱·AI

数据服务

- Appen
- scale
- ALEGION by SanctifAI
- sama
- gretel
- 海天瑞声
- 数据堂 Datatang
- 标贝科技 DataBaker Technology
- STARDUST
- BAYOU

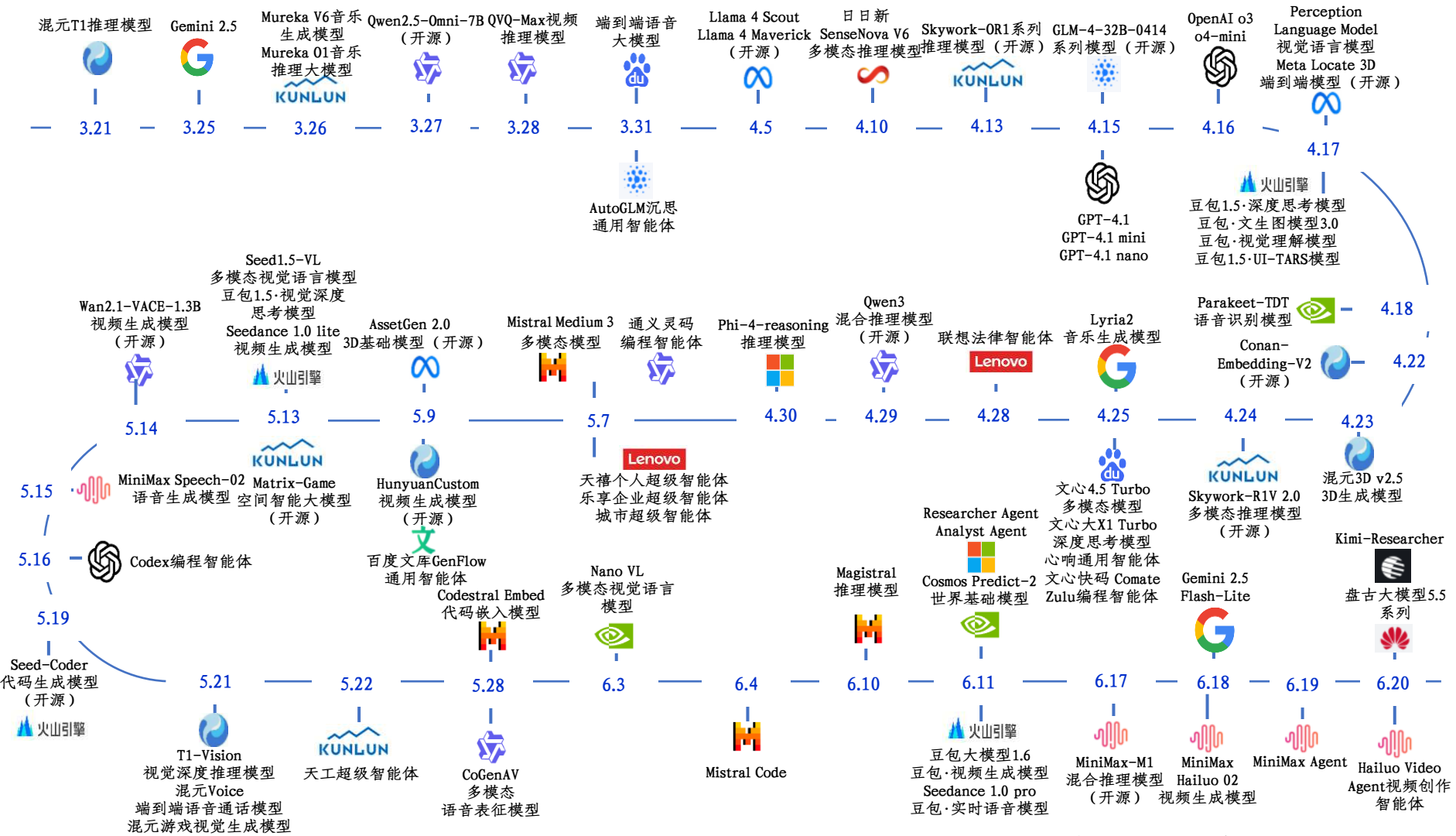
资料来源：企业/产品/应用排序不分先后，至顶智库结合公开资料整理绘制。

1.5 全球人工智能产业发展路线图（2024-2025）



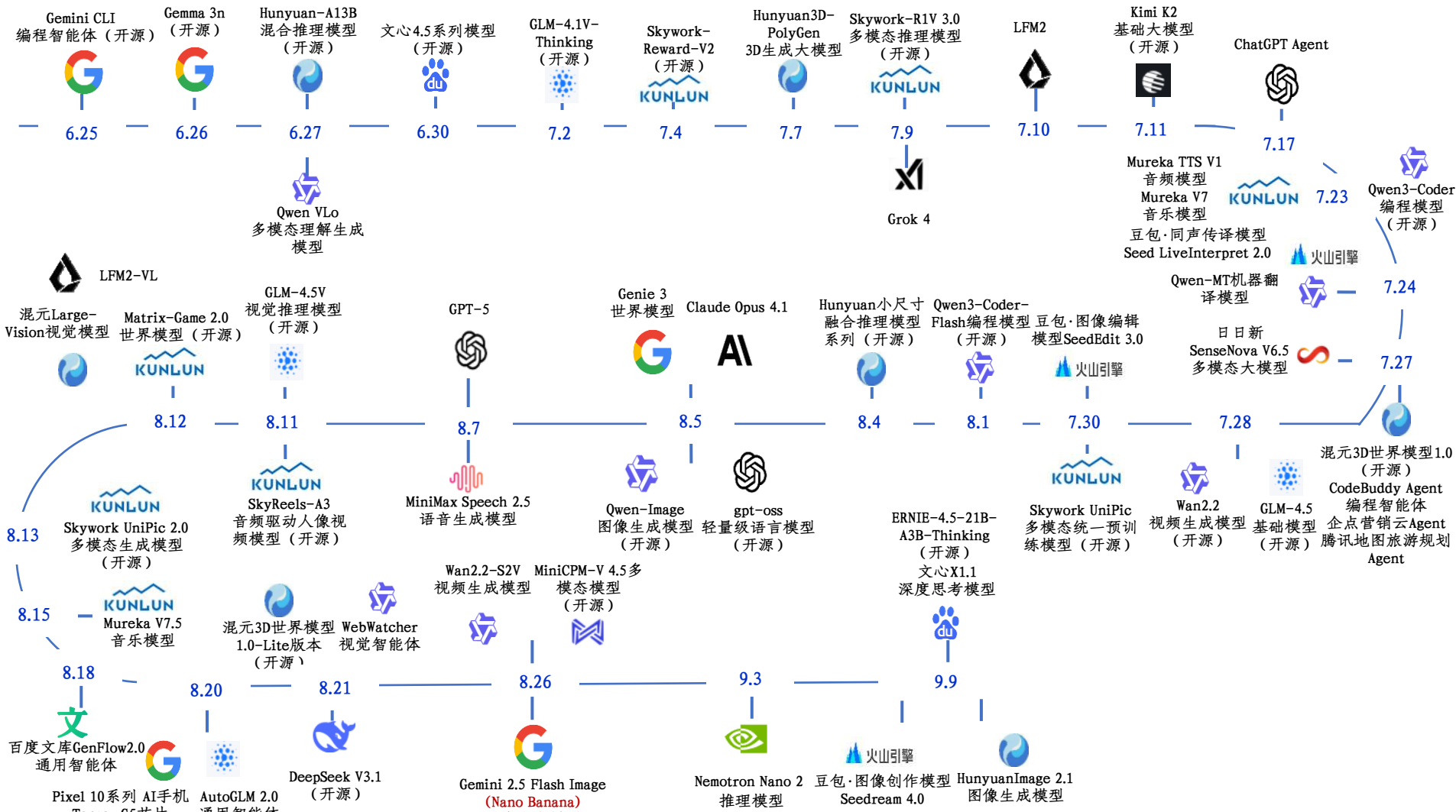
资料来源：至顶智库结合公开资料整理绘制。

1.5 全球人工智能产业发展路线图（2024-2025）



资料来源：至顶智库结合公开资料整理绘制。

1.5 全球人工智能产业发展路线图（2024-2025）



资料来源：至顶智库结合公开资料整理绘制。

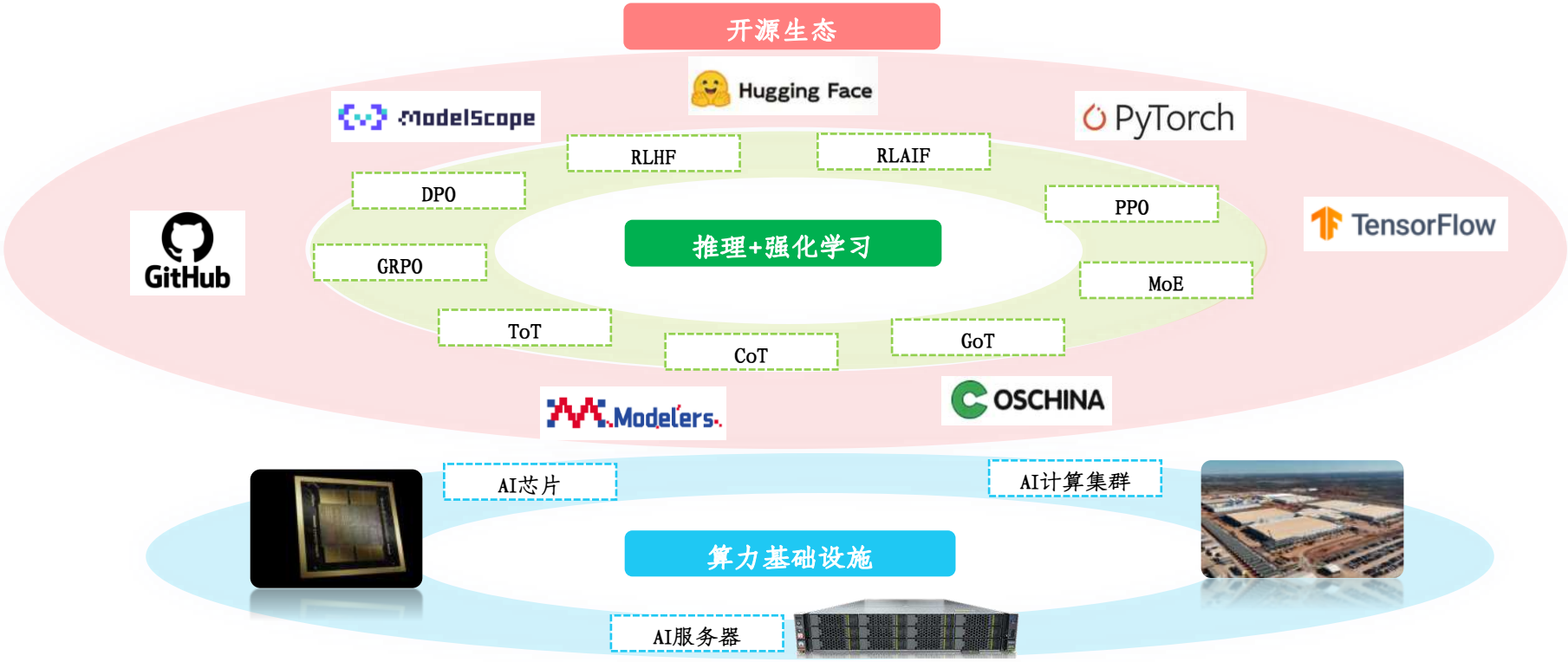
2. 迈向AGI的关键领域

推理模型　　强化学习

算力基建　　开源生态

2.1 驱动AGI发展关键因素：推理+强化学习+算力基建+开源生态

通用人工智能的不断演进主要由四大因素驱动。一是模型推理能力的不断跃升。DeepSeek相关模型在性能上达到全球开源模型的顶尖水平。**二是强化学习的不断迭代。**如GRPO等新型算法通过组内相对奖励机制，解决传统PPO算法对价值函数的依赖问题。**三是算力基建的持续投入。**美国“星际之门”计划未来四年投资5000亿美元构建AI基础设施，为大规模AI模型训练和推理提供算力基础。**四是开源生态的广泛构建与繁荣共享。**Hugging Face汇聚超6000个可部署开源模型，推动前沿技术快速转化为生产力，构建全球协作的创新网络。

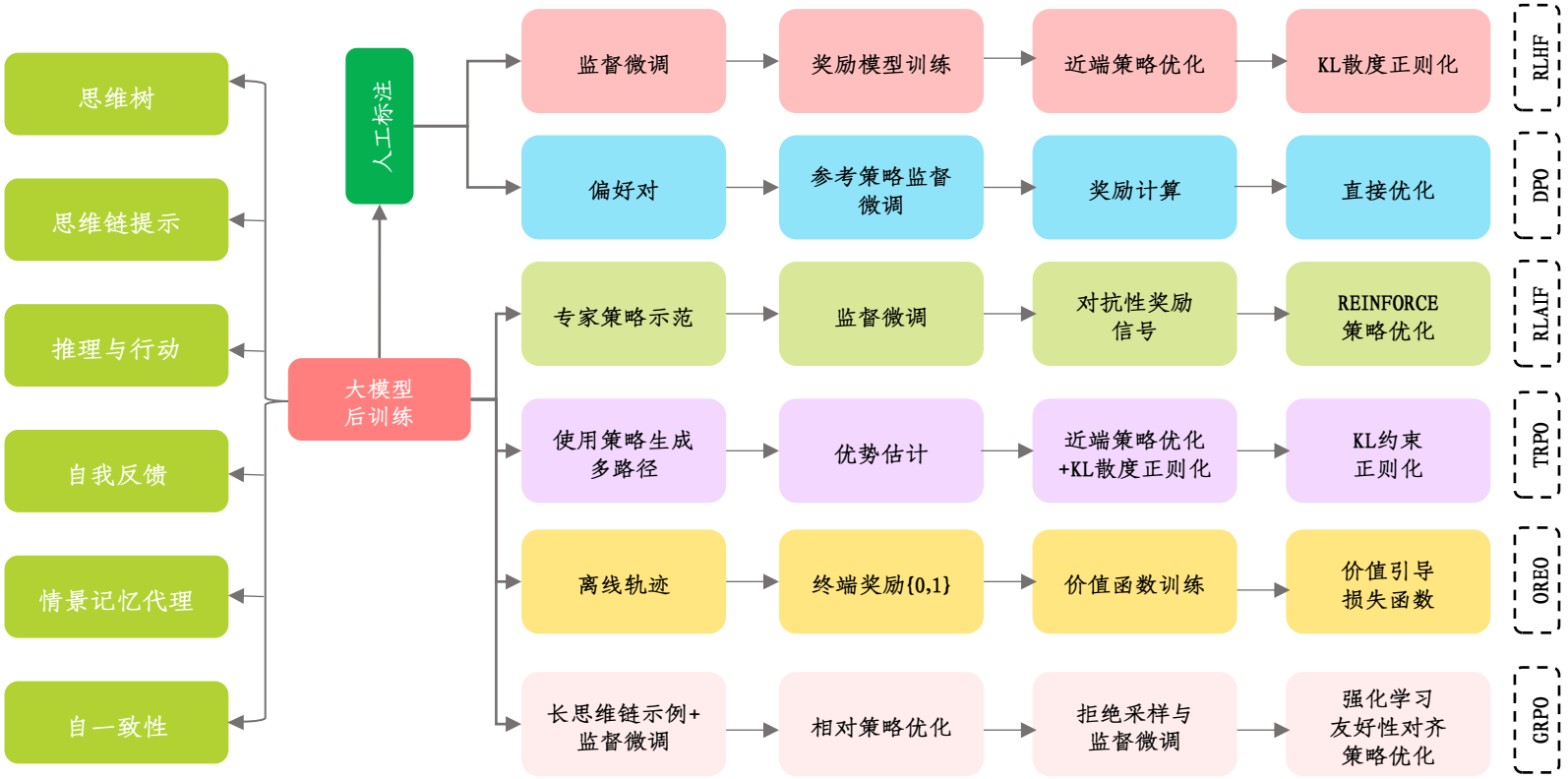


资料来源：至顶智库结合公开资料整理绘制。

2.2 推理路径：通过各类算法机制提升模型推理能力

大模型基于一套系统化技术路径来提升其推理能力。以思维链（CoT）为基础实现分步推理，通过自我反馈和情境记忆形成动态调整机制，并借助自一致性校验确保逻辑一致性。在训练层面，采用监督微调（SFT）与基于人类反馈强化学习（RLHF）相结合的方式，借助优势估计和终端奖励完成策略更新。同时，引入对抗性奖励信号、KL 惩罚以及价值函数训练以实现策略优化的目标。

推理时间推理（Inference-time Reasoning）

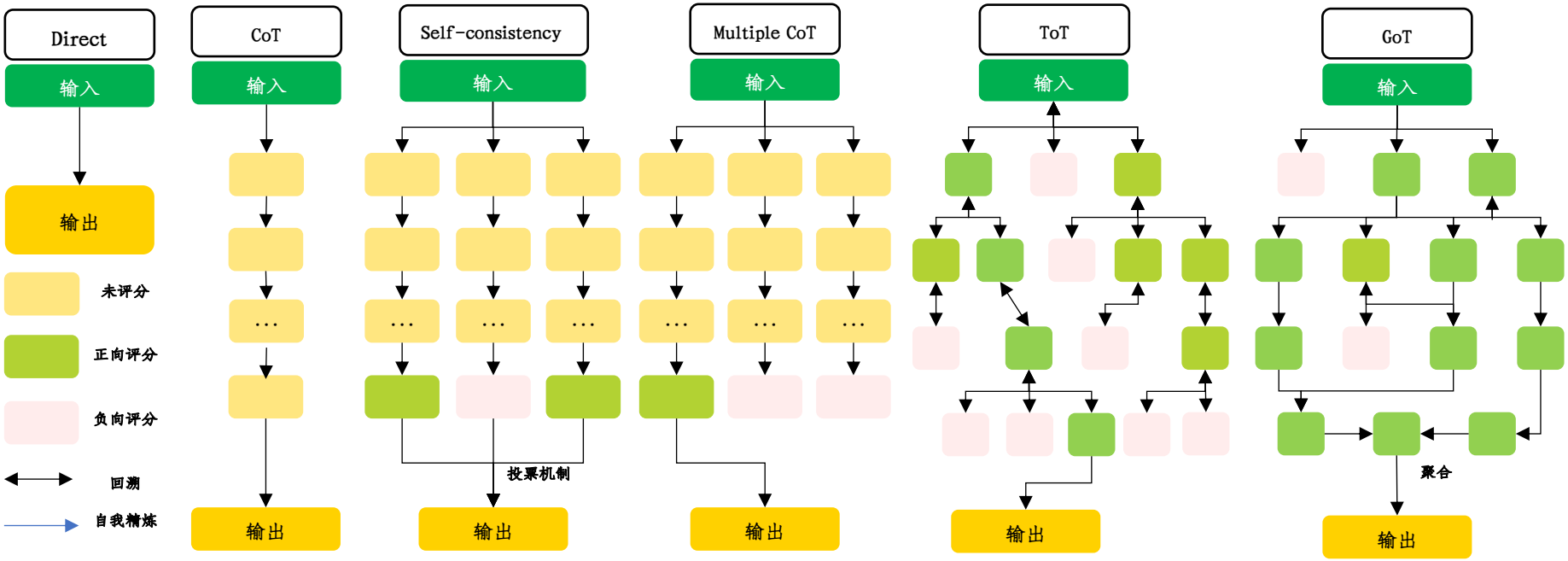


资料来源： *LLM Post-Training: A Deep Dive into Reasoning Large Language Models*，至顶智库整理绘制。

2.2 推理路径：多种范式增强模型应对复杂任务的能力

大模型通过多种推理路径适应不同任务需求。基础范式包括直接输出（Direct）和思维链（CoT）推理。自一致性（Self-consistency）与多重思维链（Multiple CoT）通过生成多条推理路径并采用投票机制，为不确定性任务进行方案择优。面对需要多路径探索的复杂任务，思维树（ToT）引入树状结构，支持分支评估与回溯机制，实现不同路径之间的探索。最新提出的思维图（GoT）则突破树状结构的限制，利用图结构实现路径间的动态聚合与信息重组，为更复杂的非线性推理问题提供更优的解决思路。

大模型推理路径

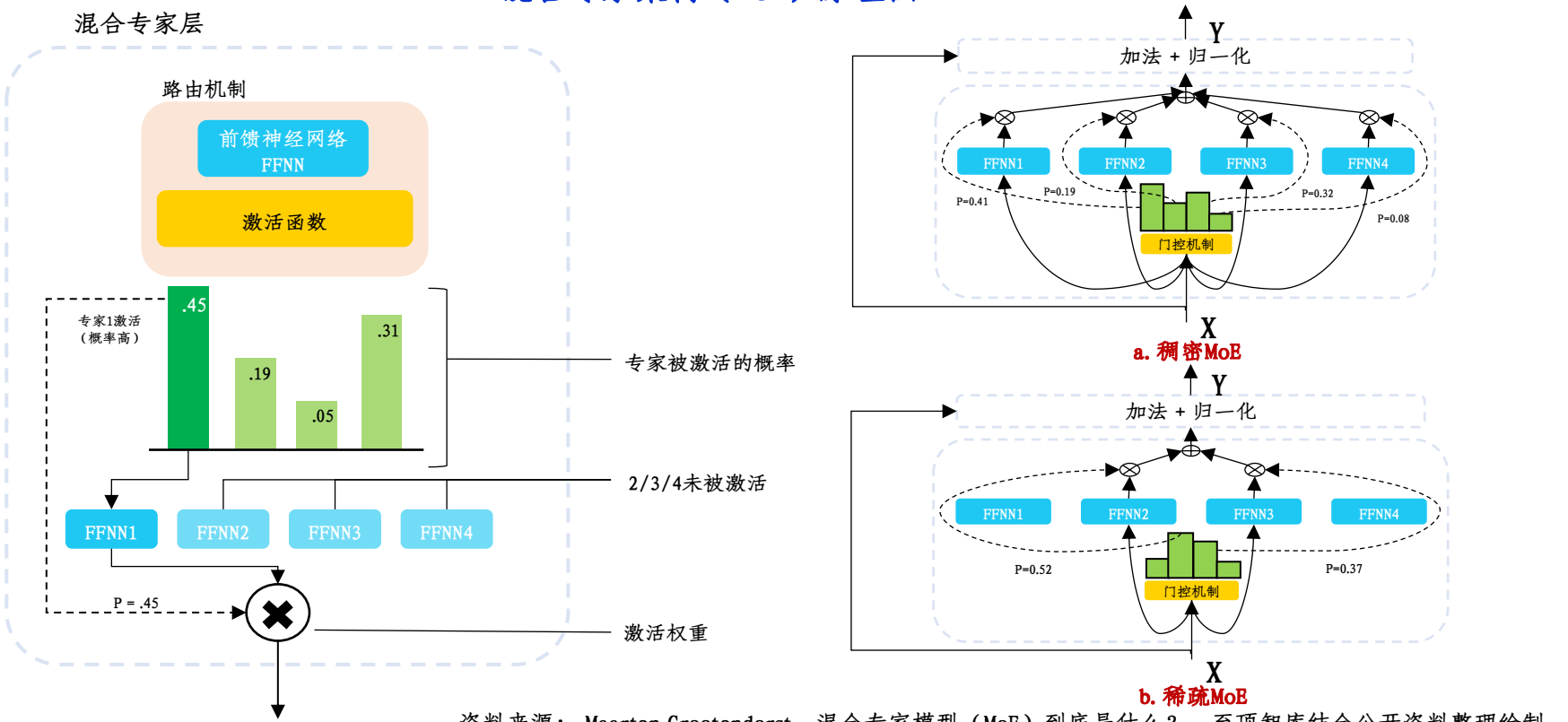


资料来源： *LLM Post-Training: A Deep Dive into Reasoning Large Language Models* ，至顶智库整理绘制。

