# 大模型驱动的汽车行业群体智能技术白皮书

清华大学自然语言处理实验室 | 易慧智能 | 面壁智能

**正式版**

**大模型落地前夜 · 业内首次发布**

每一个汽车从业者需要了解 · 专属汽车行业的群体智能白皮书

# 前序




## 孙茂松

清华大学计算机科学与技术系长聘教授，
欧洲人文和自然科学院外籍院士，ACL Fellow


诞生于 1956 年达特茅斯会议的「人工智能」，是人类璀璨文明史中最年轻的学科之一，但她自诞生以来就不断与其他学科、行业深度交融，对人类社会产生日益深远的影响。2023 年以来大语言模型（LLM）技术进一步加速了这个进程，特别是最近，大模型驱动的智能体成为业界公认的赋能技术。为此，我们团队提出了岗位孪生、业务孪生和组织孪生的概念和技术框架，旨在综合运用大模型的通用能力和智能体技术的灵活适配特性，打通大模型赋能行业应用提质增效的最后一公里，实现智能科技服务人类。本次发布的白皮书，深入浅出地介绍了大模型驱动的智能体技术，特别是面向汽车行业提出体系化解决方案，对于未来通用人工智能赋能汽车行业提供了有益参考。智能科技日新月异，需要主动拥抱变化、勇于探索的先行者，弄潮儿向涛头立，让我们共同努力携手迎接即将到来的通用智能时代。


## 刘知远

### 清华大学计算机科学与技术系副教授


当前，人工智能正从学术领域跨越到实际应用的新阶段，大模型驱动的群体智能技术正成为推动革新的核心动力。目前大模型已能够构建出更具通用性和适应性的智能体，这些智能体不仅能独立执行复杂任务，还能在群体中协同作业，展示出远超单体智能体的集体智慧。在汽车行业，群体智能的应用不仅能够大幅提升生产效率，优化用户体验，更是孕育新的商业模式。

清华大学自然语言处理实验室，长期深耕自然语言处理的前沿核心技术研究，在大模型、AI Agent、群体智能等方面取得了系列具有国内外影响力的学术和实践成果。此次联合易慧智能、面壁智能发布这本白皮书，系统地阐述了大模型技术在汽车产业的应用前景和实践路径，为我们提供了一个全面、深入的视角。

我们正步入一个由大模型驱动的“Internet of Agents”智联网时代，这个时代将由智能体的群体协作和互动定义，它们不仅服务于人类，更将与人类共创更加智慧和可持续的未来。我们期待通过这本白皮书，与业界同仁共同探索和实践，将大模型群体智能的潜力转化为现实，共同推动汽车行业迎接智能化的挑战和机遇，开创智能汽车行业的崭新篇章。

## 李伟

### 易慧智能总裁

在时代变革的浪潮中，汽车行业正经历着前所未有的机遇与挑战。用户需求日益多样化，触点愈发丰富，这一切都要求我们必须以前瞻性的眼光和创新的手段来应对。幸运的是，我们处在一个AI技术飞速发展的时代，大语言模型为我们提供了理解和应对这些挑战的全新视角。

易慧智能是汽车行业领先的 AI 产品与业务解决方案提供商，拥有庞大的用户群体与深厚的行业影响力，依托于海量行业知识库、用户数据和丰富的行业场景认知为基础，利用大数据和AI技术深度赋能汽车产业链，不断完善服务于汽车全产业链的能力，助力车企与经销商集团实现智能化转型，提升客户在行业内的整体竞争力。易慧智能凭借卓越资源背景，融合学术、技术实力派巨头——联合清华大学自然语言处理实验室与面壁智能，致力于将最尖端的 AI 科技及学术研究与最佳的技术落地业务实践相结合，打造大模型驱动的群体智能协同平台，为汽车企业提供群体智能与组织孪生解决方案及一站式的运营实施服务。

这本白皮书是我们对群体智能和组织孪生技术在汽车行业应用的务实探索，也是我们帮助汽车行业全面实现智能化的具备开拓性的一步。

## 肖超

易慧智能 COO

回顾过去的一年，汽车行业经历了白热化的竞争与疯狂内卷，但是如何深入洞察用户的真实需求，做到数字定义产品；在面对用户在纷繁媒体触点下的失焦，面对新媒体即时互动的沟通品牌一致性和用户体验如何保障；用户被拉长的购买决策周期，潜客孵化过程中如何降低各种不确定性因素导致的流失等等；都是在产品定义和营销创新方面亟须创新的话题。

随着 ChatGPT 的发展全球都掀起的智能化热潮，2024 年更被行业誉为 AI 应用的元年。基于单体 Agent 已经具备规划、执行、感知、记忆和工具使用的能力，多个 AI Agent 协同在人类目标设定下、提供必要的数据和算力资源，已经可以完成复杂任务，我们看到了 AI Agents 群体智能协同工作在行业应用中场景丰富前景广阔。

对汽车行业拥有丰富场景认知和适用性判断的易慧智能，以 AI Agents 的汽车行业应用为切入点，将领先的 AI 科技与最佳的业务实践相结合，为客户定制的大模型驱动的群体智能协同工作平台，提供精研的组织孪生解决方案和卓越的 AI 数字员工运营服务，探索 AI 技术加持下降本增效的新路径，助力汽车行业的企业实现智能化发展的最后一公里。

## 李大海

面壁智能 CEO

以大模型为核心的 AGI 革命将会成为人类历史上第四次重大技术变革，这场变革可以和蒸汽革命、电力革命乃至信息革命相提并论，并将持续至少 20 到 30 年的发展时间。面壁智能始终相信，人工智能将深刻改变我们的世界。若干年后，整个人类社会的生产和生活将会因 AGI 革命的演进而发生翻天覆地的变化。

如今，人类拥有大模型这种智能工具，却面临场景定义不清、行业落地困难的诸多痛点，而智能体则是解决行业落地痛点的最重要手段。大模型好比汽车引擎，一辆汽车除了引擎，还需要底盘、转向系统、内饰以及其它所有的必要组件才能上路。为发挥好大模型的潜力，面壁为大模型这个“引擎”挂载了一系列前沿技术，如工具学习、超长记忆等，甚至通过多智能体协作的方式提高智能体的能力边界，探索 AI 赋能人类的最优解。

面壁智能致力于探索大语言模型、AI Agent、群体智能和数字组织孪生技术的最新发展趋势及其对汽车行业的深远影响。大语言模型作为人工智能领域的一大突破，提供了更加强大和精准的语言理解和生成能力。结合 AI Agent 技术，这些模型不仅能够处理复杂的自然语言任务，还能在更广泛的场景中自动执行任务，提高决策效率。

我们非常荣幸拥有易慧智能这样一位在汽车营销领域积累深厚又不断进取的合作伙伴，并有幸分享来自清华大学自然语言处理实验室最前沿的技术视野。这本白皮书阐述我们在汽车营销领域前沿的智能体场景应用探索，期待能为 AGI 事业的发展贡献一份力量。

# 目录



## 前序

孙茂松（清华大学计算机科学与技术系长聘教授，欧洲人文和自然科学院外籍院士，ACL Fellow）

刘知远（清华大学计算机科学与技术系副教授）

李伟（易慧智能总裁）

肖超（易慧智能COO）

李大海（面壁智能CEO）

# 前言

随着科技的飞速发展，汽车行业正面临着颠覆性的变革。从传统的燃油车到电动汽车，从手动驾驶到自动驾驶，从机械座舱、电子座舱到智能座舱，每一次的技术突破都在推动着汽车行业的进步。在智能化、网络化、电动化的趋势下，汽车不仅仅是一种出行工具，而是一个承载了众多创新技术的移动智能终端。在发展与变革的过程中，大语言模型和群体智能对车企在生产、销售、营销等各环节均带来前所未有的机遇和挑战，群体智能与组织孪生解决方案也从纸上谈兵变为可在行业中实现推进落地。