2.3 混合专家架构（MoE）：推动推理效率与模型性能突破

MoE（Mixture of Experts）架构中，输入数据通过前馈神经网络（Feed-Forward Neural Network，FFNN）与激活函数处理，再由门控机制为每个专家分配激活概率。在稠密MoE架构中，所有专家均被激活参与计算，最终输出为各专家结果的加权和；在稀疏MoE架构中，仅激活其中若干专家（如图中激活FFNN1），以提高推理效率并降低计算资源开销。该机制实现在保持模型性能的同时，优化推理效率，适用于大规模参数部署。近年来，大模型已引入MoE架构以提升参数利用率和训练扩展性。

混合专家架构（MoE）原理图

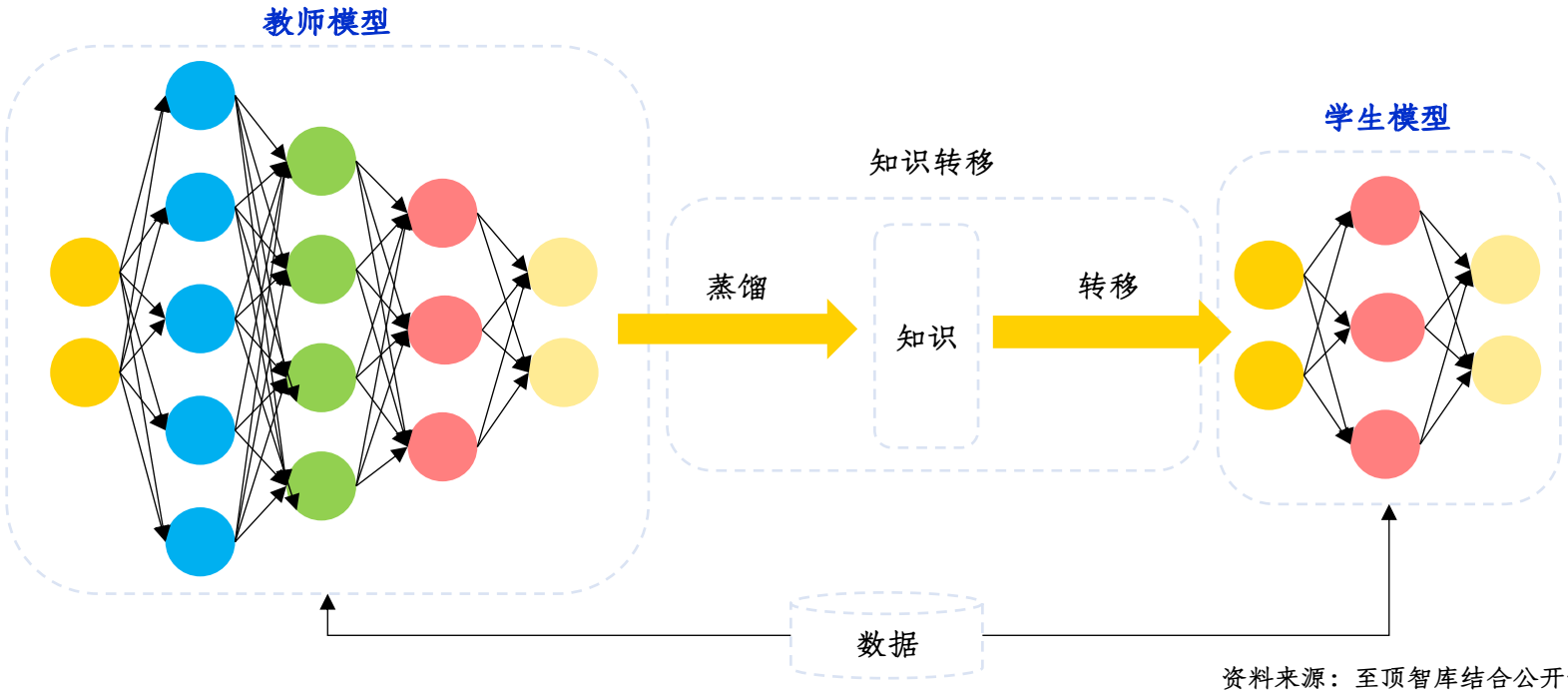


资料来源： Maarten Grootendorst，混合专家模型（MoE）到底是什么？，至顶智库结合公开资料整理绘制。

2.4 模型蒸馏：压缩计算复杂度，实现模型高性能和轻量化部署

模型蒸馏是指大型复杂模型（教师模型）的知识迁移到小型高效模型（学生模型）的技术，其核心目标是在保持模型性能的同时，显著降低模型的计算复杂度和存储需求，使模型更适合在资源受限的环境中部署。教师模型规模庞大，性能较高，但在计算和存储资源上存在较大压力；学生模型结构较为简单，通过模仿教师模型输出的软标签（概率分布）学习其知识和表示能力，软标签包含类别间相似性和内在关系的更多信息，有助于学生模型捕捉潜在特征并提升泛化能力。在实际应用中，DeepSeek-R1对Qwen和Llama等开源模型进行蒸馏，得到更高效的小模型，显著降低推理成本。此外，诸如DistilBERT、TinyBERT以及MobileBERT模型也都采用蒸馏技术，用以在保持性能的同时提升效率，推动轻量化模型的发展和普及。

模型蒸馏的原理与机制

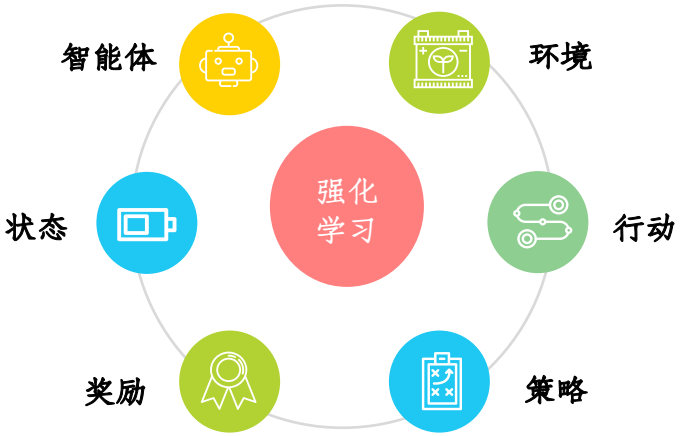


资料来源：至顶智库结合公开资料整理绘制。

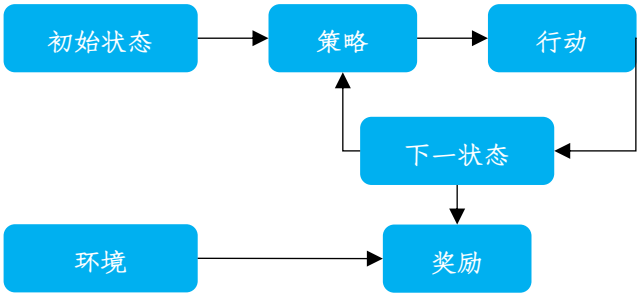
2.5 主流模型的后训练过程已普遍采用强化学习方法

当前，主流大模型利用强化学习技术进一步提效。强化学习作为机器学习领域的核心技术之一，由智能体、环境、状态、行动、奖励及策略六大核心部分组成。与监督学习和无监督学习不同，强化学习是指智能体通过执行动作来影响环境，并根据环境反馈的奖励来调整策略，以便在未来做出更好的决策。目前主流模型的后训练过程均已采用相关强化学习方法进行优化。

强化学习核心要素



强化学习通用流程



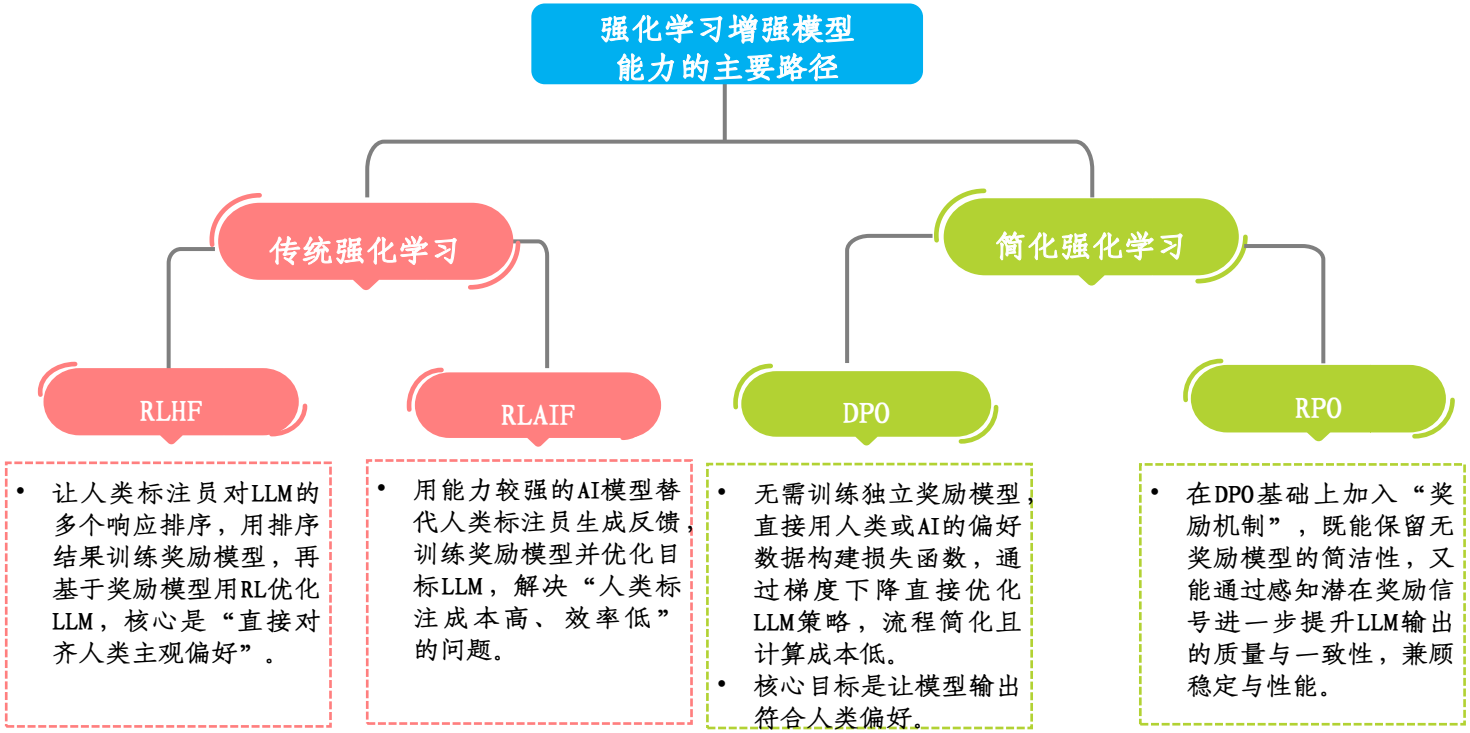
主流模型后训练的强化学习方法

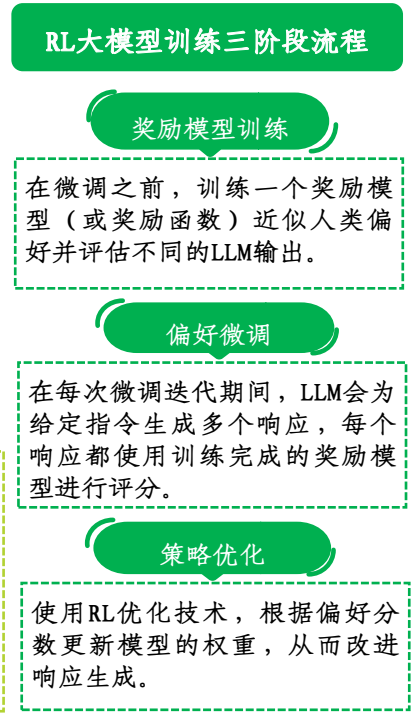
模型	企业	强化学习方法
DeepSeek-R1	deepseek	RL through CoT
Kimi-k1.5	Moonshot AI	RL through CoT
o1	OpenAI	RL througt CoT
Hermes 3	NOUS RESEARCH	DPO
Athene-70B	Nexusflow	RLHF
Starling-7B	Berkeley UNIVERSITY OF CALIFORNIA	RLAIF,PPO
Gemma2	Google	RLHF
Qwen2	Alibaba	DPO
Llama 3	Meta	DPO
Nemotron-4 340B	NVIDIA	DPO,RPO
ChatGLM	智谱·AI	ChatGLM-RLHF
DeepSeek-V2	deepseek	GRPO

资料来源：*Reinforcement Learning Enhanced LLMs: A Survey*，至顶智库整理绘制。

2.5 传统与简化强化学习成为大模型提效“双涡轮”

主流大模型在后训练阶段采用的强化学习方法主要包含两类。一类是传统强化学习方法如人类反馈强化学习（RLHF）和AI反馈强化学习（RLAIF）；另一类是简化强化学习方法如直接偏好优化（DPO）和奖励偏好优化（RPO）。强化学习在大模型训练中经历三阶段流程，即奖励模型训练、偏好微调和策略优化。借助上述方法，大模型可突破单一预设答案的局限，动态适配不同偏好，生成结构合理、契合上下文且更具创造性与高质量的内容，更加贴合用户期望。





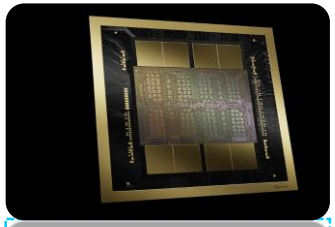
资料来源：*Reinforcement Learning Enhanced LLMs: A Survey*，至顶智库整理绘制。

2.6 海外科技巨头引领高性能AI芯片发展

近年来，AI芯片已成为驱动人工智能发展的核心引擎，尤其在大模型训练和推理中，算力、内存带宽和互联技术直接决定模型迭代更新速度。当前，国际主流芯片厂商正围绕高性能计算、低精度格式和系统级优化展开激烈竞争，推动AI芯片向更高性能演进。NVIDIA凭借其Blackwell架构与Rubin架构持续领跑，保持其在高端训练和推理芯片市场的领导地位；Google依托自研TPU深化软硬件垂直整合，强化其云计算和AI服务的底层能力；AWS通过自研Trainium训练芯片与Inferentia推理芯片的协同部署，提供云端算力解决方案。

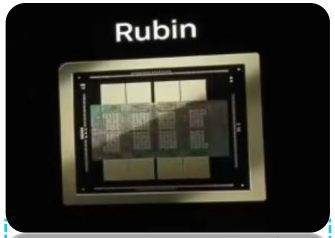
全球主流AI芯片示例

NVIDIA Blackwell



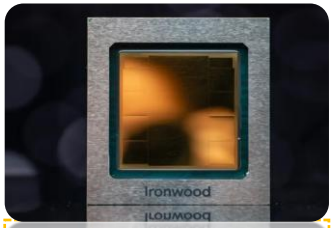
- 2024年，NVIDIA推出Blackwell架构；
- 适用于大规模推理AI场景，采用4NP工艺制程，具有2080亿个晶体管，能效较上一代Hopper GPU实现新突破，支持实时性能下的高吞吐量；
- 基于Blackwell Ultra的GB300系统计算能力是Hopper系统的65倍，性能大幅提升，推动实现收益最大化。

NVIDIA Rubin



- 2025年，NVIDIA公布其下一代超级芯片Vera Rubin；
- 整体性能是GB300 NVL72的3.3倍，集288GB HBM4、13TB/s带宽、260TB/s吞吐量于一身，实现900倍于Hopper的性能；
- Rubin Ultra计算面积翻倍，密集FP4浮点运算性能提升至100 PFLOPs。HBM容量达到1024GB。

Google Ironwood

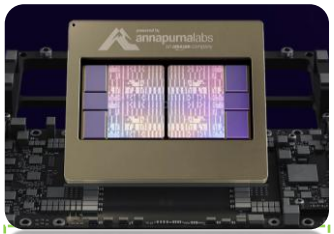


- 2025年，Google首次亮相第七代TPU Ironwood；
- 算力上每个单独的芯片峰值可达4614 TFLOPs；相较于第六代TPU Trillium，Ironwood在功耗效率上实现2倍提升，比首款Cloud TPU高出近30倍；
- Google迄今为止性能最强、可扩展性最高的定制AI加速器，首款专为推理设计的加速器。

AWS Inferentia 2

- 2022年，AWS发布第二代推理芯片Inferentia 2；
- 架构与Trainium 1相似，但NeuronLink-v2互连端口更少。每瓦性能最多可提升50%，将成本有效降低40%。与一代相比，吞吐量提高4倍，延迟大幅降低；
- 基于Inferentia 2的Amazon EC2 Inf2可以大规模部署复杂模型。

AWS Trainium 2



- 2023年，AWS推出第二代训练芯片Trainium 2；
- 专为以高性能训练大参数基础模型和大语言模型构建。性能是第一代的4倍，内存提升3倍；
- 基于Trainium 2的Amazon EC2 Trn2专为生成式人工智能构建，单独实例包含16个Trainium加速芯片，用于训练部署数千亿至数万亿参数的模型。

资料来源：至顶智库结合公开资料整理绘制。

2.7 国内AI芯片架构持续创新，训练推理两线并进

当前，国内AI芯片正依托国产化战略快速崛起，以华为昇腾910C、昆仑芯P800、沐曦曦云C600等为代表，在推理和轻量化训练场景中率先实现规模化落地。与国外追求绝对算力峰值不同，国内企业更注重架构自主与性能优化，并通过软硬件垂直整合和性价比优势抢占市场。总体而言，国内AI芯片尽管受到外部环境制约，但目前已取得显著进展。长远来看，先进制程、软件生态、硬件稳定性以及基础架构原创性仍将是未来需要持续攻坚的重点领域。

国内主流AI芯片示例

昇腾910C

- 2024年，华为推出其最新一代芯片昇腾910C；
- 制程上910C采用中芯国际7nm（N+2）制程工艺。比起同类产品以更低成本实现更高良率、更快的量产速度；
- 2025年，华为在910C的基础上推出CloudMatrix384超节点，实现从服务器级到矩阵级的资源供给模式转变，通过全面的架构创新，在多方面全面领先。

P800

- 2025年，昆仑芯发布第三代算力产品P800；
- 显存规格优于同类主流GPU20%～50%，对MoE架构更加友好，率先支持8bit推理，单机8卡即可运行671B模型；可较好支撑DeepSeek系列MoE模型大规模训练任务；
- 采用完全由昆仑芯自研的XPU-P架构，可以实现对训练和推理任务的快速适配。

曦云C600

- 2025年，沐曦发布新一代通用GPU产品—曦云C600；
- 基于沐曦自主知识产权核心GPU IP，实现从芯片设计、制造到封装测试的全流程国产供应链闭环，集成大容量存储与多精度混合算力；
- 满足下一代生成式AI的训练和推理需求，性能强劲，全面对标国际旗舰GPU产品。

MTT S4000

- 2023年，摩尔线程发布MTT S4000；
- 基于全功能GPU架构、专为大模型打造的训推一体通用计算卡，支持FP64/FP32/TF32/FP16等完整计算精度，配备48GB高性能显存；
- 可实现多卡互联及千卡集群部署，为千亿参数大模型的训练、微调和推理提供强劲算力支撑，兼具先进的图频处理能力。

L600

- 2025年，燧原推出第四代训推一体芯片L600；
- 面向训练及推理场景，国内首创原生FP8低精度算力，拥有144GB存储容量、3.6TB/s存储带宽、800GB/s互联带宽；
- 基于L600推出全新计算系统—云燧OGX系列。其中OGX400实现单机八卡OAM全互联，拥有1152GB单机存储容量、28.8TB/s单机存储带宽、2.8TB/s单机聚合带宽，单机支持DeepSeek满血版。

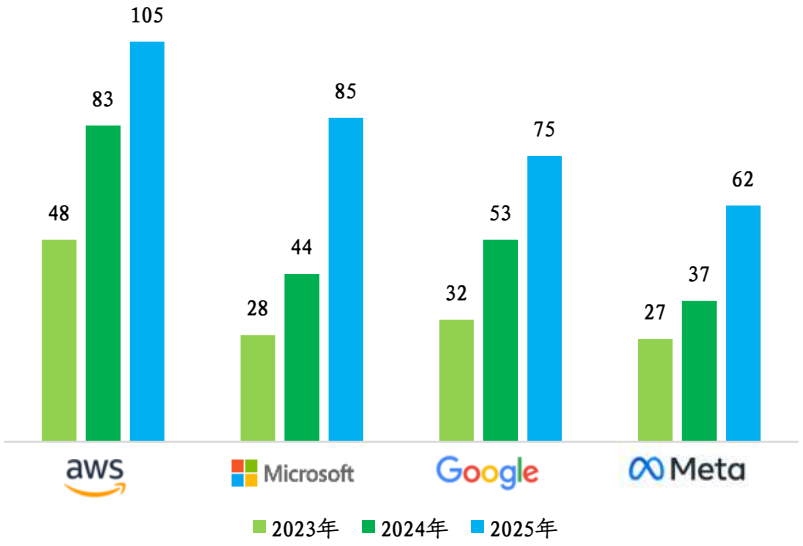
资料来源：至顶智库结合公开资料整理绘制。

2.8 海内外云厂商持续加大AI基建投资力度

近年来，国内外主要云厂商的资本开支呈现出持续攀升态势。放眼海外，美国头部云厂商投资支出持续增加，以AWS、Microsoft、Google、Meta为主的云厂商纷纷掀起投资浪潮，AWS、Microsoft、Google、Meta2025财年资本开支预计分别达到1050亿、850亿、750亿、620亿美元；聚焦国内，AI领域快速发展持续刺激国内AI基建投资，国内头部云厂商投资持续加码，投资目标已从传统数据中心转向智算中心。阿里巴巴、腾讯、百度2024财年资本开支突破新高，分别达到848亿、768亿、81亿人民币，反映出国内AI基建资本开支进入上行周期。