群体智能技术的发展，为汽车行业带来了新的机遇。通过多个智能体的协作，可以处理更加复杂和动态的任务，如智能交通系统的优化、车辆群的协调运行等。这不仅提高了汽车行业的运营效率，也为用户提供了更为丰富和智能的服务。

此外，数字组织孪生技术的应用，为汽车行业带来了革命性的变革。通过创建数字孪生模型，企业能够在虚拟空间中模拟和分析研发、生产与营销流程，从而实现更高效的资源配置和风险管理。这项技术在产品设计、生产过程优化、以及市场策略制定等方面都显示出巨大的潜力。

本白皮书全面探讨了大模型群体智能技术及其在汽车行业的应用潜力。首先，我们在第一章分析了中国汽车行业的市场现状，聚焦于消费需求的变化、供给侧的挑战以及由此产生的机遇。接着，我们在第二章深入探讨了大模型群体智能技术体系，包括大语言模型、AI Agent、群体智能和组织孪生。第三章着重于分析大模型群体智能技术在汽车行业的应用价值和实践案例。最后，在第四章详细描述了汽车行业群体智能生态矩阵及其共赢逻辑，并以对未来的展望作为总结，强调了这些技术对于汽车行业转型升级的重要性。

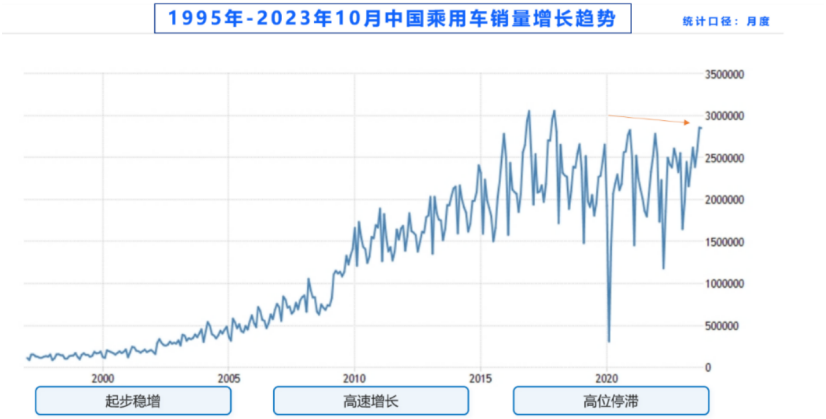
第一章

# 战略态势：人工智能时代的汽车行业发展

## 1.1 中国汽车行业市场现状

### 1.1.1 中国乘用车市场需求节奏放缓但总量处于高位

中国汽车行业发展近 30 年经历了“起步积累”、“高速增长”、“高位停滞”三个发展阶段。20 世纪七八十年代，中国对轿车实行严格管控政策，年销量仅 20 万台，需求和供给缺乏弹性，价格无法自由调控。尽管增长速率缓慢，但市场基数小，增长空间大。20 世纪九十年代进入起步积累阶段，政策上放宽对私人车市的管控，不得通过行政和经济手段干预个人购车，价格机制逐步发挥调控作用。进入 21 世纪，国家鼓励轿车进入家庭，市场价格成为调控需求与供给的核心机制。21 世纪前二十年，80、90 后见证了中国私家车市场的爆发式增长，并在 80、90 后等人群中迅速普及，从 2001 年的 236 万辆井喷式增长至 2017 年 2887.9 万辆。2018 年起销量开始转升为降，2020 年受新冠疫情的冲击对整个汽车行业产生了重大影响，打击了汽车的生产和消费。2021 年车市回温销量略有增长，随后两年呈现缓慢下降趋势。易车研究院在 2023 年车市价格战洞察报告中调研发现，汽车市场需求放缓、供给效能提升带来了 2023 年激烈的“价格战“，一季度销量同比下滑 13.66%。基于汽车的商品交易属性，交易价格由供给和需求决定，这突显出当前的供需矛盾 [1-1]。



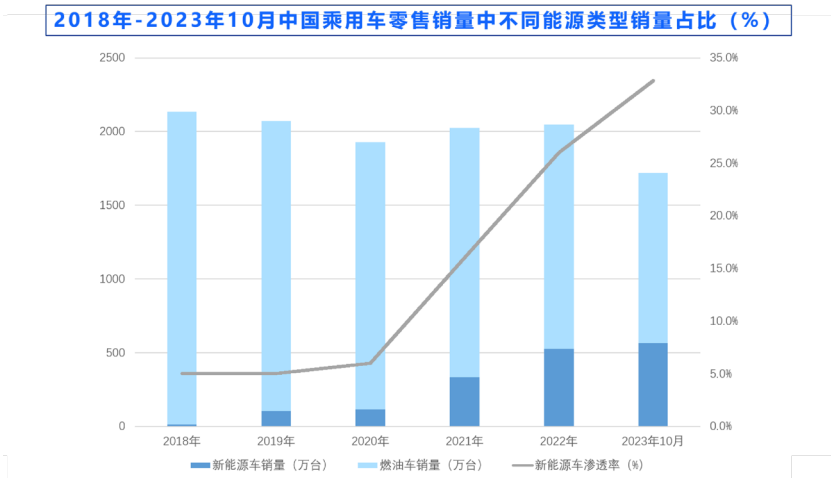

数据来源：全球经济指标

图表 1-1 1995 年 -2023 年 10 月中国乘用车销量增长趋

### 1.1.2 新能源市场逆势上扬

虽然中国乘用车市场整体处于需求增长停滞的大环境中，但细分的新能源车市场表现越加醒目。2023 年新能源汽车市场渗透率突破 30%，提前实现了《新能源汽车产业发展规划 (2021-2035)》中关于 2025 年新能源新车销量达到新车总销量 20% 的目标，已经成为我国汽车行业弯道超车的重点赛道。国家政策的扶持给新能源汽车发展带来众多有利条件，财政部、税务总局、工信部在 2023 年 6 月联合发布的《关于延长和优化新能源车辆购置税减免政策的公告》，将新能源汽车车辆购置税减免政策延长 4 年至 2027 年 12 月 31 日；2023 年 10 月由科技部发布的《关于支持新能源汽车产业高质量发展的若干政

策实施》等一系列政策的颁布，推动了新能源汽车市场繁荣发展、刺激消费需求，旨在推动汽车产业的技术研发、创新、转型和升级。



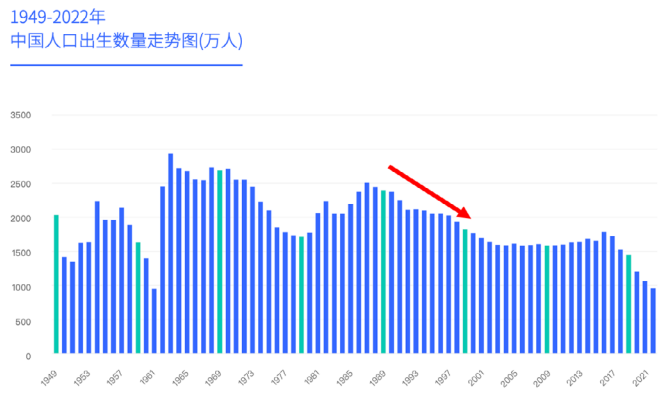

数据来源：全国乘用车市场信息联席会，统计整理

图表 1-2 2018 年 -2023 年 10 月中国乘用车销量增长趋势

## 1.2 汽车市场需求侧洞察

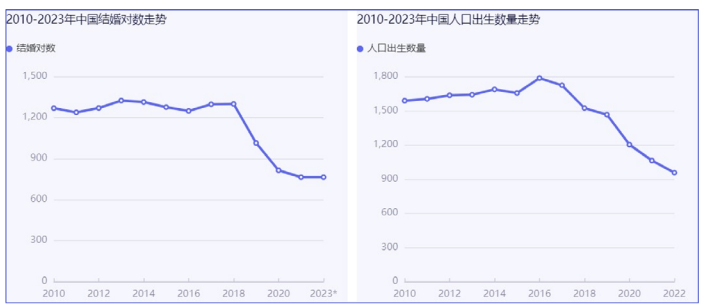
### 1.2.1 人口结构的变化，影响整体购车人群减少

易车研究院调研发现，结婚、生子是中国老百姓的关键购车需求节点 [1-1]。2008 年开始，80 后的“结婚购车浪潮”是中国车市（特别是首购车用户）的主要推动力。2018 年后，90 后开始大规模进入车市，90 后人数减少购买潜力不及 80 后；结合图表 1-4 我们发现近五年大家对结婚和生子积极性持续走低，一定程度上降低了首购车人群需求。