国外云厂商2023-2025年资本开支情况

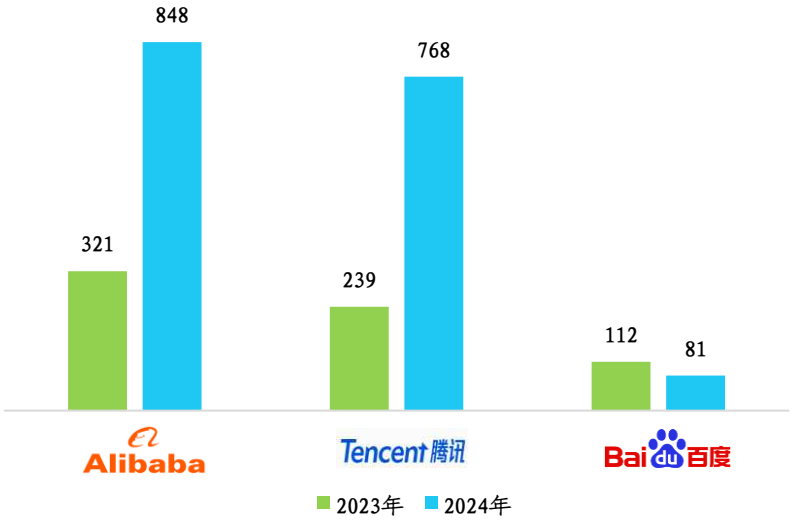
（单位：十亿美元）



数据来源：各企业财报，WSJ，2025年资本开支为中金预测，至顶智库整理绘制。

国内云厂商2023-2024年资本开支情况

（单位：亿元）



数据来源：各企业财报，WSJ，至顶智库整理绘制。

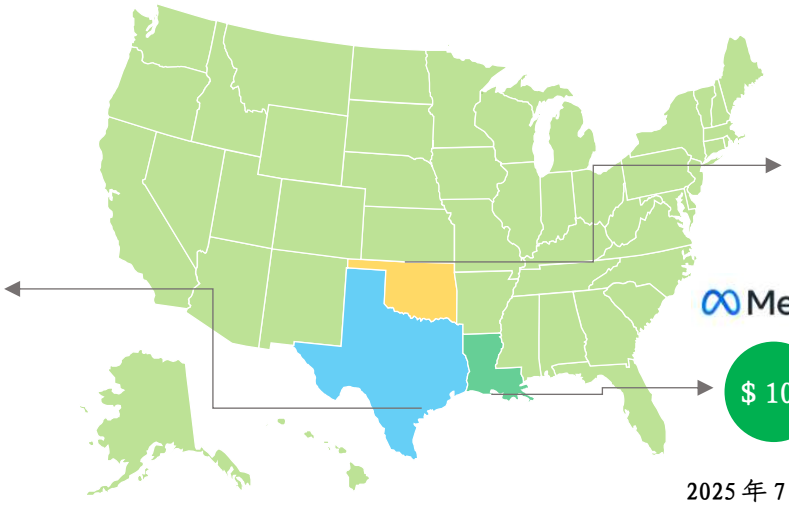
2.9 美国科技巨头持续加码，算力投资稳固攀升

美国科技巨头积极推进智算中心建设，AI算力需求不断攀升。以OpenAI、Google、Microsoft和Meta等为主的头部云厂商在AI基建领域持续加大投资力度，刷新基建投资额新高。OpenAI主导的“Stargate”项目累计投入已达1000亿美元，预计投资金额将达5000亿美元；Google上调年度资本支出至750亿美元；Meta计划向“Hyperion”集群投资100亿美元。以上数据反映出美国在模型训练与推理方面的需求持续高涨。总体来看，美国科技巨头通过大规模投资和技术创新，不断巩固其在全球人工智能领域的领先地位。此外，美国算力领域的投资参与主体日渐多元，成为助力AI基建落地的重要力量。

美国科技巨头智算中心建设情况

OpenAI SoftBank Group ORACLE MGX $ 500B

2025年1月，OpenAI宣布正式启动“Stargate”AI基建项目。该项目位于得克萨斯州，迄今累计投入已达1000亿美元，以实现AI分布式训练的目标。



Google Cloud $ 9B

2025年8月，Google宣布将在未来两年投资90亿美元，用于俄克拉何马州的云和人工智能基础设施。服务于新数据中心园区建设与现有设施扩建。

Meta $ 10B

2025年7月，Meta发布将于路易斯安那州开展的“Hyperion”集群规划，预计投资100亿美元，将进一步扩展Meta的AI基础设施，支持更复杂的模型训练和推理任务。

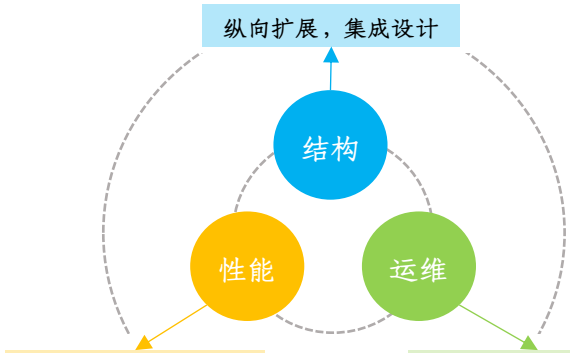
资料来源：至顶智库结合公开资料整理绘制。

2.10 国内超节点方案推动AI计算集群性能实现新突破

超节点是一种通过单节点内增加芯片数量，具备超高互联带宽、纵向扩展与集成化等优势的方案，在性能、成本、组网、运维等方面表现突出。超节点能够提供超高互联带宽与超低通信时延，有效支撑并行计算任务，缩短模型训练周期，提升整体可靠性。华为推出的Atlas 900 A3 SuperPoD（昇腾384超节点），通过总线技术实现384个NPU之间大带宽低时延互联，优化资源调度以满足AI训练与推理需求；浪潮信息发布的元脑SD200，可实现单机内运行超万亿参数大模型，并支持领先大模型机内同时运行及多智能体实时协作与按需调用；昆仑芯发布的超节点方案通过硬件创新提升全互联通信带宽，支持IB/ROE跨域低延迟传输，助力万卡级智算集群建设。超节点方案正推动AI计算集群向更高效、可靠的方向发展。

超节点特征

单节点内增加芯片数量，组装设计高度集成。



超节点案例

HUAWEI 昇腾384超节点

- 2025年7月，华为首次展出昇腾384超节点，即Atlas 900 A3 SuperPoD；
- 该产品基于超节点架构，跨节点通信带宽提升15倍，通信时延下降10倍，业界唯一突破Decode时延15ms，千亿稠密模型训练性能可达传统集群2.5倍以上；
- 通过系统工程优化，实现资源高效调度，更好满足模型训练和推理对低时延、大带宽、长稳可靠的要求。

浪潮信息 元脑SD200

- 2025年8月，面向万亿大模型训推，浪潮信息研发SD200超节点服务器，采用创新的多主机低延迟内存语义通信架构，以开放系统设计聚合64路本土GPU芯片。
- 基于融合架构2.0，采用多主机三维网格（3D Mesh）设计，实现单机64颗本土GPU芯片的高速互连。配合远端GPU虚拟映射技术，突破多主机统一编址的难题。

昆仑芯 KUNLUNXIN 昆仑芯超节点

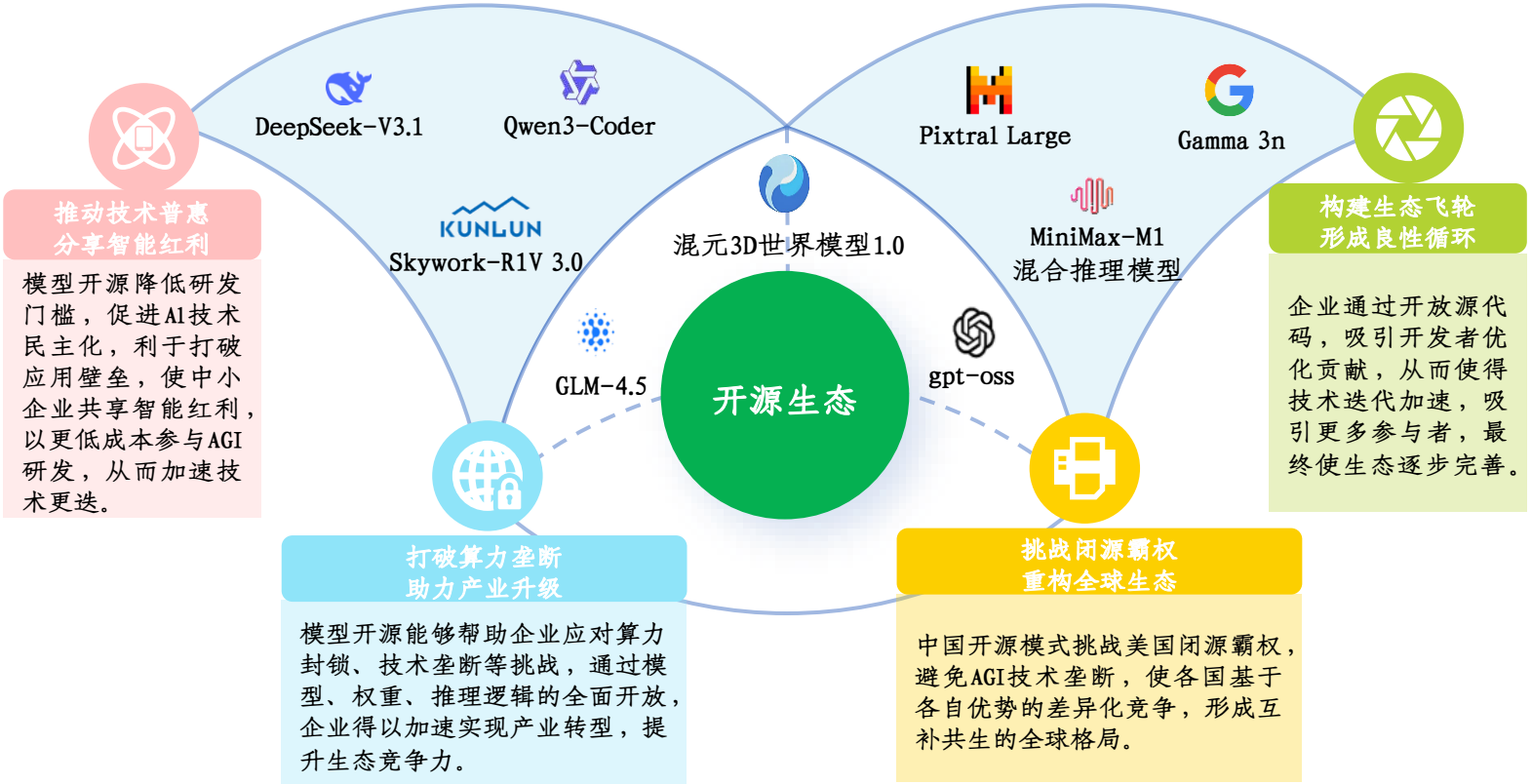
- 2025年4月，昆仑芯面推出超节点新品，为AI算力集群性能优化提升提供全栈解决方案；
- 通过硬件架构创新，该产品实现全互联通信带宽提升8倍，MoE大模型单节点训练性能提升5-10倍，单卡推理效率提升13倍；
- 支持IB/RoCE通信，实现跨柜高带宽、低延迟数据传输，支持万卡以上规模的智算集群构建。

资料来源：华为、浪潮信息、昆仑芯、至顶智库结合公开资料整理绘制。

2.11 开源生态加速AGI时代到来

近年来，开源生态成为推动AGI发展的核心引擎。中国AI企业密集开源高性能模型，如阿里通义Qwen3系列、DeepSeek-R1等。通过开放模型架构与训练框架，企业可降低算力依赖成本；开源驱动的技术民主化进程可打破闭源垄断，构建差异化生态；开源社区汇聚全球开发者协作，以“生态飞轮” 效应加速技术迭代，使AGI产业真正实现普惠共享。

开源生态对于产业发展的重要意义



资料来源：至顶智库结合公开资料整理绘制。

3.智能体技术及应用进展

智能体特征　　智能体技术架构

通用智能体　　行业智能体

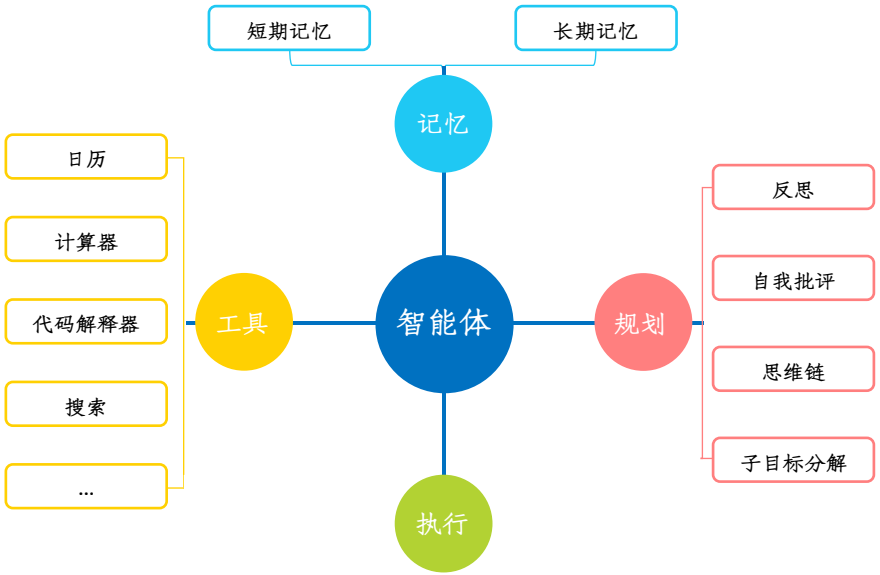
企业智能体　　MCP协议

A2A协议

3.1 智能体特征：自主感知、规划执行

智能体（AI Agent）是一种基于大模型的高自主性智能系统，可凭借强大的语言理解能力与内容生成能力实现对环境的实时感知，并通过自主规划与调用工具完成复杂目标。智能体具备记忆、自主规划、工具调用与自动执行复杂任务四大核心能力。相较于AI助理，智能体在工作目标导向、交互维度、任务处理范围、自主程度、记忆、工具使用等方面均实现突破，凭借其特有的自主感知与规划执行能力，实现从智能助理被动调用单一功能到智能体自主解决复杂任务场景的转变。

智能体核心特征



资料来源：至顶智库结合公开资料整理绘制。

智能助理与智能体对比

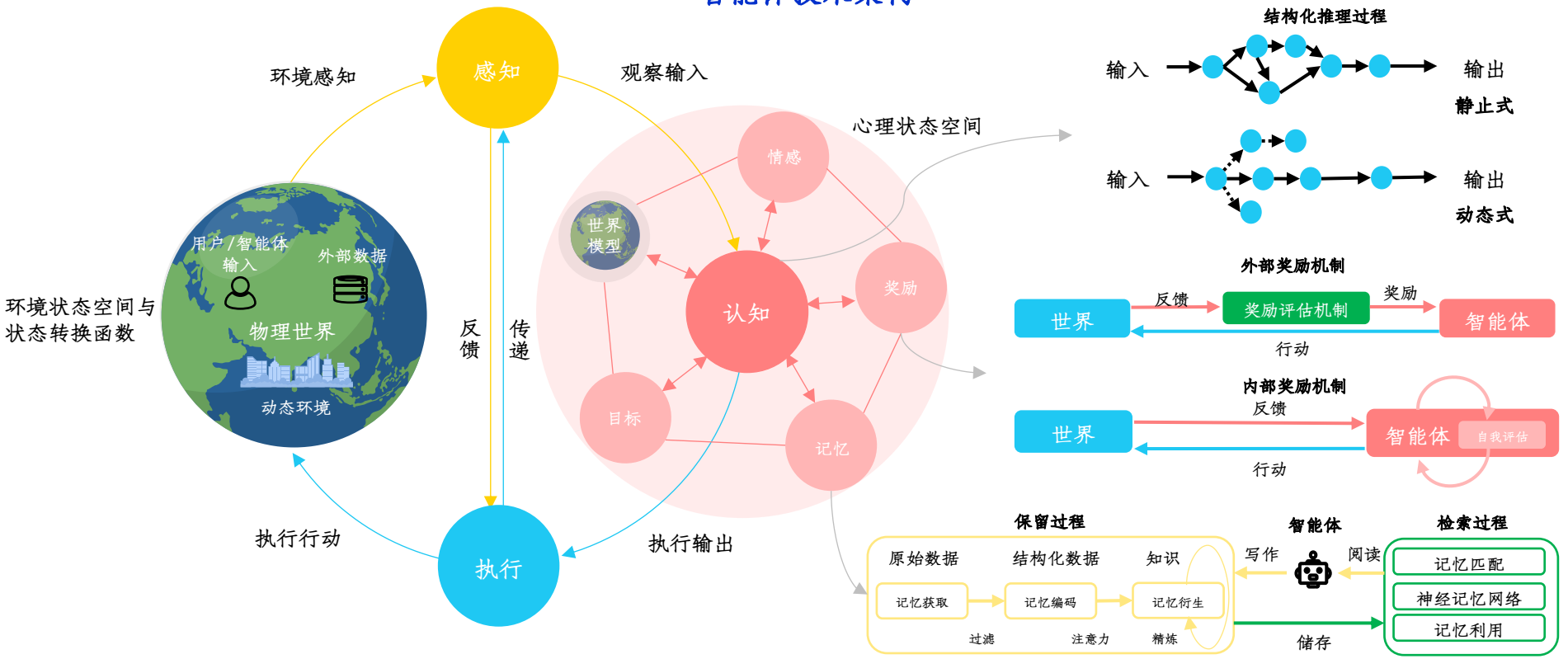
	智能助理	智能体
主要目标	完成用户指令、提高效率	自主感知、规划并达成长期目标
交互方式	多轮对话+轻量级工具调用	多模态感知+连续行动+环境交互
任务范围	相对封闭、短期任务	开放、复杂、跨系统任务链
自主性	有限自主，需用户确认	高度自主、可主动规划、反思、纠错
记忆能力	用户长期记忆（偏好、历史）	跨对话、跨任务长期记忆+经验沉淀
工具使用	调用API与本地应用	任意工具组合（浏览器、数据库、物理设备）
用户群体	个人	个人/企业/政府

资料来源：至顶智库结合公开资料整理绘制。

3.2 智能体技术架构：涵盖感知、认知与执行三大模块

智能体的技术架构主要由感知、认知与执行三大模块组成。其中，感知模块负责处理环境感知，接收用户输入、外部数据以及动态反馈等一系列信息并进行解析；认知模块分为情感、奖励、记忆、目标与世界模型五个部分，作为智能体的认知基础，与执行模块协同运转，使智能体能够完成“感知—规划—工具调用—行动—反思”的全链路自主任务流程，最后将任务结果重新输出到物理世界。

智能体技术架构

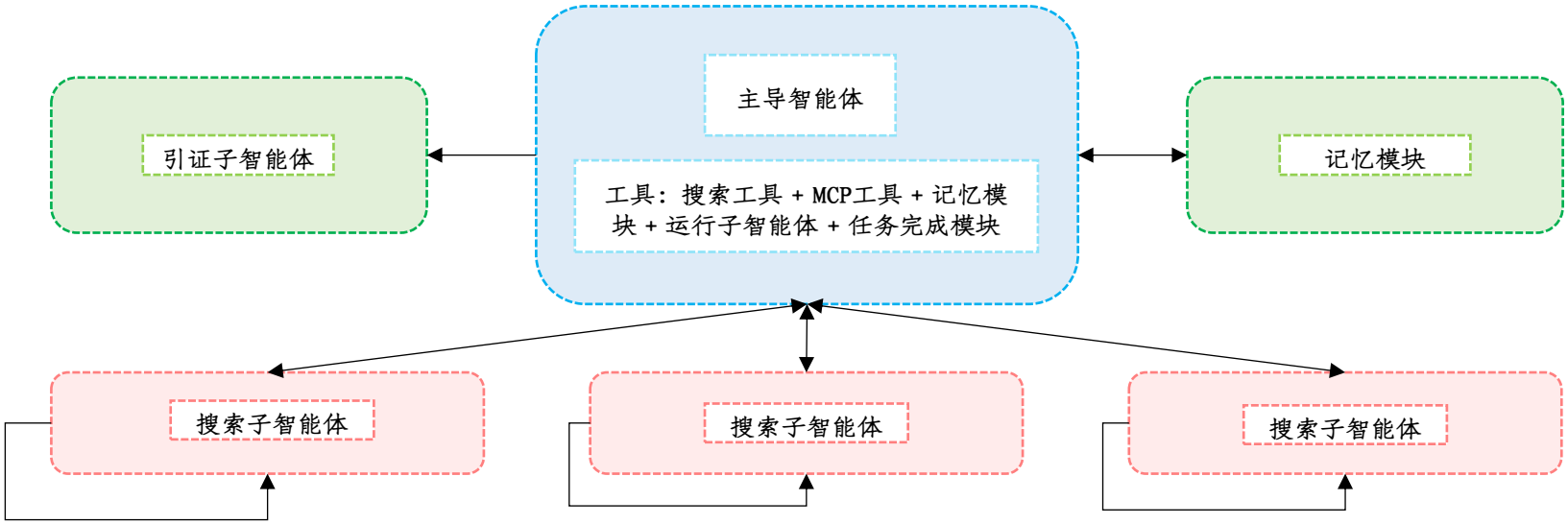


资料来源：*from Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems*，至顶智库整理绘制。

3.3 多智能体系统：实现任务的并行处理与信息整合

多智能体系统通过主导智能体协同子智能体叠加工具调用，实现任务的并行处理与信息整合。以Anthropic多智能体研究系统架构为例，以主导智能体为核心，依托搜索工具、模型调用协议工具（MCP）、记忆模块、搜索子智能体及任务完成模块协同完成研究工作。主导智能体可调用搜索子智能体执行多方向检索，并调度引证子智能体插入文献引用，以增强信息可信度。记忆模块在整个研究过程中持续存储并更新状态，保持上下文的连贯性与一致性。用户请求在系统内部分配，经由多个子智能体并行探索与引用增强后，生成最终报告并反馈给用户。

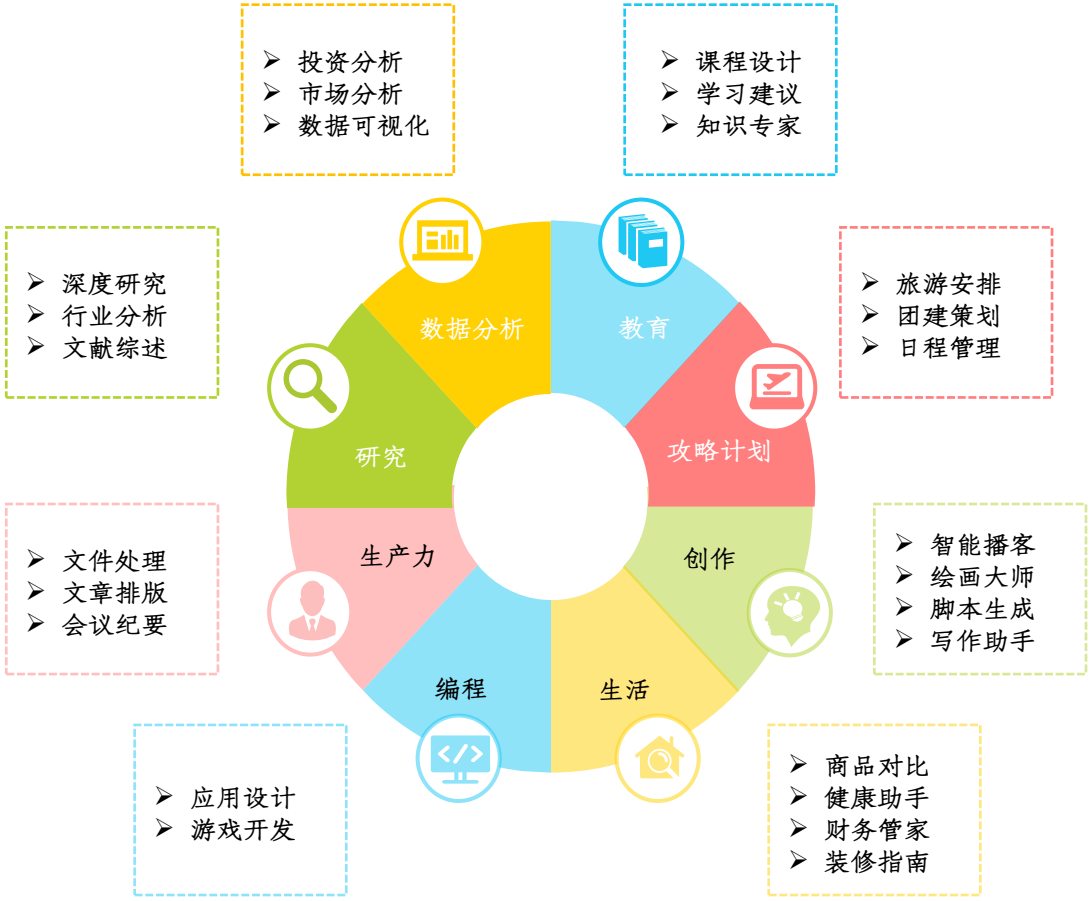
多智能体系统架构



资料来源：*How we built our multi-agent research system, Anthropic,* 至顶智库结合公开资料整理绘制。

3.4 通用智能体：应用场景泛化，需求精准适配

通用智能体已广泛覆盖生产生活的多个场景。既涵盖数据分析、教育、研究、编程、生产力等多元任务板块，提升日常办公效率；又渗透攻略计划制定、内容创作、生活助手等个性化场景，为用户提供精准决策建议，体现“全场景覆盖、全需求适配”的特征。



ChatGPT Agent应用案例—制定出行计划

资料来源：至顶智库结合公开资料整理绘制。

3.5 行业智能体：应用领域持续拓宽，场景创新不断深化

当前，行业智能体应用领域持续拓宽，场景创新不断深化。智能体凭借“环境感知—自主决策—动态执行”的核心能力，已在金融、医疗、工业、教育、政务、电信等领域得到广泛应用，显著提升各行业效率并创造新服务模式。在金融领域，智能体根据金融机构的独特需求和长尾业务进行深度定制，具有高度的针对性和专业性；在工业领域，智能体成为支持行业发展的“数字大脑”，推动新型工业化的核心引擎；在教育领域，智能体为师生教与学提供实时、个性化、启发式服务。

行业智能体典型应用场景

医疗领域，Hippocratic AI 旗下的 Healthcare Agent 能够处理多项非诊断性但耗时的任务，包括预约准备、术前术后电话沟通等。其也可根据用户需求提供入院、服务设施等信息。

Hippocratic AI — Do No Harm —

工业领域，西门子发布的Industrial Copilot有设计、规划、工程等五大功能，可独立执行完整工业工作流程。设计工程师可借助其处理复杂任务。

SIEMENS

教育领域，科大讯飞发布的星火教师助手具备对话、教学设计等多个模块，推动教师教学创新、提高思维能力。教师仅需语音下达指令，系统便能迅速结合班级数据诊断班级学情。

科大讯飞 IFLYTEK

政务领域，中国移动旗下移动云发布的政务智能体集通用问答、事项指引、个性定制、热点报告四大功能于一体。可对用户提问“秒回”并提供可视化指引，推动基层政务平台减负。

中国移动 China Mobile

法律领域，联想联合图灵法思共同发布的联想法思AI律师助手包含争议解决、法务咨询等八大功能，提高法务处理效率。用户将案件材料上传后，助手会自动梳理案情、整理要素。

Lenovo 联想

容联云 CLOOPEN

金融领域，容联云容犀Copilot&Agent，其七个业务智能体覆盖智能质检、知识管理、坐席辅助与业务分析等核心场景，实现营销-销售-客服-运营全链路提质增效。

明略科技 MINGLIAKE TECHNOLOGY

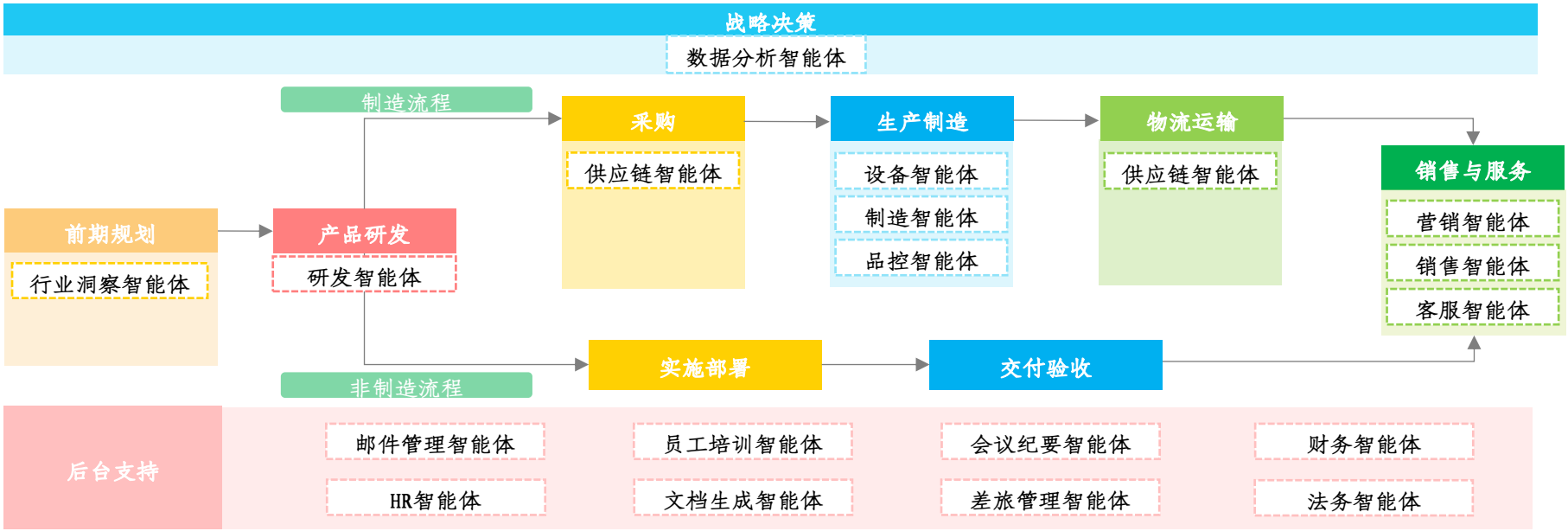
营销领域，明略科技发布的DeepMiner智能体包含智能规划、数据连接、预置知识、报告生成等核心功能，助力企业即时获取、整理、分析、洞察数据，提升工作效率，推动数据向精准决策转化。

资料来源：至顶智库结合公开资料整理绘制。

3.6 企业智能体：优化各部门工作流，促进智能化升级

企业智能体具备任务处理能力，作为具备自主决策能力的“数字员工”，为企业级用户提供多样化解决方案。一方面，不同行业先后涌现出适配自身工作流的智能体解决方案。在制造业中，企业智能体助力优化生产流程；在互联网行业中，企业智能体可被用于前期产品研发并协助销售服务。另一方面，企业内部办公场景中，智能体可作为企业后台支持部门的智能助手，协助处理日常行政、人力资源、财务管理等工作，帮助企业降本增效。企业智能体展现出广泛应用价值，正成为推动企业数字化转型和效率提升的关键力量。

智能体在企业内部的应用示例

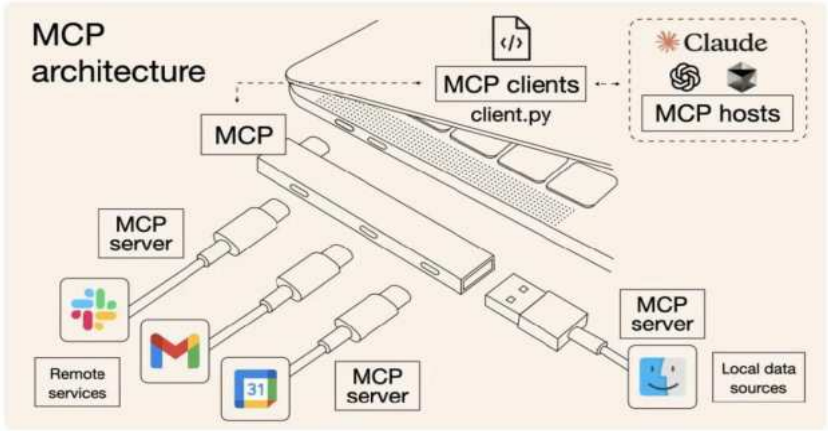


资料来源：至顶智库结合公开资料整理绘制。

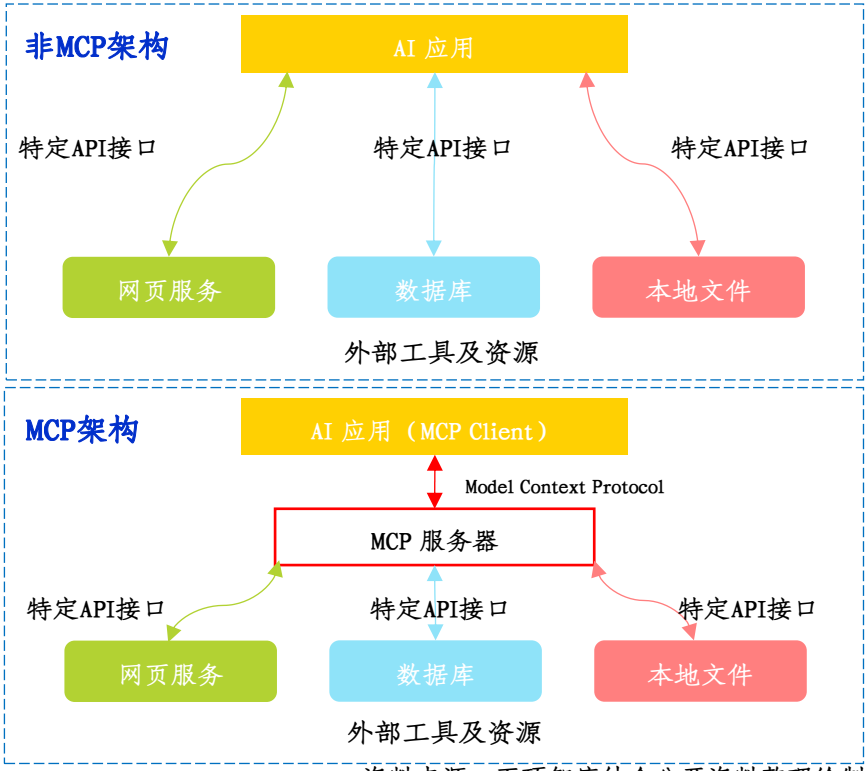
3.7 MCP协议：助力AI模型与不同数据源和工具实现无缝交互

MCP（Model Context Protocol）是由Anthropic提出的一套标准化交互协议，旨在创建一个通用标准，使AI应用程序的开发和集成变得更加简单，使得开发者能够以一致的方式将各种数据源、工具和功能链接到AI模型，为Agent开发提供支持。作为重要交互协议，MCP使得AI模型和应用开发进一步解耦，显著降低Agent的开发门槛。MCP Server标准化封装，本质上为模型及智能体提供更细粒度、轻量化的工具调用能力，降低智能体对复杂工具的调用门槛。

MCP协议主要特征



MCP架构主要由MCP主机、MCP客户端及MCP服务器三部分组成。AI基于MCP实现与外部工具及数据源无缝通信：用户发起一系列Prompt后，MCP主机上运行的MCP客户端，向MCP服务器发送初始请求，结合数据源（远程服务、数据库、本地文件等），经工具选择、API调用等操作给出初始响应与通知等操作。MCP主机、客户端、服务器三个核心组件协同运作，实现AI应用与外部工具和数据源间的无缝通信，满足用户需求 。

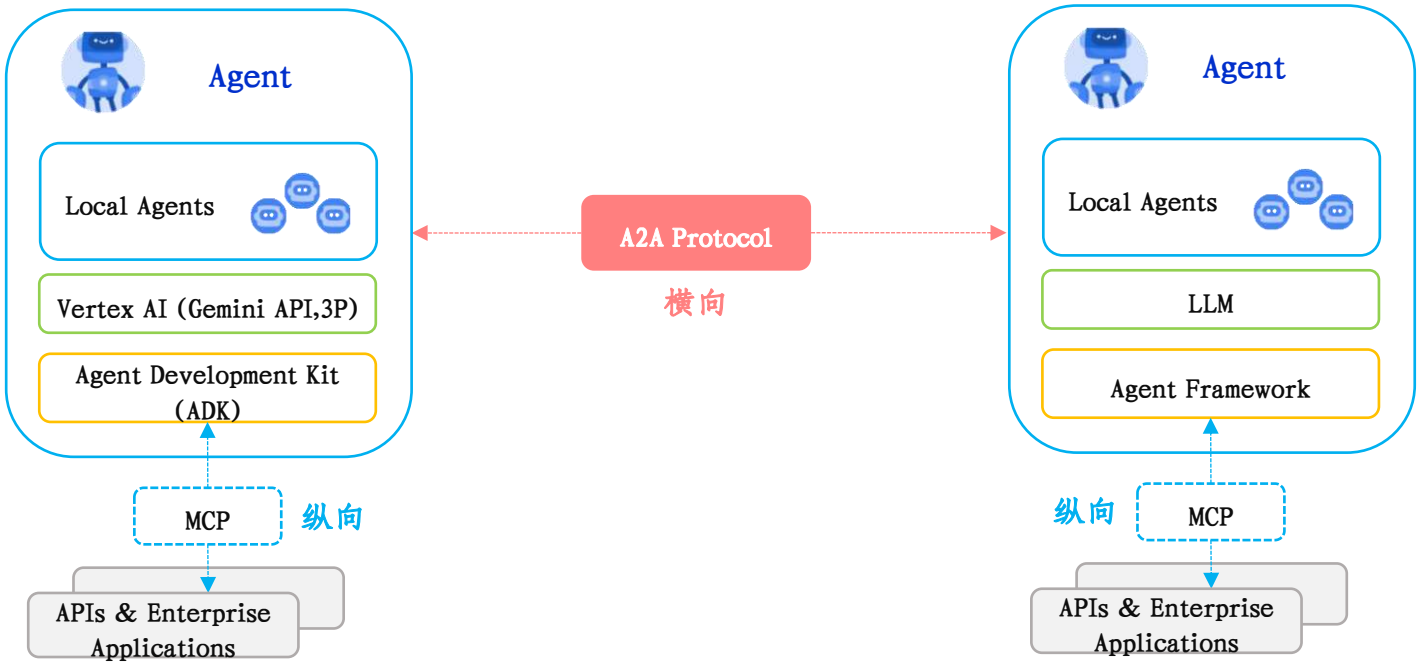


资料来源：至顶智库结合公开资料整理绘制。

3.8 A2A协议：助力不同智能体实现横向协作

2025年4月，Google发布A2A（Agent2Agent Protocol）开源框架协议，旨在促进不同智能体之间的横向协作。该协议可在无需框架或供应商的情况下实现智能体在不同生态系统间的协作，由此推动智能体在更广泛场景中的落地。A2A协议帮助智能体超越孤立的数据系统和应用程序从而完成协作，进一步提升智能体的自主性和生产力。此外，A2A协议支持音视频流等多种交互模态，既能高效处理即时任务又能支持深度研究。与MCP协议不同，A2A协议侧重解决大规模多Agent部署问题，是对纵向解决智能体工具调用问题的MCP协议的有效补充。

A2A与MCP协同工作示例

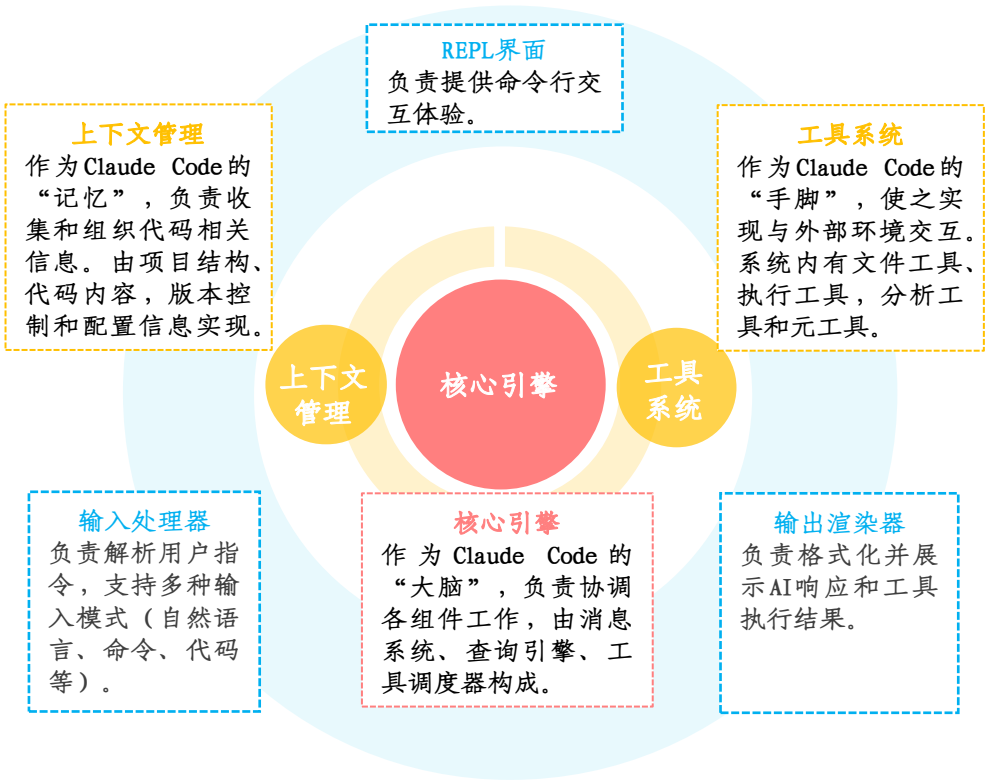


资料来源：至顶智库结合公开资料整理绘制。

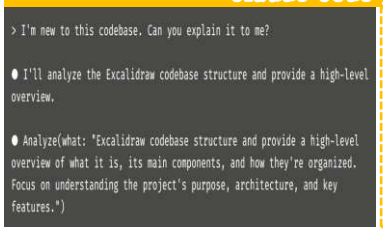
3.9 Anthropic：Claude Code变革应用开发范式

Claude Code是由Anthropic开发的全新编程智能体工具。旨在通过自然语言指令帮助开发者高效完成代码编写、调试和项目管理等任务，可被直接集成于开发者工作环境中。从系统架构来看，Claude Code主要由核心引擎、执行层和交互层构成。核心引擎充当“大脑”的作用，负责协调工作；交互层是用户与Claude Code的接触点；执行层负责执行环境交互、收集信息等任务。在功能上，Claude Code可实现代码解析、错误鉴别、代码重构等功能，能力突出。