数据来源：国家统计局（统计口径），数源整理：易车研究院《2023 年车市价格战洞察报告》[1-1]

图表 1-3 中国人口走势图表



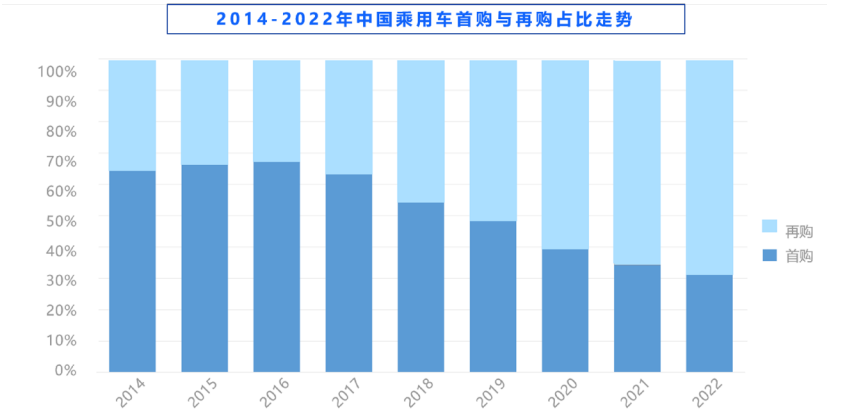

数据来源：国家统计局（统计口径），数源整理：易车研究院《2023 年车市价格战洞察报告》[1-1]

图表 1-4 2010 年 -2023 年中国结婚对数和出生人口走势图表

## 1.2.2 首购车用户减少，再购车用户比例增加

首购车用户呈现下降趋势，再购逐渐成为核心增长动力且均价有所提升，给中高端车型带来更多机会。根据易车研究院 2023 年《家庭拥车数量洞察报告》2014 年至 2022 年首购和再购的数据，再购市场有较大的潜力空间 [1-2]。同时根据麦肯锡 2023 年中国汽车消费者

调研，有 54% 的受访者表示在再购车时考虑升级价格区间 [1-3]，促使中高端汽车市占率的提升。




数据来源：易车研究院（以家庭为单位）《家庭拥车数量洞察报告》[1-2]

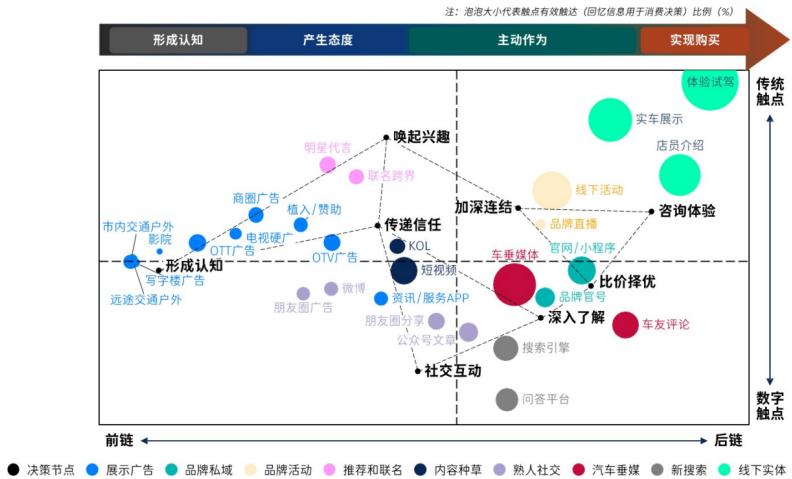
图表 1-5 中国乘用车首购与再购占比走势图表

### 1.2.3 新媒体时代用户获取信息触点多、注意力碎片化

近年来消费者获取汽车资讯呈现多渠道、多触点的特点，在常态化触媒包围下，品牌主都在想方设法地抢占用户注意力。群邑联合易车发布的《2023 全域链路时代汽车营销变革白皮书》中提到，整个用户消费旅途中涉及多达 29 个消费触点 [1-4]。

用户注意力从原先聚焦于汽车垂直资讯平台与汽车厂商官网，持续且不同程度地分散到各个泛娱乐类短视频平台、知识分享及社交媒体平台、搜索引擎、新闻资讯平台和综合视频平台等。厂商需依据各平台的用户画像和推送逻辑，不断向用户推送车系种草内容或竞品拦

截信息抢占用户注意力。从被动获取信息到主动筛选、糅合信息，实际延长了用户从形成认知到产生购买行为的时间，用户注意力被分散的同时也增加了转化难度。因此，企业亟待信息整合，为用户提供高效精准的内容，打造品牌认知的一致性。




信息来源： 群邑，易车《2023 全域链路时代汽车营销变革白皮书》

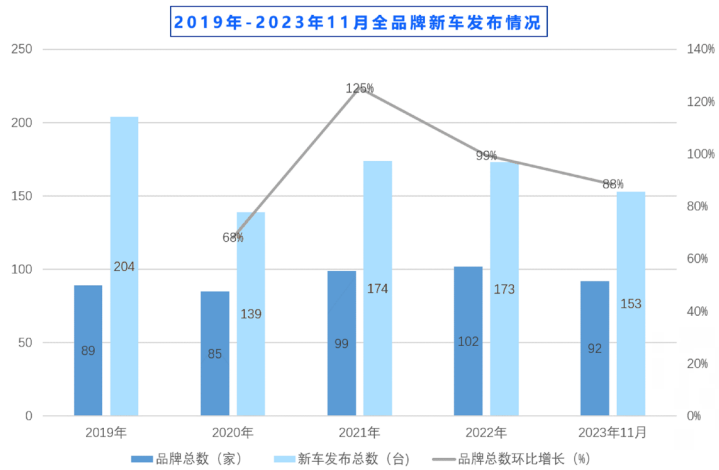
图表 1- 6 汽车营销场景媒介触点与决策点关系图表

## 1.3 汽车市场供给侧洞察

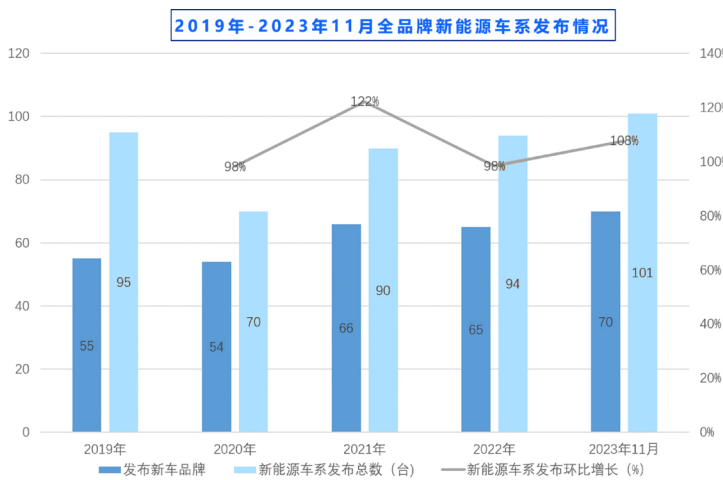
### 1.3.1 品牌格局呈现群雄纷争，新入局者有机会打造全新市场格局

2019-2023 年汽车品牌与车型迅猛增长，在新产品数目不断扩张的同时，旧有格局也悄然发生变化，给新入局者提供了发展机会，也为汽车市场注入了新的活力。快速涌入的新产品给消费者更多的选择