ANTHROP\C **Claude Code技术架构图**



Claude Code 应用案例一



Claude Code能迅速解析代码库，智能搜索可实现无需手动选择文本亦能理解项目结构与上下文关系。

Claude Code 应用案例二

Claude Code可完整处理从阅读问题、编写代码、运行测试到提交PR的完整工作流程。

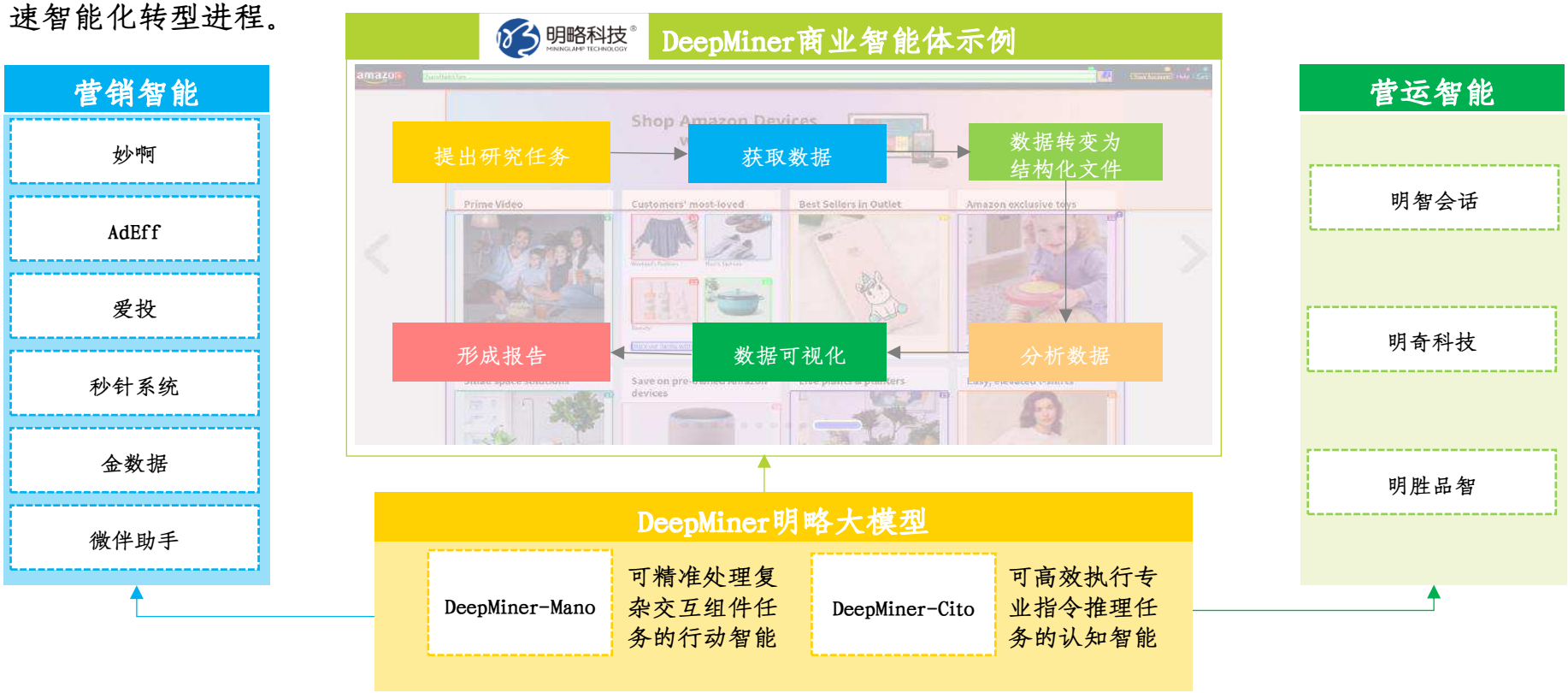
Claude Code 应用案例三

Claude Code可凭借对代码库与上下文关系的深度理解高效执行真正可用的多文件编辑工作。

资料来源：Anthropic，至顶智库结合公开资料整理绘制。

3.10 明略科技：DeepMiner商业数据分析智能体为企业精准决策赋能

明略科技作为数据智能应用软件提供商，为企业用户提供以数据分析为核心的产品服务。以明略大模型（DeepMiner-FA、DeepMiner-Mano、DeepMiner-Cito）为支撑，在营销与营运领域分别提供妙啊、AdEff、爱投、秒针系统、金数据、微伴助手，以及明智会话、明奇科技、明胜品智在内的全栈智能化产品及解决方案。基于明略大模型构建的DeepMiner全球商业数据分析智能体，拥有多智能体协作、复杂任务模版化等能力，具备智能规划、数据连接、预置知识和报告生成等多种功能。明略科技相关产品协同发力，旨在推动全球企业加速智能化转型进程。



资料来源：明略科技，至顶智库整理绘制。

4.智能硬件及AI应用进展

AI眼镜　　AI手机

AI PC　　智能汽车

Deep Research　　AI编程

4.1 智能硬件：AI眼镜

Meta、小米、夸克、雷鸟、Rokid、影目等推出多款AI眼镜，产品销量呈现爆发式增长。Meta发布的AI眼镜兼具时尚外观和实用功能，在全球范围引发广泛关注。在2025世界人工智能大会上，阿里夸克推出的AI眼镜具有多重优势：硬件层面，夸克AI眼镜进行系统重构，采用双芯双系统设计；AI能力层面，夸克AI眼镜实现从基础语音指令到自然对话系统的跨越；生态整合层面，夸克AI眼镜深度融合阿里及支付宝生态，为用户带来更智能、便捷的穿戴体验。阿里夸克依托在硬件领域的积累，基于通义千问模型能力，打造实用好用的AI眼镜。

主流AI眼镜对比

品牌及产品	AI功能	计算平台	续航能力
Meta AI眼镜	AI助手、拍摄、音乐播放、免提通信、翻译	骁龙AR1	典型续航4h
小米 AI眼镜	AI助手、拍摄、语音通话、支付	骁龙AR1+低功耗处理芯片	典型续航8.6h
夸克 AI眼镜	AI助手、拍摄、音乐播放、语音通话、出行提醒、导航、支付、购物、翻译、提词	骁龙 AR1+低功耗协处理器	——
雷鸟RayNeo X3 Pro	AI助手、拍摄、音乐播放、导航、翻译	骁龙AR1	5h录音时长 3h音乐播放 0.6h视频拍摄
Rokid Glasses	AI助手、拍摄、音乐播放、导航、翻译、提词	骁龙AR1	12h日常使用 4h连续蓝牙电话 6h音乐播放
影目INMO AIR3	AI助手、拍摄、导航、翻译、购物、社交	紫光展锐穿戴式芯片	典型续航3h

资料来源：至顶智库结合公开资料整理绘制。

Meta AI眼镜功能

- AI助手：使用Meta AI基于实时环境获取一系列建议和解答。
- 拍摄：按下捕获按钮或说“嘿Meta，拍照”，即可拍摄超高质量的照片和视频。
- 音乐：内置蓝牙扬声器与麦克风提供音频播放与捕捉服务。
- 免提通信：通过AI眼镜实现文本发送、语音及视频通话的免提操作。
- 翻译：提供法语、意大利语、西班牙语和英语的实时翻译服务。

夸克 AI眼镜功能

- AI助手：基于夸克多模态大模型，实现百亿级图片检索与专业讲解。
- 拍摄：基于Super RAW超级暗光增强模式优化夜景拍摄效果，AI算法实现超帧边分，高性能IMU实现稳定高清精准防抖。
- 导航：通过导航软件随时随地定制近眼显示导航系统。
- 支付：通过眼镜扫码与声纹识别实现安全支付。
- 出行提醒：夸克眼镜联合飞猪旅行等应用，绑定用户行程，时刻提醒出行计划。
- 购物：夸克眼镜联合拍立淘，实现海量商品搜索与比价。

资料来源：至顶智库结合公开资料整理绘制。

4.2 智能硬件：AI手机

AI手机的应用场景不断拓展，智能化能力不断增强。国内外主流手机厂商不断创新，苹果、三星、小米、华为、OPPO、vivo等推出搭载各类AI功能的新款机型。相关AI能力涵盖从基础功能（如文案写作、图像生成）到进阶功能（如识图问答、自动导航）等方面。具体来看，苹果的Visual Intelligence通过拍照并结合AI模型提供信息反馈；小米的自动导航通过提取消息的地理位置并自动发送至地图实现导航。

Apple Intelligence

写作工具（Writing Tools）：
支持“改写(Rewrite)”“校对(Proofread)”和“摘要(Summarize)”功能。

图像创作与生成：
可通过文字提示生成图像，支持动画、插画和涂鸦等风格。

视觉智能（Visual Intelligence）：
在iPhone 16/16 Pro及更新机型中，可长按相机按钮拍摄物体图片并发送至ChatGPT或进行网络搜索。

Siri智能升级：
提升自然语言处理能力，具备上下文理解能力及更自然的对话体验；支持文字输入Siri、在屏幕显示时响应上下文请求等交互方式。

超级小爱

识图问答

屏幕圈搜

自动导航

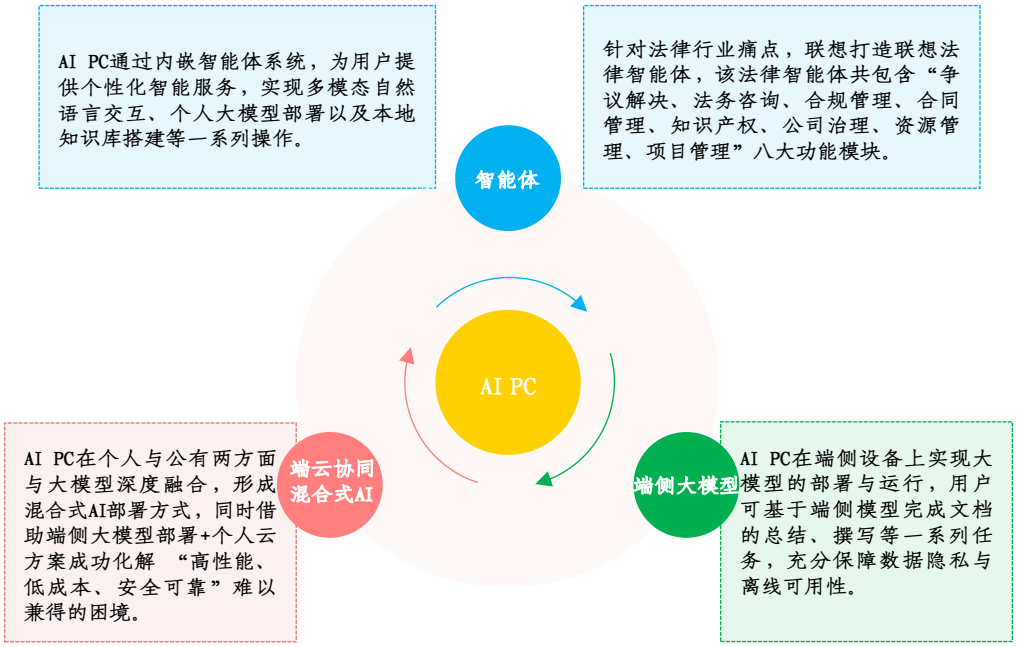
文档问答

资料来源：至顶智库结合公开资料整理绘制。

4.3 智能硬件：AI PC

AI PC领域中，国内外厂商现已相继推出一系列AI PC产品。现有AI PC呈现内嵌智能体、端侧部署AI大模型与端云混合式AI部署方式三大典型特征。AI PC通过内嵌智能体与端侧部署大模型，进一步提升用户在工作和学习场景中的智能体验。同时，端侧大模型+个人云的部署方案在保证AI工作的高性能、低成本与随时可用性的同时，保障用户的数据隐私。联想现已形成覆盖笔记本、台式机、工作站与服务器的全方位产品矩阵，不断丰富软件生态，展现AI PC在各类场景应用的广阔空间。

2025 AI PC典型特征



资料来源：至顶智库结合公开资料整理绘制。

Lenovo 联想 AI PC产品矩阵

产品类型	产品型号	产品功能
笔记本	YOGA Air 15 Aura YOGA Pro 14 YOGA 360 14 小新Pro 16 ThinkPad P1/P16/P16v/T14p/X1 Carbon ThinkBook 14+/16+/16p	➢ 内置联想天禧个人超级智能体。基于本地异构算力运行，通过大动作模型模拟人类的操作方式在PC上执行各类任务。 ➢ 同时，天禧智能体全面接入DeepSeek-R1联网满血版大模型，在响应速度、复杂指令执行与跨端任务执行等方面性能得到显著优化。
台式机	ThinkCentre P900/P900c Neo Ultra M460 T490	
工作站	ThinkStation PX ThinkStation P7/P8 ThinkStation P620/P520/P5	➢ 为开发者测试、微调或推理最新一代推理型AI模型提供充足支撑。
服务器	ThinkSystem	➢ 实现万亿参数的HPC和AI计算，为科学模拟、AI训练、大规模建模等场景提供硬核支撑。

资料来源：至顶智库结合公开资料整理绘制。

4.4 智能汽车：端到端自动驾驶技术体系

端到端自动驾驶为高级别自动驾驶发展提供全新的技术路径。基于统一的神经网络从传感器数据输入直接到控制指令输出的连续学习与决策过程。核心技术主要体现在数据采集标注、训练算法、模型设计训练、模型压缩等方面。端到端自动驾驶模型依赖数据闭环实现算法性能持续提升；训练算法旨在通过数据闭环建立原始传感器输入到驾驶规划控制指令的映射，提高自动驾驶系统性能和安全性；模型设计训练分为“分模块模型”和“单模型架构”两种；模型压缩主要包括模型剪枝、模型量化和编译优化，旨在降低网络计算需求，提升车载计算平台的运行速度。

端到端自动驾驶技术体系

数据采集标注

典型训练算法

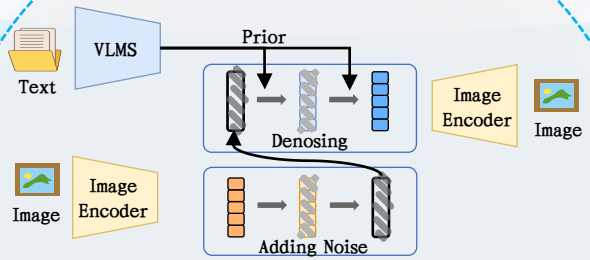
监督学习

监督学习是对人类驾驶员示范数据的模仿，依托模型输出和驾驶员决策控制量的差值对驾驶模型进行迭代更新。

强化学习

强化学习不依赖于带有驾驶操作标签的驾驶数据，通过车辆与环境的交互优化和探索试错，逐步提升驾驶性能。

融合多模态大模型的端到端自动驾驶



具备文字、语音、图像处理能力的多模态大模型（VLM）融入自动驾驶系统，能够直接从传感器数据中提取丰富的环境信息，并借助语言模型理解人类指令并生成可解释的决策过程，最后将多模态信息转化为驾驶操作指令。

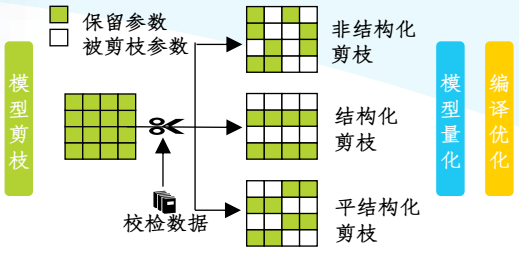
模型设计训练

分模块“端到端”模型

感知 预测 规控

单模型“端到端”架构

模型压缩技术

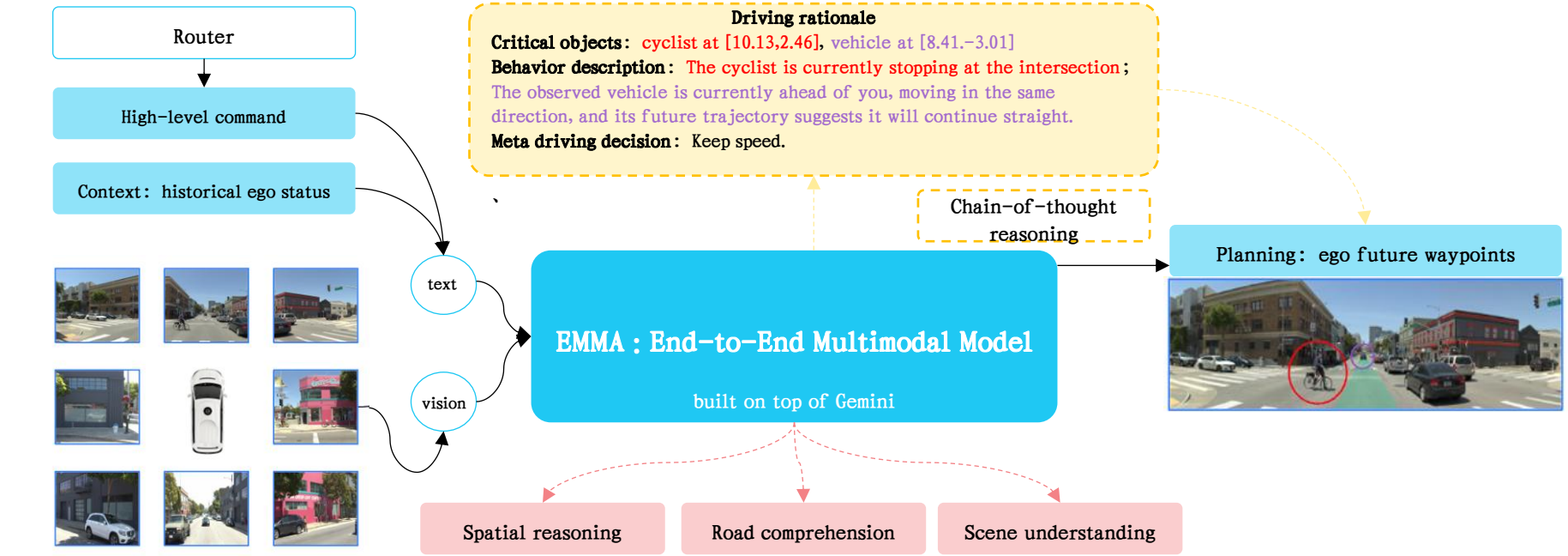


资料来源：李升波，江昆等：汽车智能驾驶技术发展与趋势展望，前瞻科技，至顶智库结合公开资料整理绘制。

4.4 智能汽车：端到端自动驾驶演进路径—VLA模型

视觉-语言-动作模型（VLA）由VLM模型演变而来，其结合视觉、语言和动作三种能力，旨在实现从感知输入直接映射到控制输出的完整闭环能力，不仅关注环境感知，也关注规划与控制问题。Waymo发布的EMMA模型具备同时处理文本、图像、视频等多模态输入，将驾驶任务定义为视觉问答（VQA）问题，最终生成多种驾驶输出形式（如规划轨迹、感知对象、道路图元素等）。EMMA充分利用Google Gemini模型储备的世界知识更好理解驾驶过程中的动态变化，作为VLA模型在自动驾驶领域的初步实践。

Waymo：EMMA端到端自动驾驶模型技术架构

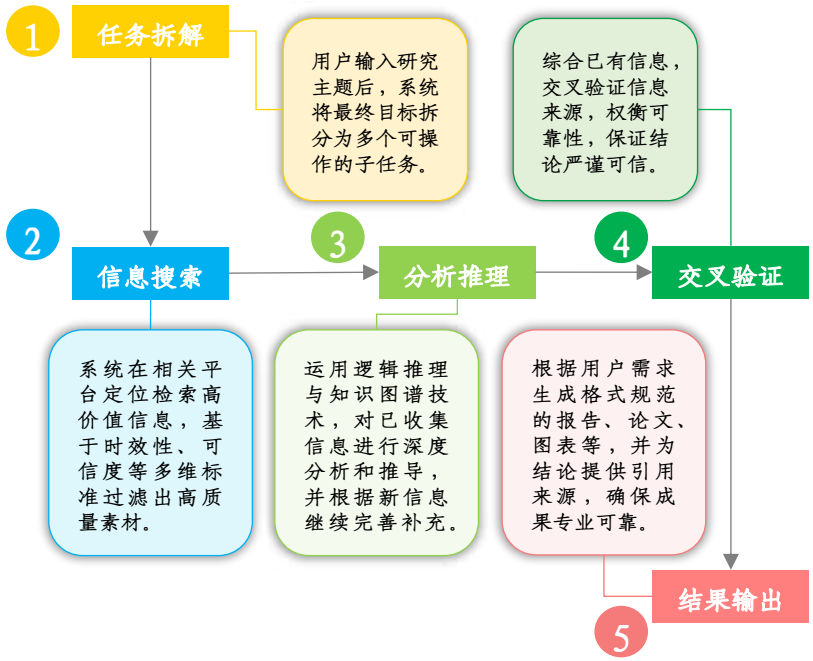


资料来源：*EMMA: End-to-End Multimodal Model for Autonomous Driving, Waymo*，至顶智库结合公开资料整理绘制。

4.5 Deep Research：从信息搜集到深度洞察的跃迁

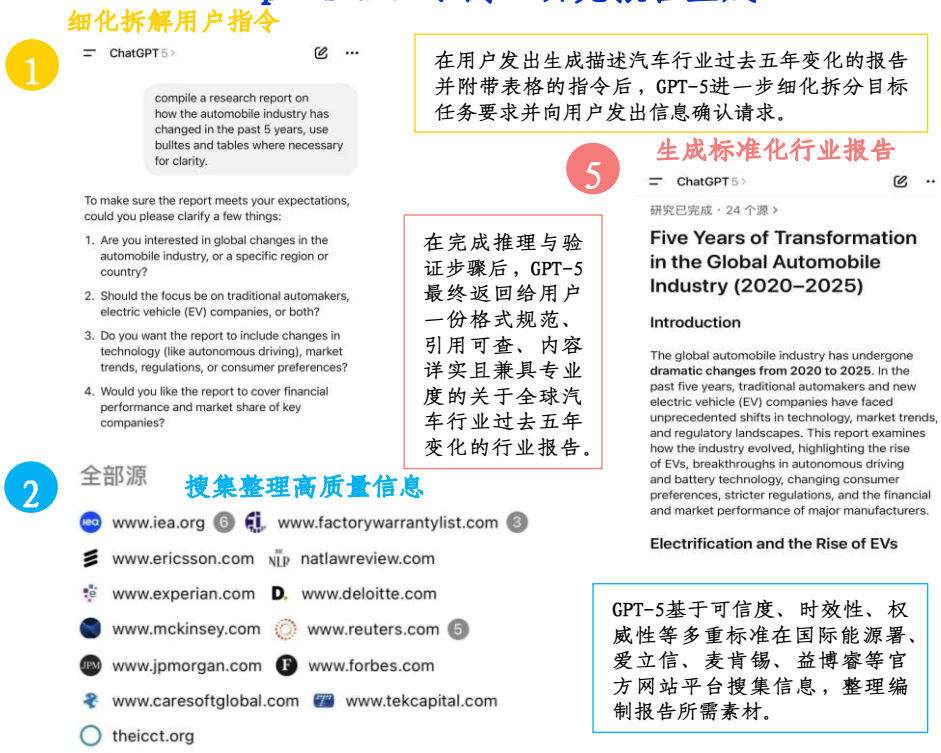
Deep Research（深度研究）作为当前信息处理与知识发现领域的重要功能，核心价值在于突破传统信息检索的局限，实现从信息搜集到深度洞察的跃迁。深度研究由任务拆解、信息搜索、分析推理，交叉验证和结果输出五个核心步骤构成，可实现输出结果的专业严谨。该功能正逐步成为现代知识工作者（如研究人员、分析师、决策者、产品开发者）不可或缺的智能伙伴，能够大幅提升信息处理效率与认知深度，成为释放人类专业潜能、驱动复杂问题解决与知识创新的关键应用。

Deep Research主要工作流程



资料来源：至顶智库结合公开资料整理绘制。

Deep Research示例：研究报告生成



资料来源：OpenAI GPT-5，至顶智库绘制。

4.6 AI编程：从辅助工具到智能体，重构软件开发范式

当前，AI编程正在深刻重构软件开发范式，其核心能力已突破传统辅助工具的边界，形成覆盖AI编程多环节的解决方案。从全球格局来看，AI编程已进入规模化应用阶段，国外以基础模型创新和自主智能体开发为主导，重视通用性与开发者体验；而国内重视工程化工具链整合，强调行业落地与本土化适配，形成差异化竞争格局。从编程环节来看，基于大规模预训练模型的代码生成系统能够准确理解开发者意图，完成从自然语言描述到可执行代码的转换，同时支持多编程语言和复杂算法实现，可针对特定代码库提供优化建议，实现代码补全、错误识别到代码优化的闭环。

国内外主流AI编程应用及功能分布

资料来源：至顶智库结合公开资料整理绘制。

5.全球AI企业最新布局

NVIDIA　　Google

OpenAI　　Anthropic

阿里巴巴　　DeepSeek

5.1 NVIDIA：全球领先的人工智能基础设施提供商

NVIDIA发布的AI高性能芯片以及计算设备，作为推动全球人工智能发展的关键基础设施，持续发挥重要作用。

2022年以来，NVIDIA相继推出基于Hopper架构和Blackwell架构的高性能计算产品线涵盖H100 Tensor Core GPU、Blackwell Ultra GPU等。GTC 2025大会提出的AI计算产品路线图，计划在2026年发布下一代AI芯片Rubin，Rubin提供50 PFLOPs密集FP4计算能力，而Rubin Ultra的密集FP4浮点运算性能更是提升至100PFLOPs，AI算力性能有大幅提升，持续推动超大规模计算以及AI模型训练和推理能力的提升。

NVIDIA AI 基础设施

Hopper 架构（2022）

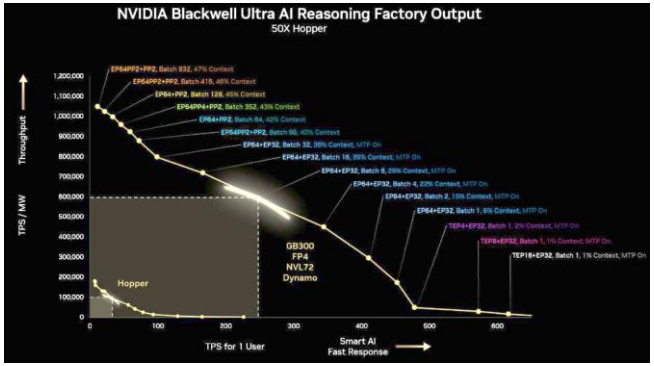
NVIDIA H100 Tensor Core GPU

由NVIDIA Hopper架构驱动的H100 Tensor Core GPU具有800亿晶体管，适用范围广泛，涵盖从小型企业到百亿亿级高性能计算，再到万亿参数的人工智能模型。

第四代张量核心与A100相比，芯片间速度最高可达6倍，包括每个流式多处理器（SM）的加速、额外SM数量以及更高的时钟频率。

全新Transformer引擎结合软件和专门设计的Hopper张量核心技术，专门用于加速Transformer模型的训练和推理。

Blackwell 架构（2025）

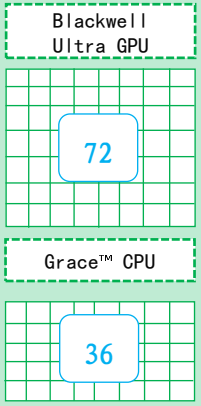


NVIDIA Blackwell

NVIDIA Blackwell架构适用于大规模推理AI场景，能效比上一代Hopper GPU高出30倍，并支持实时性能下的高吞吐量。

通过优化Blackwell Ultra GPU并行化策略（专家/张量/流水线）和跨GPU管理，基于Blackwell Ultra的GB300系统推理效能相比Hopper系统实现50倍AI工厂产量或生产力提升，同时保持低延迟，实现收益最大化。

DGX GB300



DGX SuperPOD AI 超级计算机

搭载DGX GB300系统，NVIDIA DGX B300作为AI基础设施平台，为数据中心提供高效的生成式AI和AI推理功能。

DGX GB300系统

采用NVIDIA Grace Blackwell Ultra超级芯片，搭载36颗NVIDIA Grace™ CPU和72颗NVIDIA Blackwell Ultra GPU。Blackwell架构GPU具有2080亿晶体管。采用台积电4NP工艺。

Rubin（2026）

NVIDIA Rubin

GTC 2025大会，下一代AI芯片Rubin亮相，Rubin提供50PFLOPs密集FP4计算能力，相比B300提升超过3倍。Rubin Ultra的密集FP4浮点运算性能提升至100PFLOPs。

资料来源：NVIDIA官网，NVIDIA H100 GPU技术文档，NVIDIA Blackwell架构技术文档，至顶智库整理绘制。

5.1 NVIDIA：全球领先的人工智能基础设施提供商

CUDA（Compute Unified Device Architecture）作为NVIDIA于2006年推出的专有并行计算平台与编程接口（API），允许开发者利用NVIDIA GPU执行科学计算与高性能计算，目前CUDA支持超过900个库。CUDA-X建立在CUDA之上，是一套由NVIDIA提供的GPU加速微服务、工具及库的集合，专门用于加速数据处理、人工智能与高性能计算（HPC）场景应用。CUDA-X涵盖数学运算库、并行算法库、图像视频库、通信库、深度学习库等，拥有超过400个加速组件，通过GPU带来计算性能提升。

NVIDIA CUDA-X

Frameworks: PyTorch

Cloud Machine Learning Services: accenture, aws, databricks, Google Cloud, Microsoft

Deployment: aws

Machine Learning | Deep Learning Training | Deep Learning Inference

CUDA-X AI

CUDA

Workstation: Dell, hp, Lenovo

Server: cisco, Hewlett Packard Enterprise, IBM, Lenovo

Cloud: aws, Alibaba Group, Google Cloud, Azure

NVIDIA CUDA-X

NVIDIA CUDA-X基于CUDA®构建，包含用于构建应用的多种微服务、库、工具和技术，CUDA-X可为人工智能、高性能计算(HPC)等应用领域带来性能的提升。

CUDA-X库基于CUDA构建，凭借400多个库，开发者可以使用CUDA平台构建、优化、部署和扩展各类应用。

AI应用开发涉及数据处理、特征工程、机器学习、验证和部署等环节，每个环节要处理大量数据和执行大规模计算任务。

CUDA-X AI包括：cuDNN（深度学习基础库）、TensorRT（推理优化引擎）、NeMo Retriever（企业级检索增强生成工具）等工具，支持从数据预处理到模型部署的全流程加速。

资料来源：NVIDIA官网，GTC 2025，至顶智库整理绘制。

5.1 NVIDIA：全球领先的人工智能基础设施提供商

NVIDIA Llama Nemotron模型、NVIDIA NeMo和NVIDIA NIM为专业开发者和企业提供智能体构建解决方案。Llama Nemotron推理模型基于Llama模型构建并提供AI推理功能，NVIDIA在后训练期间对该推理模型系列进行增强，以提升多步数学运算、编码、推理和复杂决策能力。NVIDIA NeMo可以借助一系列工具构建和维护智能体。微软将 Llama Nemotron 推理模型和 NIM 微服务集成到 Microsoft Azure AI Foundry，为客户提供增强服务的选项，如针对 Microsoft 365的Azure AI Agent Service。

NVIDIA NeMo+NIM+Llama Nemotron助力开发者构建智能体

NeMo Curator
NeMo Curator通过大规模处理文本、图像和视频数据进行训练和定制，提高生成式AI模型的精度。

NeMo Guardrails
NeMo Guardrails是可扩展的护栏编排平台，用于确保LLM交互的安全性、精度和主题相关性。

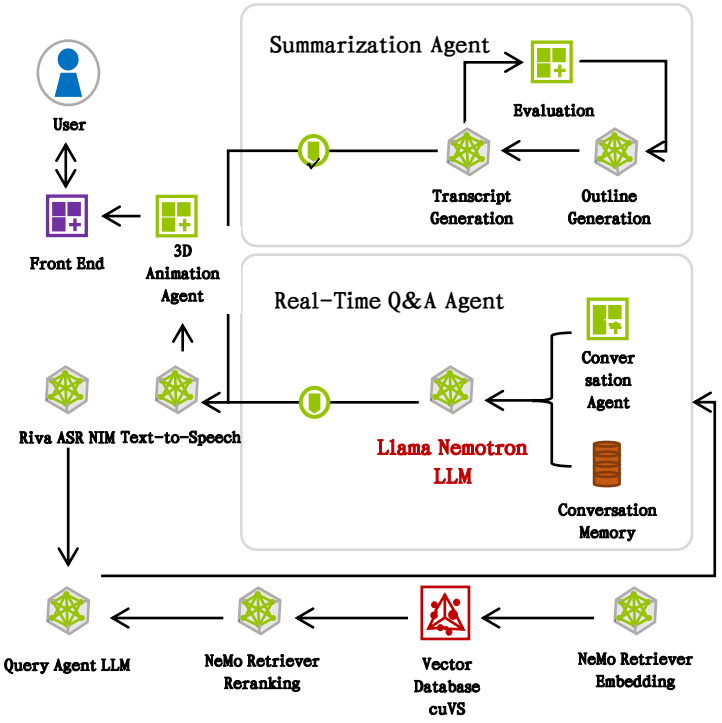
NeMo Customizer
NeMo Customizer作为可扩展的高性能微服务，可针对特定领域用例简化LLM的微调和对齐。

NeMo Retriever
NeMo Retriever是微服务的集合，可提供信息检索，并具有较高的准确性和更大的数据隐私性。

NeMo Evaluator
NeMo Evaluator用于在平台通过学术和自定义基准测试对生成式AI模型和工作流进行评估。

NVIDIA NIM

- Prebuilt container and Helm chart
- Industry standard APIs
- Domain specific code
- Optimized inference engines
- Support for custom models

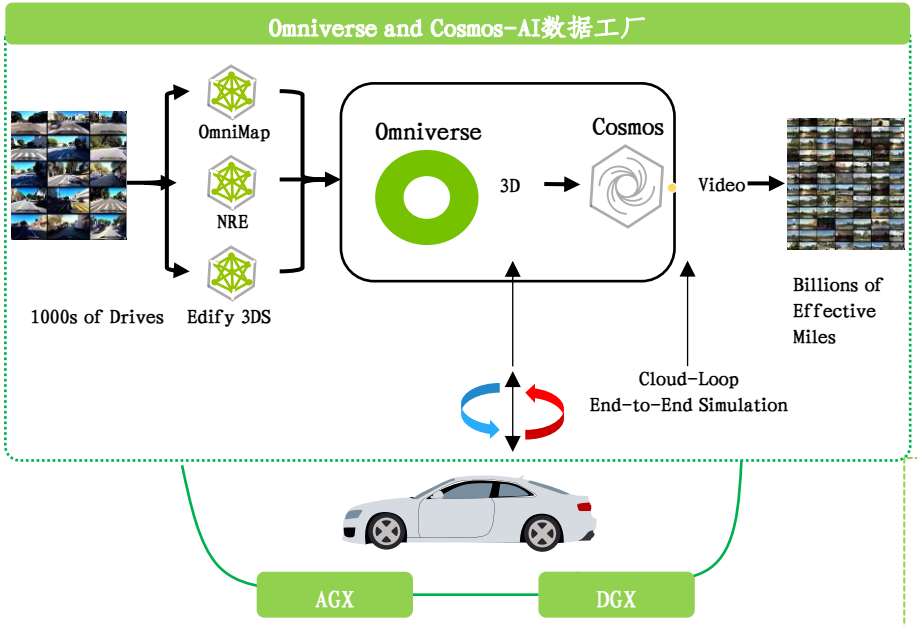


资料来源：NVIDIA官网，CES 2025，至顶智库整理绘制。

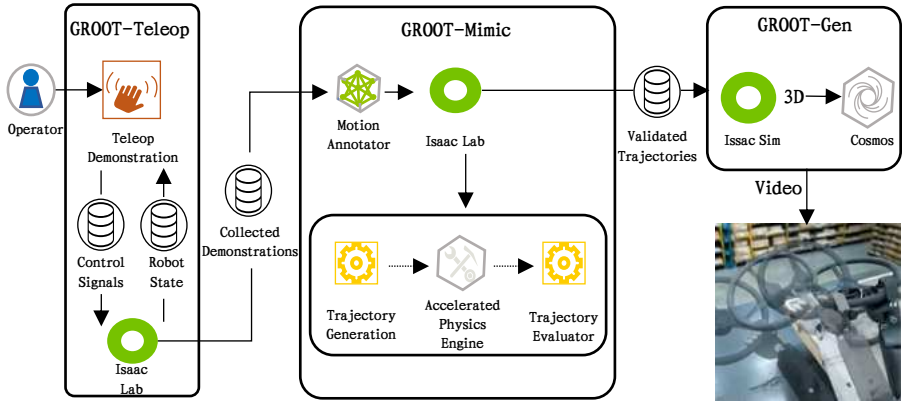
5.1 NVIDIA：全球领先的人工智能基础设施提供商

NVIDIA提供完整方案助力智能驾驶和具身智能发展。汽车场景中，NVIDIA为汽车智能化提供三类计算系统：用于AI模型训练的NVIDIA DGX、用于测试驾驶和生成合成数据的系统Omniverse以及车载超级计算机DRIVE AGX。利用Omniverse和Cosmos创建的“AI数据工厂”，通过合成驾驶场景大幅扩展训练数据。NVIDIA将数百次的驾驶场景扩展为数十亿的有效里程，大幅增加实现安全和先进自动驾驶功能所需的数据集规模。机器人场景中，Isaac GR00T涵盖机器人基础模型、运动与数据合成系统、仿真框架等，帮助开发者从少量人类示范数据中产生大规模数据集，推动具身智能快速发展。

NVIDIA 智能驾驶三类计算系统



NVIDIA Isaac GR00T运动合成系统

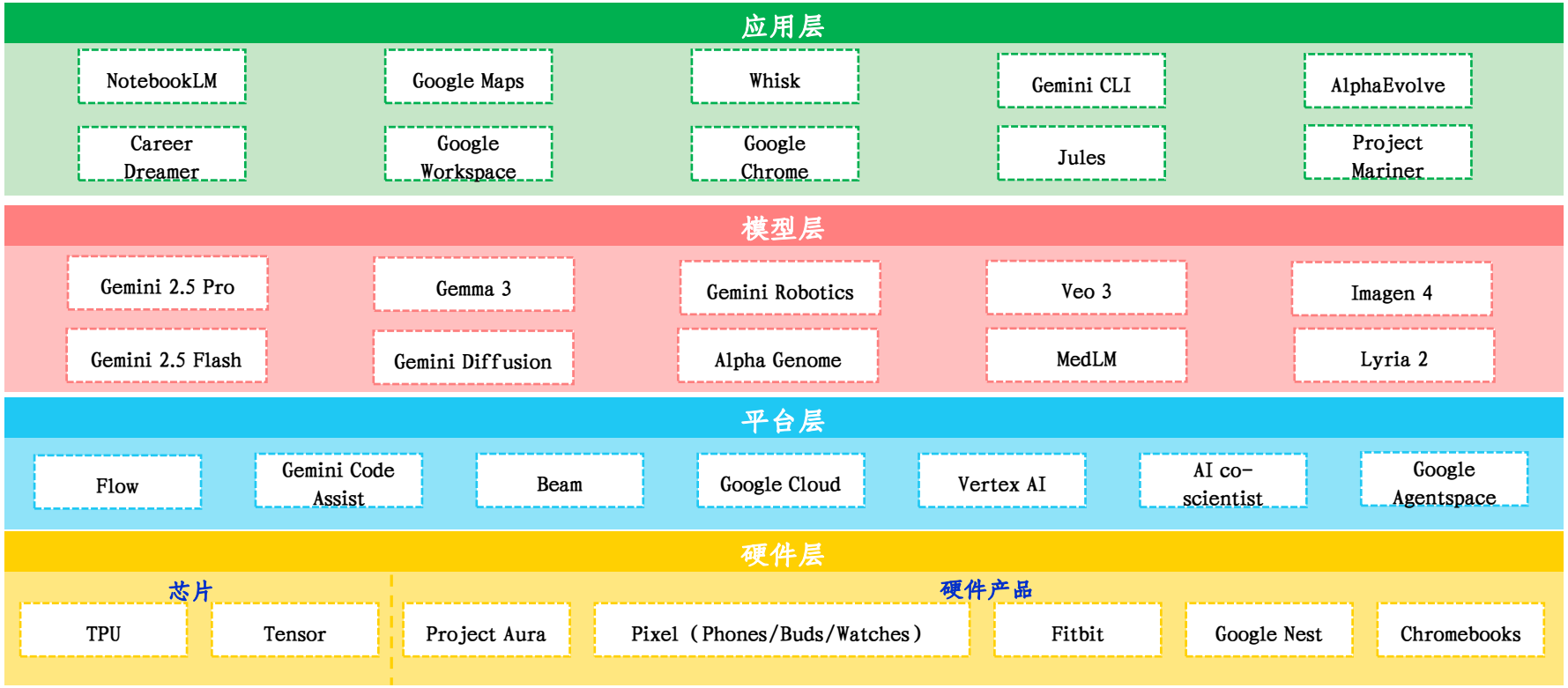


1）操作人员使用Apple Vision Pro进入数字孪生系统；
2）操作人员通过少量远程操作演示捕捉动作轨迹；
3）通过Isaac GR00T运动合成系统将少量运动轨迹扩展为更大规模的数据集；
4）利用Issac Sim和Cosmos所构建的GR00T-Gen进行3D图像放大；
5）在软件系统中进行测试验证，最终部署到物理机器人。

资料来源：NVIDIA，CES 2025，至顶智库整理绘制。

5.2 Google：“硬件—平台—模型—应用”全方位构建AI生态

Google的人工智能产品布局已形成四层协同发展的完整生态,包括硬件层、平台层、模型层以及应用层。2025年，Google持续推动四个层面的产品研发与更新，在各层面分别实现一系列重大创新。其中模型层面，Google前后推出Gemini 2.0系列以及Gemini 2.5系列大语言模型，在多模态理解、推理能力上实现跨越式提升；应用层，Google积极响应AGI发展热潮，推出了Gemini CLI、Project Mariner等一系列高自主性的AI Agent产品。



资料来源： Google，至顶智库结合公开资料整理绘制。

5.2 Google：基于Gemini 2.5大模型底座，持续增强多模态能力

2025年6月17日，Google正式发布Gemini 2.5 Pro、Gemini 2.5 Flash，并推出Gemini 2.5 Flash-Lite实验版本。Gemini 2.5模型采用稀疏MoE（Mixture-of-Experts）架构，原生支持文本、视觉和音频输入的多模态处理。2025年8月26日，Google发布重磅文生图模型Gemini 2.5 Flash Image（Nano Banana），在图像质量、编辑控制和应用场景上有大幅改进，不仅可以对人物和宠物进行精准编辑，保持特征一致，还能实现多图合成、多轮次修改与风格迁移等复杂操作，并融入现实世界知识。

Gemini 2.5系列模型介绍

Gemini2.5

Gemini 2.5 Pro

Gemini 2.5 Pro是目前谷歌最智能的思考模型，在推理方面一系列基准测试中达到SOTA水平；编程方面，在Gemini 2.0的基础上实现了巨大的飞跃——Gemini 2.5 Pro擅长创建视觉上引人注目的web应用程序和代理代码应用程序，以及代码转换和编辑。

同时，Gemini 2.5 Pro创新性引入“Deep Think”增强推理模式，模型能够在做出反应之前考虑多种假设，极大提高Gemini解决困难推理问题的能力。

Gemini 2.5 Flash

Gemini 2.5 Flash是一款基于蒸馏技术，尺寸更小的混合推理架构模型。2.5 Flash实现对成本与效率同时兼顾，在推理、多模态、代码和长上下文的关键基准上性能表现得到了改进，同时效率更高，将token使用量减少20%-30%。

同时，Google也推出了Gemini 2.5 Flash-Lite的预览版，2.5 Flash-Lite是2.5系模型中速度最快、性价比最高的一款模型，相比于2.0 Flash-Lite，在数学、编码等方面性能提升显著。

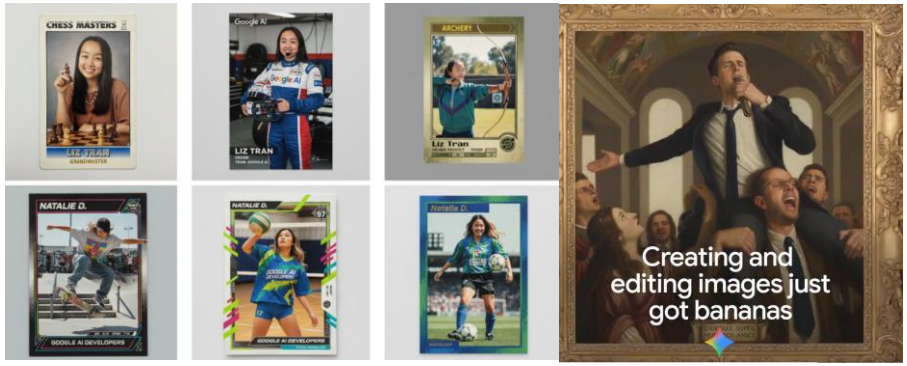
文生图模型Gemini 2.5 Flash Image（Nano Banana）

角色一致性

支持将同一角色置于不同场景、展示产品在多角度与多环境下的效果，或生成统一的品牌资产，同时确保主体外观高度一致。

精准图像编辑

支持通过自然语言进行精准的局部编辑与目标性转换。仅需输入简单文本提示，即可实现背景模糊、删除照片中的人物、调整主体姿势，等多样化操作。



融入世界知识

通过整合Gemini的世界知识，突破语义理解的局限，开辟全新应用场景。

多图像融合

能够理解并融合多张输入图像。通过单一提示将物体置入新场景、为房间重新设计配色或纹理，或生成逼真的多图融合图像。

资料来源：Google，至顶智库结合公开资料整理绘制。

5.3 OpenAI：从核心模型迭代到智能体演进，引领AI技术突破

OpenAI是全球领先的人工智能科技公司，以其在大模型领域的突破性进展而闻名。公司致力于开发和推广安全、有益的AI技术。目前，OpenAI已推出一系列创新产品，包括ChatGPT Agent、Codex、GPT-4o、OpenAI o3、o4-mini、GPT-5及Operator等，既实现多模态模型与推理模型的技术升级，又在智能体领域实现全新突破，组成多元的产品体系。2025年5月，OpenAI斥资65亿美元收购前苹果首席设计官Jony Ive的AI硬件初创公司io。

ChatGPT Agent

结合深度研究的思考和分析能力、Operator的操作执行能力，以及ChatGPT的智能与对话流畅性，能够主动从其智能体技能库中选择合适的工具，利用自身计算机系统完成任务。