空间，满足不同消费者多样化的需求，其中新能源品牌近年来在汽车市场上表现抢眼。（参考图表 1-7、1-8）




图表 1- 7 汽车市场品牌总数和发布新车发布总数走势图表




数据来源：各品牌官方发布渠道，统计整理

图表 1- 8 新车发布品牌中新能源车系发布总数走势图表

近 5 年来整体市场份额波动较大，行业洗牌加速且尚未形成稳定格局，恰好是新入局玩家凭借敏锐的市场洞察和创新能力，在市场中迅速崛起的好时机。如图表 1- 9 展示近 5 年中国乘用车品牌销量 TOP10 中，有 5 家连续 5 年跻身前十榜单。


数据来源：全国乘用车市场信息联席会，统计整理

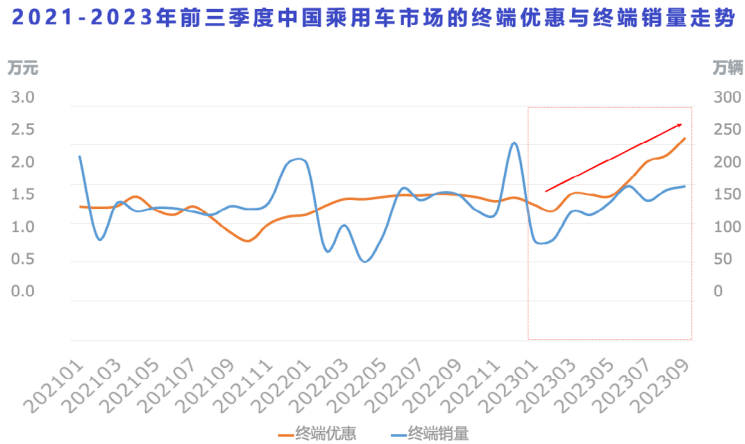
图表 1- 9 新车发布品牌中新能源车系发布总数走势图表

## 1.3.2 价格内卷带来经营利润下滑

我们认为“价格战”是把双刃剑，企业可以利用价格优惠吸引消费者注意，在激烈的市场竞争中快速抢占份额，但同时也会压缩部分利润空间。从 2021 年 -2023 年 9 月中国乘用车市场终端优惠与终端销量走势图表，不难发现优惠幅度与销量基本呈正比。

2021-2022 年，中国乘用车市场每辆车的平均优惠幅度在 1.5-2 万元之间。2023 年初由特斯拉率先打响价格战，最高降幅超过 4.8 万，

随后众多新能源品牌和传统车企也纷纷跟进，通过降价、限时促销等方式来吸引消费者。2023 年二季度末，平均每辆车的终端优惠突破了 2 万元，三季度末更是逼近了 2.6 万元，这种优惠的规模是前所未有的。




数据来源：易车车型库，出处：易车研究院，《2023 年车市价格战洞察报告》[1-1]

图表 1- 10 2021 年 -2023 年前三季度中国乘用车市场终端价格与销量走势。

销量提升并不等同于企业利润提升。 2023 年上半年，从国内 10 家上市车企业对外公布的财报数据看，多数车企上半年营收、净利润均呈现上涨趋势，但也有 3 家公司归母净利润同比出现下滑。

| 2023上半年国内上市车企财报 | | | | | | |
|---|---|---|---|---|---|---|
| 上市车企 | 总营收（亿元） | | | 归母净利润（亿元） | | |
| | 2023上半年 | 去年同期 | 同比 | 2023上半年 | 去年同期 | 同比 |
| 上汽集团 | 3,265.55 | 3,159.93 | 3.34% | 70.85 | 69.10 | 2.54% |
| 比亚迪 | 2,601.24 | 1,506.07 | 72.72% | 109.54 | 35.95 | 204.68% |
| 北京汽车 | 990.47 | 836.79 | 18.37% | 28.46 | 21.58 | 31.85% |
| 吉利汽车 | 731.82 | 581.84 | 25.78% | 15.71 | 15.52 | 1.22% |
| 长城汽车 | 699.71 | 621.34 | 12.61% | 13.61 | 56.01 | -75.69% |
| 长安汽车 | 654.92 | 565.74 | 15.76% | 76.53 | 58.58 | 30.65% |
| 广汽集团 | 651.88 | 484.48 | 27.12% | 29.66 | 57.51 | -48.42% |
| 理想汽车 | 474.40 | 182.95 | 159.31% | 32.23 | -629 | - |
| 东风集团股份 | 456.77 | 443.96 | 2.89% | 12.70 | 55.00 | -76.91% |
| 一汽解放 | 330.15 | 228.72 | 44.35% | 4.01 | 1.70 | 135.87% |

数据来源：上市公司车企财报公开信息整理

图表 1- 11 2023 年上半年中国主流上市车企业财务状况

2023 年上半年，经销商集团受到价格战影响，亏损面积增大。降价销售新车压缩了利润空间，毛利润和毛利率均不及 22 年同期表现。也有部分消费者为搭乘购置税减半的福利，在 22 年底前提前透支了消费需求。综合因素使得经销商经营压力进一步加大。

| 2023上半年经销商集团毛利润、毛利率对比 | | | | | |
|---|---|---|---|---|---|
| 经销商集团 | 毛利润（亿元） | | 毛利率（%） | | |
| | 2023上半年 | 2022上半年 | 2023上半年 | 2022上半年 | 同比（%） |
| 中升集团 | 69.7 | 84.67 | 8.5% | 9.8% | -1.3% |
| 永达汽车 | 34.23 | 36.76 | 7.5% | 9.8% | -2.3% |
| 广汇宝信 | 9 | 13.3 | 5.7% | 9.1% | -3.4% |
| 正通汽车 | 6.3 | 9.5 | 5.1% | 8.6% | -3.5% |
| 和谐汽车 | 5.8 | 6.9 | 7.2% | 8.7% | -1.5% |
| 新丰泰集团 | 2.66 | 4.16 | 5.1% | 8.1% | -3.0% |
| 美东汽车 | 10.01 | 13.34 | 7.1% | 10.5% | -3.4% |
| 广汇汽车 | 61.48 | 65.99 | 9.2% | 10.0% | -0/8% |
| 国机汽车 | 14 | 13.3 | 6.5% | 7.8% | -1.3% |
| 世纪联合控股 | 0.33 | 0.53 | 4.6% | 6.9% | -2.3% |

数据来源：经销商集团财报公开信息整理

图表 1- 12 2023 年上半年中国部分经销商集团利润状况

### 1.3.3 直营模式成为用户运营新抓手，降本增效正当时

打造行业领先的成本优势、实现一致的品牌体验持续影响用户心智，是企业经营效能提升的两个关键胜负手。

易慧智能实地走访汽车销售门店发现，人力成本居高不下，引入传统工具化应用也并未带来预期的经营效能提升，此外，人员服务专业度问题及为保障品牌一致性带来了大量额外成本问题，是经销商与品牌直营店面临的普遍挑战。品牌在销售模式上，正加速从主流经销商模式到直营模式，再到混合经营模式进行积极探索，以达到降本增效的目的。

经销商模式通过经销商网络销售和服务车辆，仍是当前汽车厂商的主流销售渠道。自负盈亏的经销商模式，具备覆盖性广、细分性强等特点。但也会导致恶意竞争，服务水平良莠不齐。在我们实地走访中，经销商门店人工邀约试驾，仍是潜客孵化的主要手段。面对严苛的邀约数量、服务通话质量和转化率考核，经销商顾问在有限精力内仅能做到应付考核，对中低意向的客户基本放弃维护，导致大量潜客流失。即便专业类应用工具越来越多，但学习成本极高，多半是摆设，主机厂无法获得用户真实数据反馈，难以带来经营效益及效率提升。