Codex

作为基于云的软件工程智能体，可以执行各种任务，如编写功能、回答有关代码库的问题、修复错误和提出拉取请求以供审核；每项任务都在本地云沙箱环境中运行，并预装存储库。

GPT-4o

在GPT-4基础上改进，实现文本、语音、图像三种模态分析能力，模型实用能力大幅提高，可实现更自然的人机交互。

OpenAI o3 & o4-mini

与先前推理模型版本相比使用体验也更加自然、对话感更强，尤其能够参考记忆和过往对话，使回答更加个性化，更贴合需求。

GPT-5

由GPT-5、GPT-5 mini、GPT-5 Pro三个版本构成，集深度思考与日常模式于一体，可根据用户问题按需思考，在编程、推理、多模态等方面表现突出。

Operator

使用浏览器查看网页，用户可在Operator中通过为所有网站或特定网站添加自定义指令，实现个性化的工作流程。

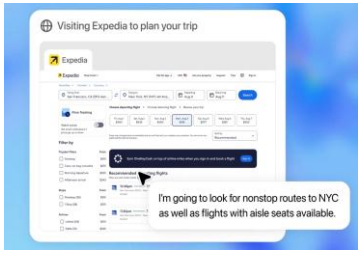
资料来源：至顶智库结合公开资料整理绘制。

5.3 OpenAI：从核心模型迭代到智能体演进，引领AI技术突破

2025年7月，OpenAI发布通用智能体ChatGPT Agent，该产品将Operator的网站交互、Deep Research的信息整合，以及ChatGPT的智能推理与流畅对话的三项优势融汇一体。ChatGPT Agent聚焦迭代式、协作式工作流程，交互性和灵活性显著提升，实现智能体能力的关键升级。同年，OpenAI发布o系列模型的最新成果o3，该推理模型擅长多模态理解，能够组合使用ChatGPT中的所有工具并有效应对多面性问题，是ChatGPT向更加自主方向迈进的重要举措。

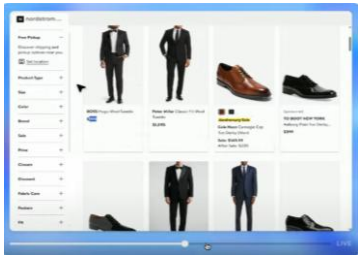
ChatGPT Agent介绍及示例

ChatGPT Agent能够自主利用多种工具进行规划，帮助人们完成复杂任务，包括自动浏览用户日历，生成可编辑PPT，运行代码等。还能连接用户Gmail、GitHub网站获取信息并解决问题，使用API来访问各种应用。Agent凭借自身强大的自主思考与行动能力，正逐步拓展至日常和专业场景中。



示例1：回应用户旅行线路规划和航班座位搜集的请求，根据用户指令在网页内行动并完成对应操作。

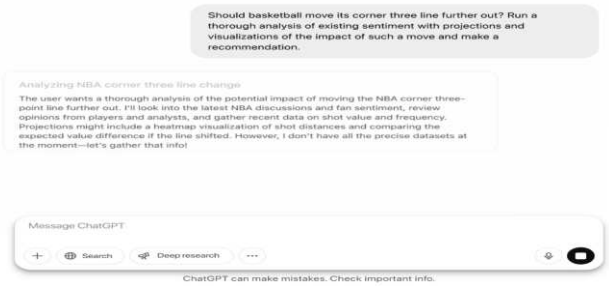
示例2：执行查询年度财务报告的操作，智能浏览网站、筛选结果，交付可编辑的幻灯片和电子表格，总结研究成果。



示例3：帮助用户提供可视化的西服商品选项，根据用户指令智能反馈给用户可选结果，轻松实现基于用户指令的任务推进。

OpenAI o3介绍及示例

OpenAI o3通过强化学习训练模型使用工具，能够有效应对视觉推理等开放式场景。此外，o3首次直接将图像融入思维链，实现视觉推理与文本推理的结合，其多模态问题解决能力突出。



该推理模型可根据用户提供的情境，如分析篮球运动轨迹—编写Python代码构建预测，生成图表或图像，调用多个工具，解释预测情境背后的关键因素。

资料来源：至顶智库结合公开资料整理绘制。

5.3 OpenAI：GPT-5实现多模型协同与编程能力突破

2025年8月，OpenAI发布最新一代多模态模型GPT-5，其中包括GPT-5、GPT-5 mini和GPT-5 Pro三个版本。GPT-5将非推理模型与推理模型融为一体，实现由单一模型向多模型协同方向演进。此外，GPT-5在编程与代码、数学与逻辑推理、多模态理解、健康咨询等方面均表现亮眼，实际问题解决能力突出。GPT-5的发布标志着大模型技术正从单纯追求“规模”，转向追求“效率与规模并存”的更成熟阶段。

GPT-5

GPT-5是一个协同工作的智能系统，由两个核心模型组成，并由实时路由器(Real-time Router)根据用户指令将计算资源在两者中进行分配。

- **The Smart, Efficient Model**：处理与用户日常交互的标准模型。
- **GPT-5 Thinking**：处理复杂难题的深度推理模型。

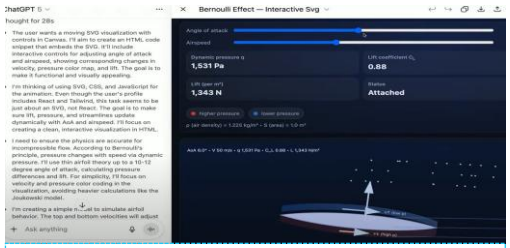
GPT-5 Pro

GPT-5 Pro是面向企业版与高级版开放的版本，作为OpenAI o3-Pro的替代品，通过并行测试时计算技术带来更全面且高质量的答案，是OpenAI对最具挑战性任务给出的解决方案。

GPT-5 mini

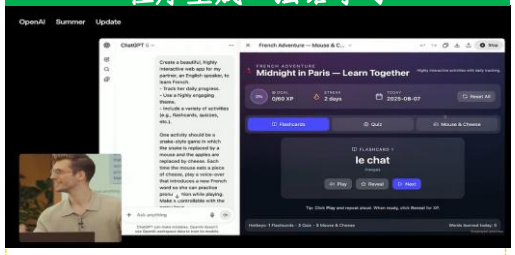
GPT-5 mini是GPT-5的轻量化版本，对GPT-5的核心功能进行保留的同时降低资源需求。

科学解释—伯努利效应



GPT-5对用户指令中的伯努利现象进行详细解释，并根据要求调动按需思考能力，完成深度推理、代码生成、图形结构设计等关键环节，最终生成动态SVG交互式演示图。

程序生成—法语学习



GPT-5发挥自身强大编程能力，根据输入的指令，及时响应用户法语学习需求，生成完整网页代码，呈现语言学习网站并内嵌趣味教育游戏，完成复杂应用程序设计。

语音交互—韩文学习

GPT-5语音功能出色，通过接入ChatGPT学习模式，以引导方式辅助用户进行韩语学习，实时为用户提供准确发音，改善语言学习方式，实现语音交互的自然无缝衔接。

资料来源：OpenAI，至顶智库结合公开资料整理绘制。

5.4 Anthropic：混合推理与多模态模型的行业领军者

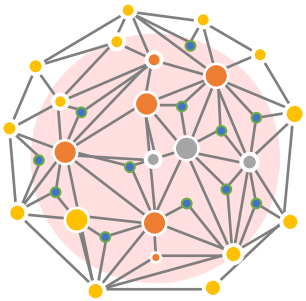
Anthropic成立于2021年，总部位于美国加州旧金山，是一家人工智能研发型企业。Anthropic通过一套预定义的价值与行为准则，引导模型在生成内容时保持高一致性与可解释性，注重长上下文、强推理与低幻觉率的协同。旗下的Claude系列大模型，已构建起层次分明、功能完备的产品矩阵。从轻量级部署的Haiku，到性价比优选的Sonnet，再到旗舰级产品Opus，该系列已覆盖多样化的应用需求，在长文档深度分析，复杂推理与多模态场景中表现亮眼。

Claude 4

2025年5月，Anthropic正式推出Claude 4系列大模型。此系列包括旗舰型Claude Opus 4和主力型Claude Sonnet 4，为代码生成、高级推理和AI智能体建立全新标准。作为混合模型的代表，Claude模型将传统大模型和推理能力结合，为用户提供迅速响应和深度思考两种模式，实现高效输出与高阶推理的有效平衡。

Claude Opus 4

Claude Opus 4的技术升级体现在针对编码和长期运行的智能体工作流的优化。作为最出色的编码模型，Opus 4擅长编码和复杂问题的解决，并能实现长时间持续的独立运行。

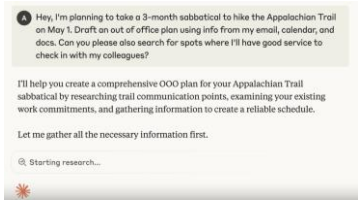


Claude Sonnet 4

Claude Sonnet 4在Claude Sonnet 3.7的基础上改进，重点在编程和推理能力方面优化，引入总结思考能力、加强工具使用能力，有效兼顾功能性与实用性。适用于复杂分析和高要求的应用场景。

Claude Research

2025年4月，Anthropic发布Claude Research功能。Research功能改变Claude获取和推理信息的方式，能够自动多角度探索用户问题并系统性解决开放式问题，为用户提供高质且全面的答案并附带详细引用数据，兼顾速度与质量。



如图，Research基于网络信息、用户日历表等内容全方位搜集信息并展开推理。

Claude Code

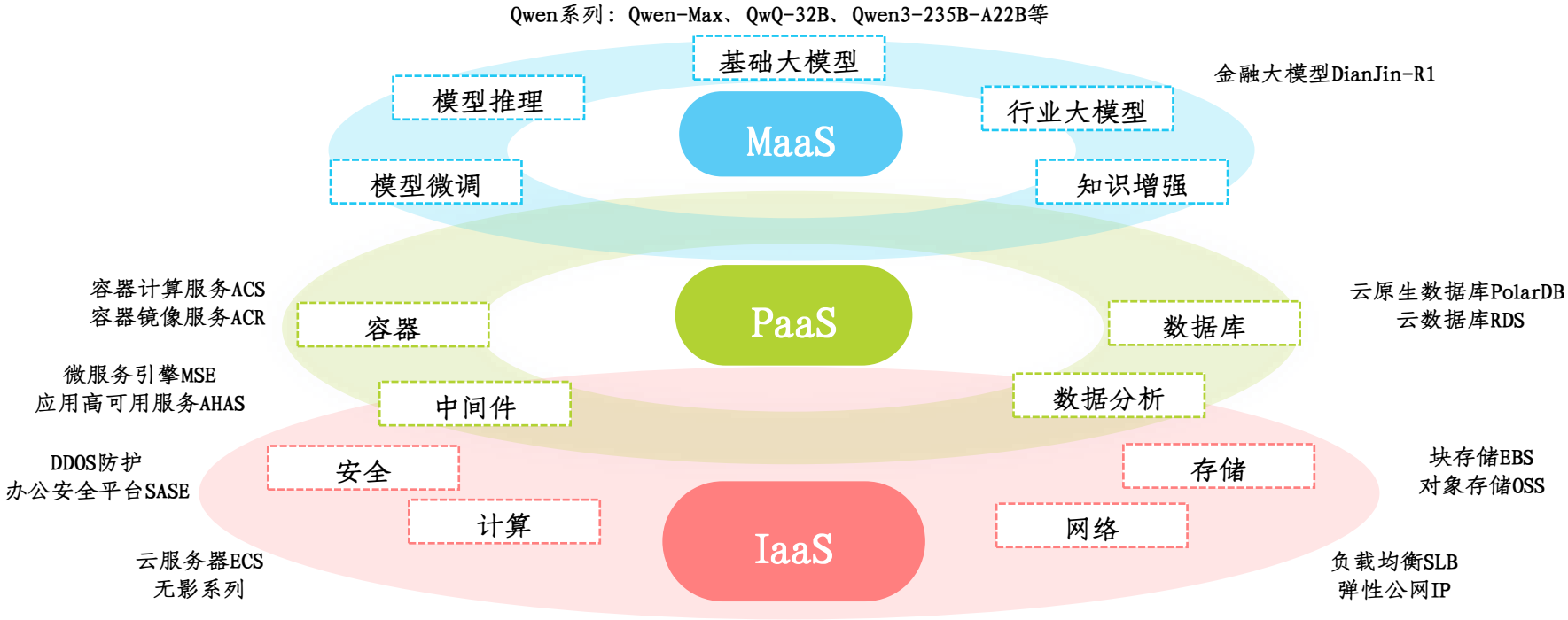
2025年2月，Anthropic推出Claude Code智能编程工具。Code工具有强大的跨文件编辑处理能力，能够搜索并阅读代码，同时理解并处理多个文件的相互关系。此外，Code还能够提交和推送代码到GitHub、执行命令行操作。

资料来源：至顶智库结合公开资料整理绘制。

5.5 阿里云：从基础设施到模型的全栈AI布局

阿里云现已构建其AI基础设施与技术服务体系，形成全栈AI智能底座。其体系可划分为IaaS、PaaS以及MaaS三个层面。阿里云正以AI为中心，全方位重构IaaS，PaaS，其中IaaS层包括安全、计算、存储等一系列基础设施服务；PaaS层，阿里云提供数据库、容器等一系列平台服务；MaaS层，阿里云已推出多款Qwen系列开源、闭源基础大模型，同时，基于模型微调，阿里云加速其大模型向其他行业的渗透节奏，目前已在政务、电力、能源、医药等多个行业部署行业大模型，推动从研发到生产等一系列工作流程的效率优化，加速政务与一系列企业数字化转型进程。

阿里云全栈AI布局

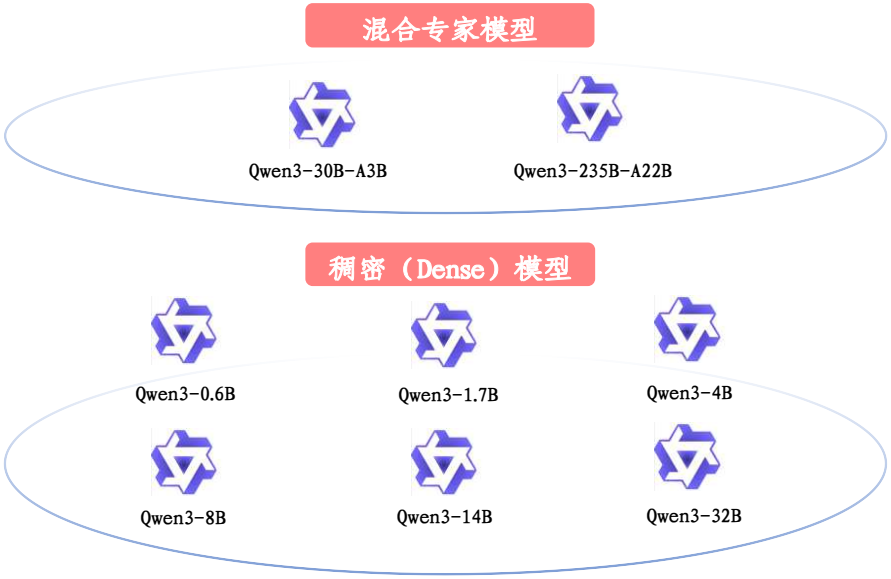


资料来源：至顶智库结合公开资料整理绘制。

5.5 阿里通义：“开源+闭源”大模型体系全面布局

阿里通义大模型秉持开源与闭源并进路线。开源端以Qwen系列覆盖多参数规模，性能领先且生态活跃，衍生模型超5万；闭源端主打Qwen-Max等旗舰模型，安全合规性强，深度赋能金融、医疗等行业，服务超9万家企业，实现技术普惠与商业落地的平衡。其中，Qwen3于2025年4月发布，是Qwen系列大型语言模型的最新成员，也是国内首个混合推理模型，将快思考与慢思考集成于一体，对于简单需求可低算力秒回答案，对复杂问题可多步骤深度思考，大大节省算力消耗。同时具备全系列、开源最强、混合推理等特性。

Qwen 阿里通义大模型体系

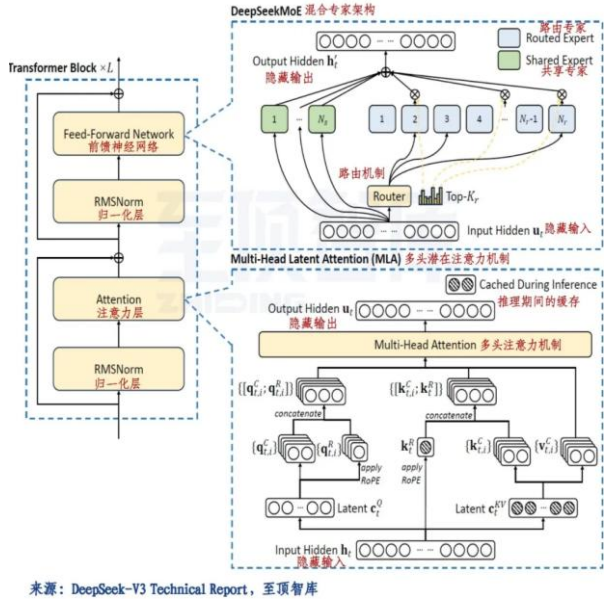


资料来源：阿里云，至顶智库结合公开资料整理绘制。

5.6 DeepSeek：MoE架构创新与推理模型的行业先锋

DeepSeek（深度求索）是一家成立于2023年的来自中国杭州的人工智能公司，其前身是国内量化投资巨头幻方量化的子公司。公司专注于开发低成本、高性能的AI模型，并在深度学习、强化学习等领域取得多项突破，特别是在混合专家架构（MoE）和多头潜在注意力机制（MLA）方面进行深入研究和创新。此外，DeepSeek坚持开源，公开模型权重和训练细节，吸引全球开发者和研究者的广泛参与。目前已发布V3和R1等多款性能突出的开源模型。

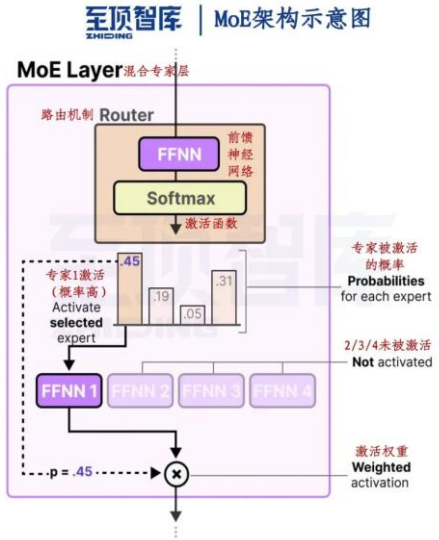
DeepSeek-V3模型架构图



来源：DeepSeek-V3 Technical Report，至顶智库

DeepSeek-V3在经典Transformer架构上进行改进，在前馈神经网络（Feed-Forward Network）引入DeepSeek MoE架构；此外在注意力层（Attention）中引入MLA机制，提高模型性能。

MoE架构示意图



来源：Maarten Grootendorst，混合专家模型（MoE）到底是什么？，至顶智库

混合专家架构（Mixture of Experts，MoE）是利用多个不同子模型（或“专家”）提升大语言模型质量的技术，主要由混合专家层和路由机制构成。

模型训练效率提升

通信优化

- 提出DualPipe算法，计算通信完全重叠；
- 双向流水线并行，降低流水线bubble；
- 跨节点通信优化，用确定性路由策略。

内存优化

- 重计算，前向计算不存，反向时计算；
- 使用CPU内存，节约GPU显存；
- 参数共享，降低内存，提高模型精度。

计算优化

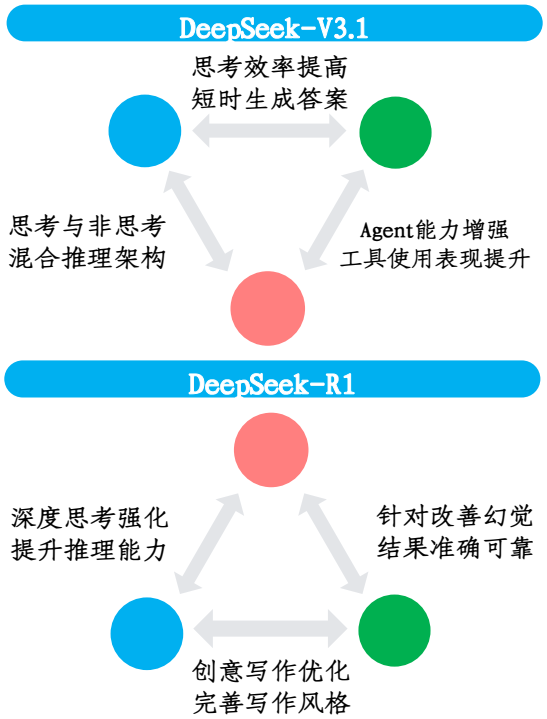
- 核心计算GEMM采用FP8,混合精度；
- 采用细粒度量化、增加累积精度、增加尾数、在线量化策略减缓outlier。

资料来源：至顶智库结合公开资料整理绘制。

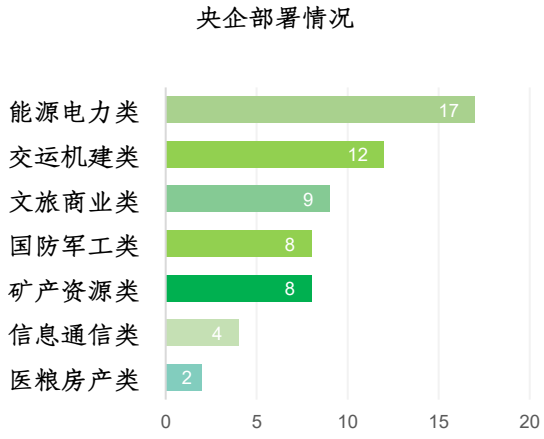
5.6 DeepSeek：模型能力持续迭代，应用部署加快落地

2025年8月21日，DeepSeek-V3.1模型发布，增加混合推理架构、提高思考效率并强化智能体支持；同年5月，DeepSeek-R1模型完成小版本升级，持续强化模型思维深度与推理能力。在落地实践方面，目前DeepSeek已助力超60家央企推动数字化与智能化转型；在智能终端领域，DeepSeek模型已深度适配手机、PC、家电、汽车等各类终端，不断提升产品智能化水平。

DeepSeek-V3.1&R1更新情况



央企部署情况



当前，DeepSeek已在多家央企及其下属机构完成部署，各行业均有体现。其应用场景广泛，涵盖办公、能源、通信、航运等多个领域，示范效应显著。

终端应用领域

DeepSeek+手机

HUAWEI　HONOR

oppo　vivo　MEIZU

DeepSeek+PC

Lenovo 联想

DeepSeek+家电

Hisense　Midea　TCL

DeepSeek+汽车

东风汽车集团有限公司 DONGFENG MOTOR CORPORATION　BYD

仰望

资料来源：至顶智库结合公开资料整理绘制。

6.AGI未来发展路径探究

世界模型　　多模态模型

持续强化学习　　非Transformer架构

具身智能

6.1 世界模型：从理解物理世界到预测未来变化

当前，世界模型成为推动人工智能加速迈向AGI的关键驱动力。世界模型不仅能够理解物理世界的现状，还能预测其未来潜在的一系列动态变化。世界模型的应用场景已覆盖具身智能、自动驾驶、游戏开发及场景生成等领域，展现出广阔的应用潜力。从研发情况来看，全球主要科技企业和研究机构已推出多个世界模型，如Google发布的Genie 3、Meta发布的V-JEPA 2、达摩院发布的WorldVLA，从架构创新到场景落地持续深化探索，推动人工智能发展迈向新阶段。

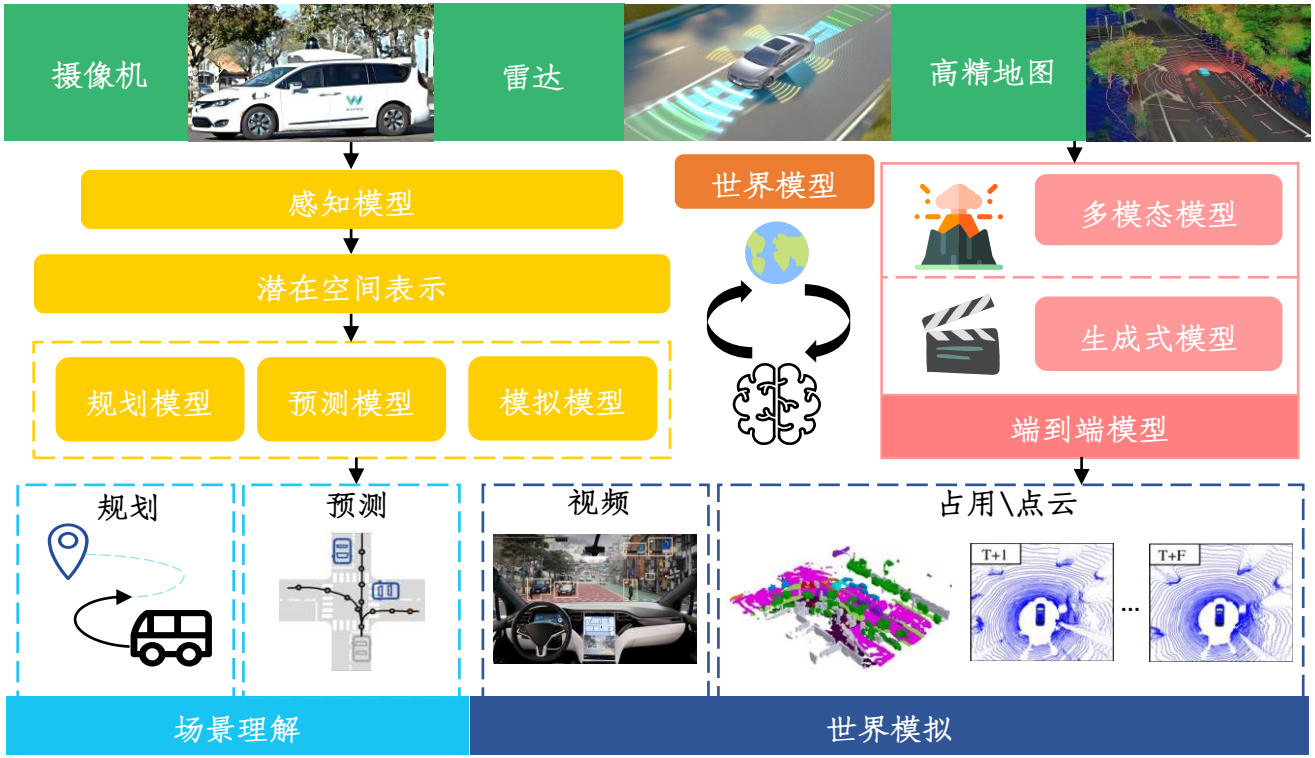
全球主流世界模型汇总

企业	模型	模型特征	应用场景	发布时间
Google	Genie 3	模拟世界物理特性、模拟物理世界、动画建模、探索地点与历史背景、实时交互性、长期环境一致性	具身智能、智能体、视频生成	2025.8
Meta	V-JEPA 2	实现最先进的环境理解与预测能力，并在新环境中完成零样本规划与机器人控制	具身智能、可穿戴AI助手	2025.6
Tencent 腾讯	混元3D世界模型1.0	3D场景生成、可导出mesh资产、场景可编辑交互、场景可漫游体验	游戏场景研发、3D世界生成	2025.7
達摩院 DAMO ACADEMY	WorldVLA	理解物理世界、动作建模、多模态交互	视频生成、具身智能、动作识别	2025.6
KUNLUN www.kunlun.com	Matrix-Game 2.0	高帧率实时交互长序列生成、多场景泛化、物理一致性	虚拟游戏世界、具身智能、影视与元宇宙内容生产	2025.8

资料来源：至顶智库结合公开资料整理绘制。

6.1 自动驾驶成为世界模型的重要落地场景

随着世界模型技术架构的不断完善，其在自动驾驶领域的应用价值也愈发凸显。世界模型通过感知、预测和规划等模块，帮助自动驾驶车辆理解和预测复杂的交通环境，从而做出可靠的决策。世界模型通过摄像头、雷达及高精地图等传感器接收实时环境数据，并借助感知模型对数据进行处理，生成潜在空间表示，为规划、预测和模拟模型提供基础，从而完成下一步驾驶的规划与预测等场景理解操作；同时，端到端模型同时处理感知数据，以视频生成的方式模拟未来可能的环境状态，以支持自动驾驶车辆的决策过程。

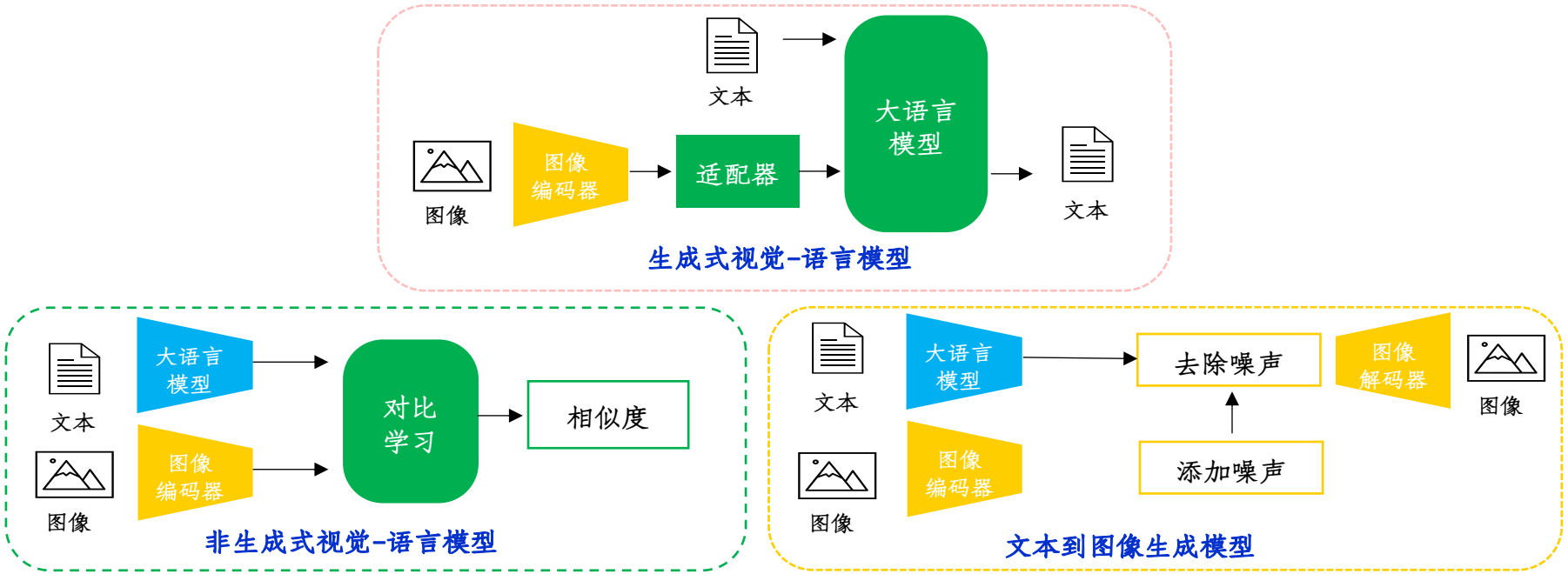


资料来源：Understanding World or Predicting Future? A Comprehensive Survey of World Models，至顶智库结合公开资料整理绘制。

6.2 多模态模型：通过整合多源数据实现协同推理

多模态模型是一类能整合文本、图像、视频等多源数据的模型。其核心在于突破单模态局限，通过跨模态交互实现信息融合与协同推理。此类模型多数依托Transformer架构，通过线性探测、稀疏自编码器等方法解析模态间关联，主要包括对比性视觉语言模型、生成式视觉语言模型及文本到图像扩散模型三大类型。在应用层面，多模态模型覆盖图像生成、视觉问答、图像检索、模型编辑、可控生成等丰富场景，能在复杂任务中展现精准干预能力。

多模态模型主要架构

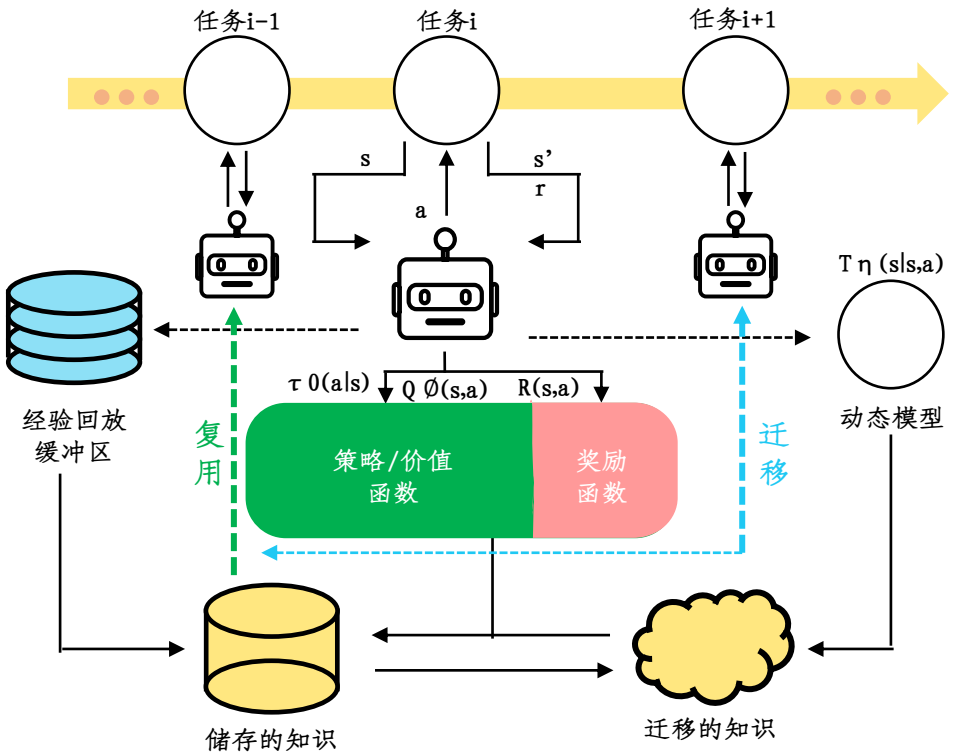


资料来源：*A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models*，至顶智库结合公开资料整理绘制。

6.3 持续强化学习：加速实现动态环境中的知识迁移

持续强化学习（CRL）是强化学习与持续学习交叉的前沿领域，旨在使智能体在动态、非平稳环境中通过序列化任务学习，避免灾难性遗忘并实现知识迁移。该方法平衡稳定性、可塑性与可扩展性，有利于最终实现类人终身学习能力。CRL方法通常围绕知识存储与迁移机制展开，主要分为四类：策略导向、经验导向、动态导向、奖励导向。

持续强化学习(CRL)结构示意图



持续强化学习(CRL)主要特征

策略导向

- 策略复用：重用完整旧策略初始化/探索新任务
- 策略分解：通过蒸馏/超网络/掩码合并多策略
- 策略合并：分解策略为共享基础+任务特定组件

经验导向

- 直接回放：存储关键经验样本重放
- 生成回放：用生成模型合成旧经验

动态导向

- 直接建模：显式学习环境动态模型
- 间接建模：用隐变量/世界模型推断动态

奖励导向

- 奖励重塑：修改奖励函数，如添加内在奖励
- 内在奖励：鼓励探索、预测误差；多奖励组合

资料来源：*A Survey of Continual Reinforcement Learning*，至顶智库结合公开资料整理绘制。

6.4 非Transformer架构：突破路径依赖的模型发展之道

随着模型规模的不断扩大和应用范围的不断拓展，Transformer架构面临诸多挑战，而非Transformer架构突围正推动模型走出一条创新发展路径。目前，非Transformer架构以状态空间模型与线性架构、液态神经网络架构、类脑与仿生架构、混合架构为代表。基于不同类型涌现出一系列性能更好、解决能力更强的模型，尤其在并行计算、推理效率上表现突出。非Transformer架构呈现混合化创新趋势，新型RNN架构崭露头角，与Transformer架构改进路线一并推动人工智能发展。

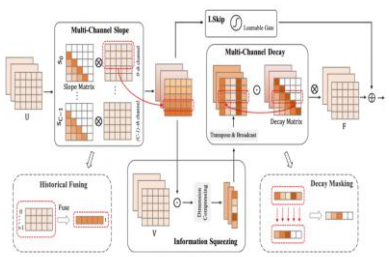
非Transformer架构发展路径

技术路径	架构名	发布方	时间	特点
状态空间模型与线性架构	Mamba-2	Carnegie Mellon University，Princeton University	2024.5	能够利用更大的状态维度，训练速度更快。在需要更大状态容量的MQAR任务上，比Mamba-1有显著改进。
	RWKV-7s	元始智能	2025.7	可实现类似MoE的优秀推理性能，同时无需占用显存，甚至无需占用内存，可让稀疏大模型真正部署到所有端侧设备。
	RetNet	MSRA	2023.7	在Transformer基础上，使用多尺度保持（Retention）机制替代标准自注意力机制，增加模型表达能力。
液态神经网络架构	LFM	Liquid AI	2024.10	基于第一性原理打造，内存使用低，推理速度快，适合解决需要深度知识背景的任务。
类脑与仿生架构	Yan	RockAI	2024.1	利用类脑激活机制与MCSD，实现比Transformer架构更高的训练效率、更强的记忆能力、更低的幻觉表达。
混合架构与轻量化方案	Hyena	Stanford University，MILA institute for AI	2023.3	由多头、分组查询注意力和排列在Hyena块中的门控卷积组成，在长上下文摘要上表现出色。

LFM架构

LFM采用与数学和信号处理领域相似的数学原理，利用液态神经网络，适合处理多种类型的数据，包括文本、音频、图像和视频。LFM能够仅用很少的数字神经元就实现高效的运作，相比其他大模型，如ChatGPT，在计算芯片的使用上更为高效。

Yan架构



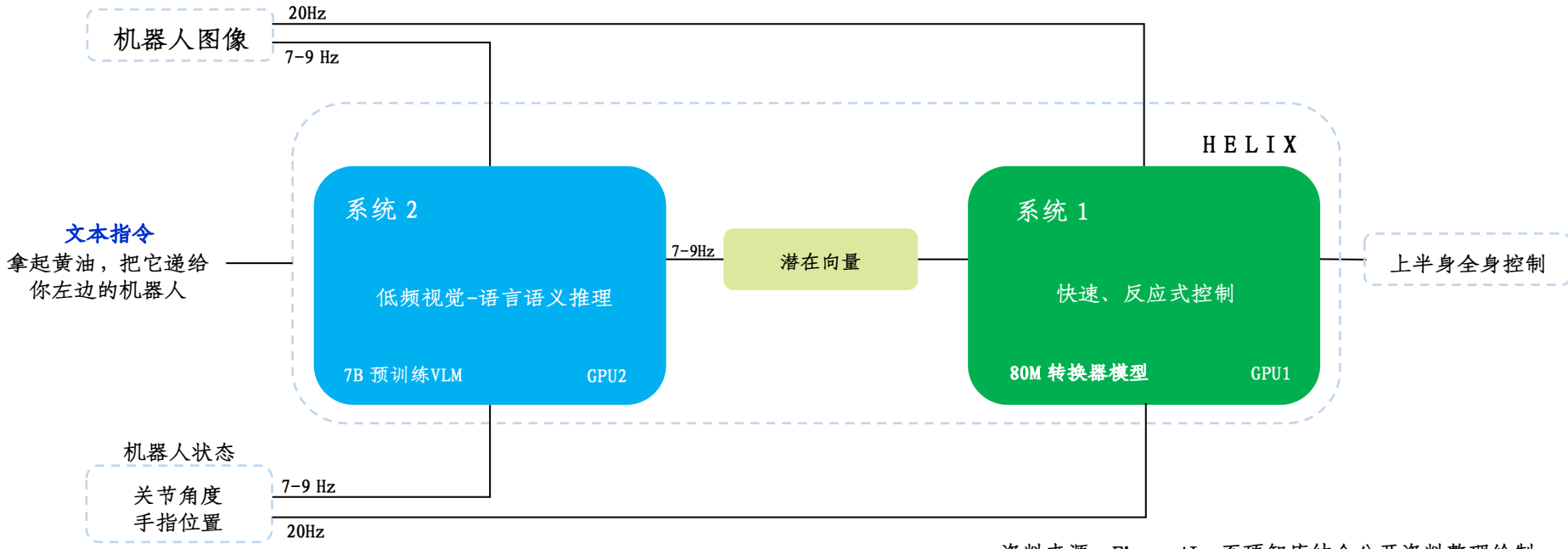
Yan架构不依赖于传统的注意力机制或RNN等序列模型。用神经元选择激活以及MCSD两个模块替换Transformer中的Attention机制，并在训练效率、推理吞吐量、资源消耗和记忆能力等多个维度上均优于传统Transformer模型。

资料来源：至顶智库结合公开资料整理绘制。

6.5 Figure AI Helix：实时协作的“慢思考-快执行”具身智能架构

Helix是Figure AI提出的首个“双系统”视觉-语言-行动（Vision-Language-Action）模型，用于实现人形机器人上半身的高频灵巧控制。其由两个核心部分组成：系统2（System2）是一个基于70亿参数的预训练视觉-语言模型，对环境图像和自然语言指令进行场景理解与语义推理，并将关键信息压缩为一个连续的潜在向量；系统1（System1）是一个约8000万参数的视觉-运动转换器模型，将来自系统2的潜在向量与机器人状态结合，输出包括手腕姿态、手指动作以及躯干与头部控制在内的连续上半身动作。这种解耦的架构使得系统能够在不同时间尺度上最优运行：系统2负责“慢思考”的高层目标推理，而系统1实现“快反应”的实时动作执行。

Figure AI Helix：双系统VLA架构

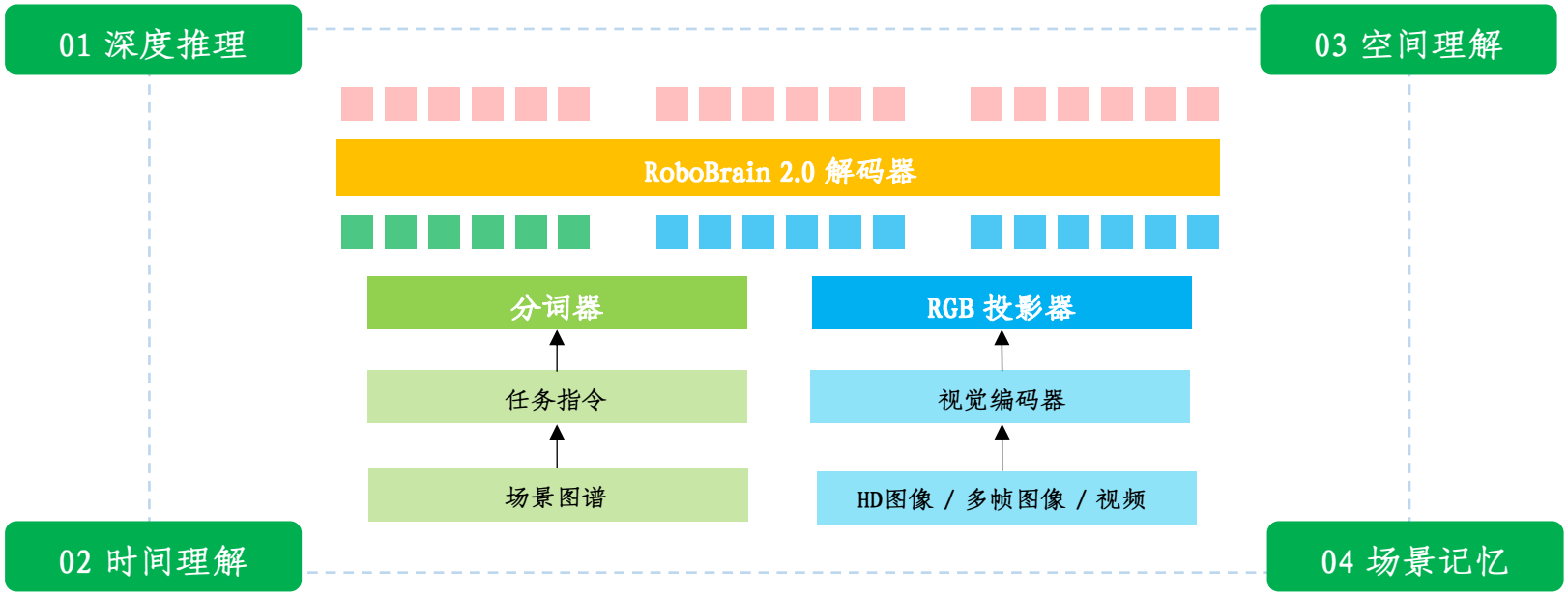


资料来源：Figure AI，至顶智库结合公开资料整理绘制。

6.6 智源RoboBrain 2.0：四大核心能力支撑长时任务执行

RoboBrain 2.0是智源最新发布的具身大脑，多项关键能力较前代均有显著突破。其在长时任务规划中能够将复杂任务拆解为多个子任务，并通过闭环反馈与监控实时检查完成情况，必要时进行重新规划。在空间理解上，新增复杂空间关系和距离推理，支持点指令与框选指令，显著提升可操作区域识别和轨迹生成的准确度，整体性能提高超过17%。在时间理解上，通过轨迹预测增强动作执行的连续性与精确度。在场景记忆方面，能够对环境中物体的位置和属性进行构建与更新，为动态环境下的持续操作提供支持。同时，RoboBrain 2.0还具备多机器人、多环境协同规划能力，可以实现跨场景泛化与自主执行，为具身智能的发展奠定更加坚实的技术基础。

智源RoboBrain2.0架构与四大核心能力



资料来源：智源研究院，至顶智库结合公开资料整理绘制。