新势力品牌入局多采用直营模式，通过品牌 APP 报价 / 下单、设立自营交付中心，没有中间商赚差价可以有效的控制价格和利润，全链路对接终端消费者，有效保障了品牌服务的一致性，优异的线下体验对促成购车不可或缺。与用户直联的环节中，厂商可以更加准确有效的掌握消费者的第一手信息、迅速获得产品反馈，帮助企业快速进行产品迭代。特别在品牌建立初期，直营模式利于品牌形象打造、提

升品牌知名度。而在发展阶段，门店建设速度跟不上下沉市场需求节奏，抢占市场份额缓慢，销量目标难以达成。且直营店多部署在核心街区，需要极高的运营成本投入在门店建设和人员培训，无疑缩小了品牌的利润空间。

基于直营模式中的实践问题，部分企业不再执着于“纯粹基因”开始对直销模式进行创新，为了最大程度提升销量向混营模式转型，引入经销商集团或代理商来提供交车和售后服务。混营模式下，品牌制定周到一致的服务和价格标准，可以杜绝经销商“偷工减料”或恶意降价。经销商的加入可以快速实现门店下沉，在激烈的竞争环境下抢占市场份额，主机厂也可以将资金投入到如核心技术研发等利润回报率更高的领域。

尽管企业经营效能提升注重人才培育和强化，但总会触碰到成本和效率的天花板。因此，企业亟需在全链路运营中打造领先行业的成本优势，从而获得更大的利润空间，同时为消费者提供有竞争力的价格。把握直连用户的契机，提供高质量的标准化服务，加强用户品牌心智建设。


图表 1-13 主流品牌销售模式列举

## 1.4 机遇与挑战

基于汽车行业加速内卷的市场竞争和消费者需求放缓的市场背景，人工智能正在重塑汽车行业的生态，对汽车企业智能化转型而言是挑战更是机遇。

**全面的成本领先是未来汽车企业竞争的基础。**在激烈的市场竞争中，汽车企业需要通过全面的成本领先策略来降低生产成本、提高运营效率，从而获取竞争优势。在人工智能时代，自动化和智能化生产成为主流趋势，这有助于降低汽车企业的生产成本和提高生产效率。例如，通过引入自动化生产线和智能仓储管理系统，汽车企业可以减少人力成本和库存成本，从而实现更高效的生产管理。同时，企业需持续投入大量资金进行技术研发和人才引进，关注全球产业链的变化，积极寻求与供应商和合作伙伴的协同降本机会，建立完善的 AI 基础设施以实现降本增效。

**一致的品牌体验和个性化的品牌沟通会成为品牌心智塑造的胜负手。**借助 AI 技术，企业可以更深入地了解消费者需求，提供个性化的品牌沟通和一致的品牌服务体验，来满足消费者对品质和服务的基本需求，从而塑造出可信赖的品牌形象。例如，根据消费者的购车习惯和偏好，为其推荐合适的车型和配置。同时，企业可以通过 AI 技术优化客户服务中心，提供高效、专业的咨询服务，提升用户满意度和用户粘性。然而，在保持品牌一致性的同时满足消费者的个性化需求，这需要企业具备精准的市场分析和精细的产品规划能力。此外，企业还需面对数据安全和隐私保护的挑战，确保消费者数据的安全与合规使用。

数据驱动的解决方案与精细化运营突破人效天花板。随着科技的进步和消费者需求的变化，传统的以“人”为中心的运营方式已经难以适应市场发展的需求。在人工智能时代，数据成为企业的核心资产。通过对数据的收集、分析和挖掘，企业可以洞察市场趋势、优化产品设计、提升服务质量。例如，利用 AI 算法分析消费者行为数据，预测未来市场趋势，提前布局产品研发。同时，企业可以通过精细化运营提高人效，降低人工成本。例如，利用 AI 技术优化人力资源管理，实现人才的精准招聘与培养。然而，数据驱动的解决方案与精细化运营也对企业提出了新的要求。企业需构建完善的数据收集和分析体系，确保数据的准确性和完整性；企业需加强数据安全保护，防止数据泄露和被滥用。

第二章

# 科技突破：迈向通用人工智能的大模型群

## 2.1 体系框架

自 2017 年 Transformer 提出之后，预训练语言模型（Pre-trained Language Model，PLM）异军突起，不断刷新各类 NLP 任务的性能上限。随着技术发展，大规模与训练语言模型参数数量不断快速提升，模型能力也飞速跃升，2022 年底，随着 ChatGPT 的发布，人们广泛意识到大模型对技术和生产力带来的无限潜力，开始讨论大语言模型是否产生了智能的“涌现”，研究基于大语言模型应用到生产生活领域的具体方法。

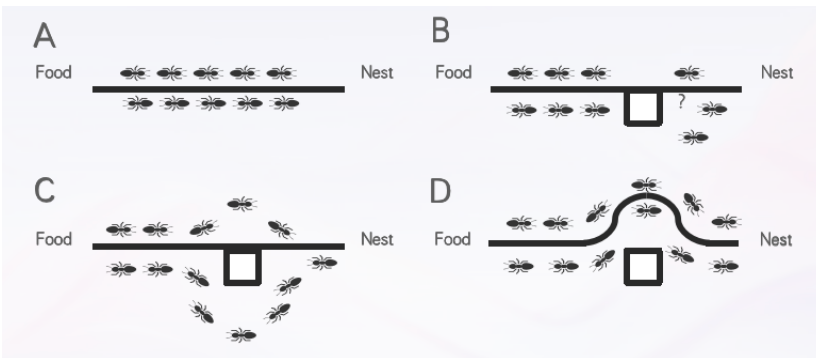
在当下，大模型技术路线已在产业界达成广泛共识，但究竟它将成为类似 Web3.0 的技术浪潮，还是一场足以绵延至少十年的产业革命，仍是一个值得深思的问题。以大模型为核心的 AGI 革命是第四次重大技术变革，它可以和蒸汽革命、电力革命、信息革命相提并论，并将持续至少 20 到 30 年，深刻改变我们的世界。若干年后，整个人类社会的生产和生活将会因 AGI 革命的演进而发生翻天覆地的变化。

如今，各行各业已清晰认识到大模型在应用中的广阔前景与价值，

然而，如何才能发挥出大模型的巨大潜力并推动生产力的发展和变革？我们可以将大模型比作汽车引擎，它为汽车提供动力。然而，要制造出一辆完整的汽车，除引擎外，还需要转向系统、底盘、内饰以及其他所有必要组件。同样，要充分发挥大模型的潜力，我们还需要在这个“引擎”基础上加入一系列高级技术，如增强的记忆能力和使用工具的能力，这样才能开拓更广泛的应用领域和想象空间。而 AI Agent（智能体）正是集合这些技术能力的载体。随着针对大语言模型的广泛研究，人们发现大模型目前存在“幻觉”等问题，导致在真实场景中落地困难。鉴于此，能够调用工具，进行复杂任务规划、执行的 Agent 技术，逐渐进入人们研究的领域。AI Agent 的出现开启了一种新的交互方式。不再是被动的执行工具，它能主动感知环境并动态响应，标志着人类智能理解的主动转变。这一创新是迈向全面人工智能（AGI）的关键步骤，反映了从传统工具使用方式向智能实体的转变。

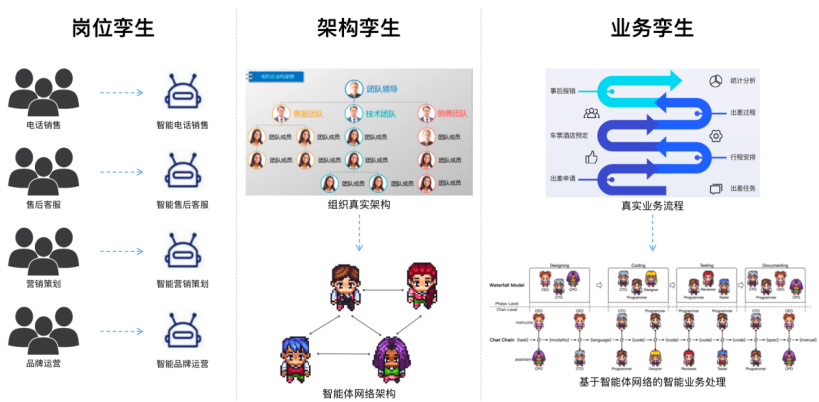

图表 2-1 AI Agent 的能力特点

智能体被定义为具备六个关键维度特征：个性化设定、智力水平、情感智能、感知能力、价值观念和成长潜力。这些特征使它们能够适应多种应用场景。为使单个智能体发挥出色的能力，需要让它们相互连接并协作，以处理和完成更为复杂的任务。实际上，无论是人类社会还是自然界，群体智能的案例比比皆是。正如我们需要团队和组织将个人联合起来一样，自然界中的蜂群、蚁群和鱼群也展示出超越个体的高级智能行为。简单个体聚集成群体时，个体间交互能够使群体涌现超越个体的智能。随着研究的深入，AI Agent 相互间，能够通过通信形成协作，完成单智能体无法完成的工作。结合能够自主理解、规划、执行、反思任务的 AI Agent 技术，群体智能的出现，大大拓展了大模型能力的上限。




图表 2-2 当蚁穴与食物的通路上出现障碍时，蚁群能够分头探索新路径，并最终采用最短路径

结合 AI Agent 和群体智能技术，我们提出了企业大模型落地的范式：组织孪生。组织孪生是一个以数字技术为核心的创新框架，它包括三个关键部分：岗位孪生、架构孪生和业务孪生。岗位孪生利用大模型技术创建个人的数字孪生虚拟人，这些虚拟人能模拟真人的交流方式，包括声音和表情，并具备“感性智能”。它们能够执行内容生成、基础交流、客户服务等工作。架构孪生则是在数字世界中映射真实公司的组织架构，通过智能体网络技术定义智能体间的交流和逻辑。最后，业务孪生通过整合大语言模型、搜索增强技术和智能体构建等，自动执行实际业务，优化业务执行效果。这个框架特别适用于复杂的行业场景，如汽车行业，提供了一个全新的数字化工作和管理方式。




图表 2-3 大模型驱动的组织孪生解决方案

# 2.2 大语言模型

## 2.2.1 大语言模型的发展演进

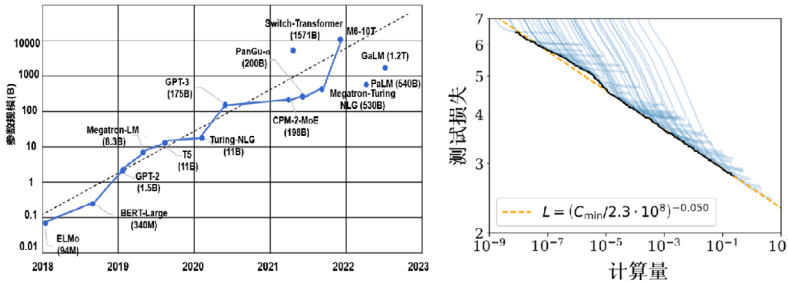
### 2.2.1.1 大语言模型基本概念

自 2018 年，以 BERT 和 GPT 为代表的预训练语言模型（PLM）技术，大幅刷新各类自然语言处理任务的性能上限，已经成为人工智能领域的主流技术范式。预训练语言模型采用“预训练 + 微调”方法，主要分为两步：1）将模型在大规模无标注数据上进行自监督训练得到预训练模型，2）将模型在下游各种自然语言处理任务上的小规模有标注数据进行微调得到适配模型。相比传统人工智能模型，预训练模型在下游应用中具有数据成本低、通用性强、综合性能好等优势。


图表 2-4 预训练语言模型“预训练 + 微调”技术范式

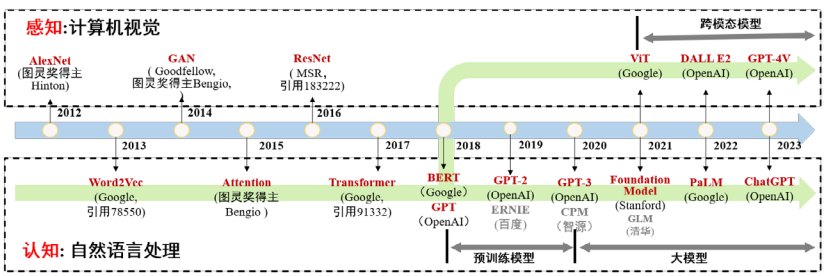
大语言模型（Large Language Model，LLM）是指大规模预训练语言模型。2020 年 5 月，OpenAI 发布了拥有 1750 亿参数 LLM 模型 GPT-3，能够完成文章撰写、对话问答、自动编程等复杂人工智能任务，

并且仅通过少量样本的学习，就达到逼近人类的学习能力，展现出迈向通用人工智能（AGI）的可行路径。由于 PLM 模型性能与模型参数、训练数据量呈现“伸缩定律”（Scaling Law）现象，即模型参数、训练数据规模越大模型性能越好，这激发了大语言模型研究热潮。大模型参数在 2018 年 -2022 年基本呈 10 倍增加趋势。国内外有许多有影响力的 LLM 被提出。




图表 2-5 更大模型更好效果

### 2.2.1.2 大语言模型发展历程




图表 2-6 大模型发展历程

图表 2-6 展示了由深度学习引导的本轮人工智能大潮里程碑式成

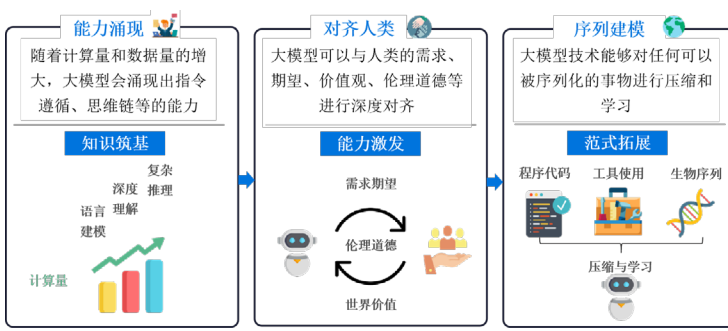
果。本轮深度学习浪潮可以最早从视觉领域发展起来，2012 年图灵奖得主 Hinton 提出 AlexNet 在大规模视觉识别挑战赛 ImageNet 评测上大幅超越现有模型，并首次在深度学习中引入 GPU 加速，激发了深度学习的研究热潮。2012 至 2016 年间，视觉领域成为深度学习的主导领域，生成对抗网络 GAN、深度残差网络 ResNet 等创新技术应运而生。同时，自然语言处理领域亦有所发展，如文本词嵌入 Word2Vec 和 Attention 机制的提出，奠定了深度学习在 NLP 领域的基础，尽管其在性能提升上并不显著。2017 年成为转折点，Google 提出的 Transformer 框架在机器翻译中取得显著进步，其分布式学习和强大编码能力受到广泛关注。继而，2018 年 Google 和 OpenAI 基于 Transformer 提出了预训练语言模型 BERT 和 GPT，显著提高了 NLP 任务的性能，并展示出广泛的通用性。这标志着“预训练 + 微调”技术范式的开端。此后，众多预训练模型相继涌现，OpenAI 以 GPT-2、GPT-3、ChatGPT 等系列模型为代表，持续引领大模型时代的浪潮。2022 年的 GPT-3，首次将模型参数规模扩展至 1750 亿，展示了少样本学习和复杂任务处理的能力，显示出实现通用智能的巨大潜力，开启了大模型时代。自 2018 年起，NLP 预训练技术成为 AI 技术发展的主导力量，并逐渐渗透到计算机视觉领域，催生了 DALL-E2、GPT-4V 等跨模态模型，进一步推动了深度学习和人工智能的发展。

此次大模型浪潮中，OpenAI 成为该领域的绝对的领导者，其提出了系列有影响力的大模型，特别是 ChatGPT 的提出，标志着大模型性能发生质变，开创了人工智能的新变革。图表 2-7 展示了 OpenAI 的系列模型发展历程。


图表 2-7 OpenAI 的 ChatGPT 发展历程

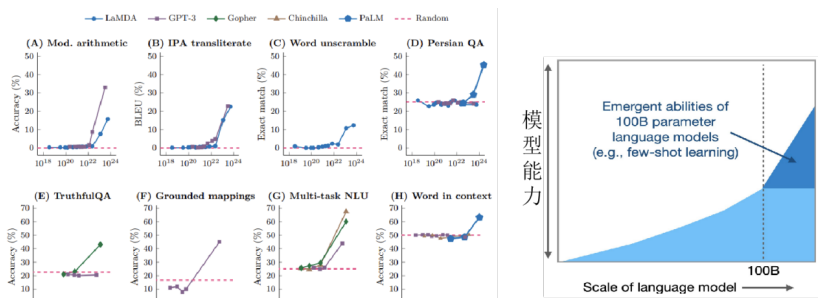
## 2.2.1.3 大语言模型能力与特点

大语言模型较传统人工智能模型，呈现出如下能力和特点，如图表 2-8 所示：




图表 2-8 大语言模型的能力与特点

• **涌现能力（Emergent Abilities）**，随着模型计算量和训练数据量的增加，大语言模会涌现出上下文学习、指令遵循、思维链推理、交互认知等能力。这里上下文学习是指给定少量演示样本，大模型就可以参考回答用户的问题，具备了一举反三能力；指令遵循是指用户给定任务描述文本指令，大模型可以找指令要求回答问题；思维链推理旨在大模型能够给出问题解答过程，通过推理过程可以提升大模型回答准确率；交互认知是指大模型具备与工具、环境等交互完成任务的能力。




图表 2-9 大模型涌现能力现象 [2-54]

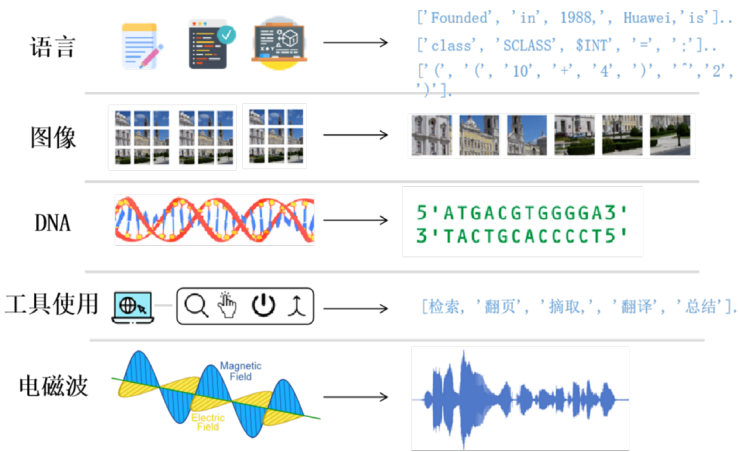
• **对齐人类**，大模型涌现能力，可以进一步与人类期望输出对齐。大模型可以与人类的需求、期望、价值观、伦理道德等进行深度对齐，通过有监督微调和人类反馈强化学习等学习人类偏好反馈，能够有效降低大模型的错误、虚假等“幻觉”内容生成，提升大模型的忠诚性、可靠性、有帮助性等，这是 ChatGPT 成功关键，也是目前解决大模型安全的关键技术。OpenAI 团队提出了超级对齐的概念，并给出了

超级对齐四年计划。


图表 2-10 对齐学习构建数字空间和人类社会桥梁

- **序列建模**，大语言模型技术能够对任何可以被序列化的事务进行压缩和学习。大语言模型采用 Transformer 架构，通过将输入转化成 token 序列实现对输入的编码和理解。目前 Transformer 架构已经成为文本、视觉、语音等各种领域的大模型的核心架构，实现了对各种模态数据编码能力。在文本之外，我们可以通过序列化方法抽象、学习理解世界中的万事万物，如语言可以转化成文本序列，图像通过切分可以划分成 patch token 的序列，DNA 可以以碱基为 token 划分成序列，Agent 的工具调用可以划分成动作执行的序列，电磁波可以转化成音频序列等。在大模型中这些序列都是词元（Token）序列。

任何可以被序列化的信息均可被大模型学习。




图表 2-11 不同领域的序列化建模

### 2.2.1.4 大语言模型发展趋势

目前，大语言模型发展的主要趋势可以概括为以下几个方向：

**更大模型参数：** 由于大模型性能与模型参数呈现“Scaling Law”（扩展定律），即在充分数据训练下模型参数规模越大模型的性能越好。同时，模型参数规模越大模型的泛化性和复杂数据的编码能力也越好，而且呈现更强的涌现能力。这激发了人们对更大模型的持续追求。许多超大规模参数模型被发布，如 OpenAI 的 GPT-3（175B）、Google 的 PaLM（540B）、智源的“悟道 2.0”（1750B）等，模型参数规模从过去的 5 年间，参数规模增长 5000 倍（2018 年几亿参数规模 BERT 发展到 2023 年万亿参数规模 GPT-4）。

**多模态大模型：**多模态数据丰富无处不在，互联网 90% 以上是图像与音视频数据，文本不到 10%。多模态协同更符合人类感知与表达方式，是机器实现类人智能重要途径。目前构建融合更多模态的大模型是当前大模型发展趋势。这一趋势是指将文本、图像、声音等多种模态的数据融合在一起，通过大模型进行处理和理解。例如，Midjourney 和 OpenAI 的 DALL-E2 能够根据文本描述生成相应的图像，而 GPT-4 可以根据理解图像和文本跨模态理解和生成。这类模型的发展，使得 AI 在视觉艺术、设计等领域的应用更加广泛和深入。

**AI for Science（大模型 +X）：**这个方向强调将大语言模型应用于科学研究中，例如药物发现、蛋白质结构预测等。大模型在这些领域的应用，不仅能够加速数据分析和知识发现，还能够提出新的科学假设和研究方向。例如，2022 年 Google DeepMind 发布基于大模型的蛋白质结构预测模型 Alphafold，预测准确性已达到与人类可比水平，取得了重大突破，极大地加速了生物医学领域的研究进程。清华大学将大模型应用于生医领域提出了 KV-PLM，将生医文献数据中分子结构通过 SMILES 表达式的形式映射到自然语言，然后对文字表达序列和生医文本进行掩码语言建模，实现了分子表达式与文本描述的桥接，在分子检索等领域任务上取得大幅提升。

**AI Agent：**是指开发能够更加自主、智能和互动的 AI 智能体。这些智能体可以在多种场景下协助人类，如个人助理、客服机器人、教

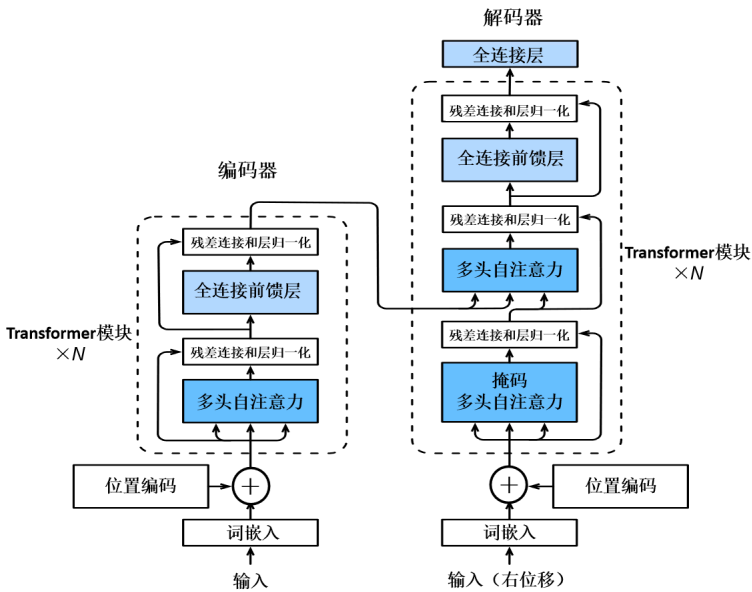
育辅助等。AI Agent 的发展不仅在于算法本身的优化，还包括对人类行为和需求的理解，以及与人类的交互能力。例如，GPT-4 等大语言模型通过智能体形式（如 ChatDev、AutoGPT、XAgent、AutoGen 等）已被应用于软件开发、创作、营销、社会模拟等多种复杂场景任务处理，展示更加强大的智能水平。比尔盖茨认为 AI Agent 是人工智能的未来。2023 年 11 月 OpenAI 开发者大会发布 AI Agent 开发平台 GPTs，用户和开发者可以定制和商业化发布自己的 Agent，将 AI Agent 发展推向了高潮。

## 2.2.2 大语言模型的模型架构

### 2.2.2.1 Transformer 架构

Transformer 架构 [2-1] 是目前大语言模型采用的主流架构 [2-2]，其基于自注意力机制 (Self-attention Mechanism) 模型。其主要思想是通过自注意力机制获取输入序列的全局信息，并将这些信息通过网络层进行传递。标准的 Transformer 如图表 2-12 所示，是一个编码器 - 解码器架构，其编码器和解码器均由一个编码层和若干相同的 Transformer 模块层堆叠组成，编码器的 Transformer 模块层包括多头注意力层和全连接前馈网络层，这两部分通过残差连接和层归一化操作连接起来。与编码器模块相比，解码器由于需要考虑编码器输出作为背景信息进行生成，其中每个 Transformer 层多了一个交叉

注意力层。相比于传统循环神经网络（RNN）和长短时记忆神经网络（LSTM），Transformer 架构的优势在于它的并行计算能力，即不需要按照时间步顺序地进行计算。




图表 2-12 Transformer 架构 [2-1]

Transformer 架构包含编码层与 Transformer 模块两个核心组件：

**编码层，**主要是将输入词序列映射到连续值向量空间进行编码，每个词编码由词嵌入和位置编码构成，由二者加和得到：

1）词嵌入，在 Transformer 架构中，词嵌入是输入数据的第一步处理过程，它将词映射到高维空间中的向量，可以捕获词汇的语义

信息，如词义和语法关系。每个词都被转化为一个固定长度的向量，然后被送入模型进行处理。

2）位置编码，由于自注意力机制本身对位置信息不敏感，为了让模型能够理解序列中的顺序信息，引入了位置编码。标准 Transformer 架构的位置编码方式是使用正弦和余弦函数的方法。

**Transformer 模块，**通过自注意力机制获取输入序列的全局信息，并将这些信息通过网络层进行传递，包括多头注意力层和全连接前馈网络层，这两部分通过残差连接和层归一化操作连接起来，Transformer 模块，由自注意力层、全连接前馈层、残差连接和层归一化操作等基本单元组成：

**1）自注意力层，**注意力（Attention）是 Transformer 模型的核心组成部分。它包含一个查询矩阵 $Q \in \mathbb{R}^{n \times d_k}$，一个键矩阵 $K \in \mathbb{R}^{m \times d_k}$ 和一个值矩阵 $V \in \mathbb{R}^{m \times d_v}$ 其中矩阵中的每一行对应一个词。注意力机制的计算方式：

$$H = \text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

此外，Transformer 采用了多头自注意力（Multi-head Attention）机制，即输入序列被线性映射多次得到不同的投影矩阵。多个尺度化后点积注意力可以并行计算，并产生多个自注意力输出。多头注意力生成多个高维的注意力表示，这使得其比单头注意力具有更强的表达能力。

**2）全连接前馈层，**在注意力层之后的全连接前馈层由两个线性变换和一个非线性激活函数组成：

$$\mathrm{FFN}(X)=\sigma(XW_1+b_1)W_2+b_2$$

FFN 作用包括两个方面：（1）非线性激活：在每个注意力模块之后引入了非线性激活函数，这有助于增强模型的表达能力；（2）信息整合：自注意力机制允许模型在不同的位置间建立联系，而全连接前馈网络则在每个位置独立地对信息进行整合，这两者结合起来，使得模型既能捕获全局（长距离）的信息，又能在每个位置进行局部的信息整合。

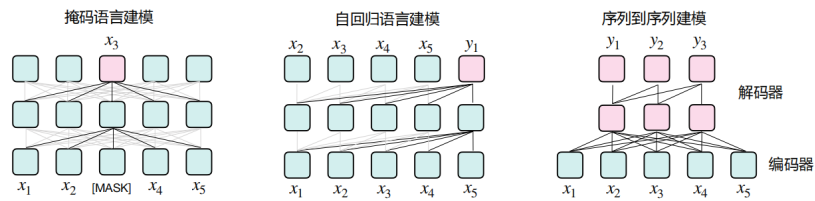
**3）残差连接和层归一化，**在每个注意力层和每个全连接前馈层之后，Transformer 都应用残差连接（Residual Connection）和层归一化（Layer Normalization）技术，这有助于在模型非常深时保留信息并确保模型性能。具体来说，对于某一层神经网络$f(\cdot)$，残差连接和归一化层定义为 $\mathrm{LayerNorm}(X+f(X))$

在 Transformer 模型被提出之后，它也衍生出了相当一部分的变体，包括在编码器和解码器中出现了不同方式的注意力机制、归一化操作、残差连接、前馈层和位置编码等。

### 2.2.2.2 大语言模型典型架构

现有的大语言模型几乎全部是以 Transformer 模型作为基础架

构来构建的，不过它们在所采用的具体结构上通常存在差异。LLM 根据架构主要分为三类：1）自回归语言模型，采用 Transformer 的编码器（Decoder），代表性模型包括 OpenAI 的 GPT 系列模型 [2-6] [2-7]、Meta 的 LLaMA 系列模型 [2-8] 和 Google 的 PaLM 系列模型 [2-9]；2）自编码语言模型，采用 Transformer Encoder 作为模型架构，代表性模型 BERT、RoBERTa 等；3）序列到序列语言模型，采用 Transformer 的 Encoder-Decoder 整体架构，代表模型包括 T5、BART 等。

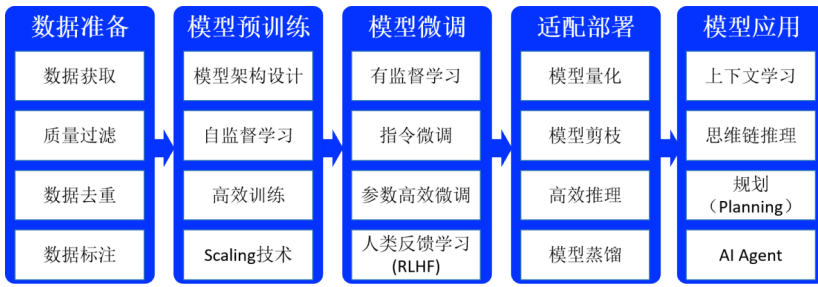



图表 2-13 大语言模型的三种典型架构 [2-3]

目前 LLM 在国际上也被认为是实现通用人工智能的“基础模型”（Foundation Model），在国内也被称为“大模型”。2022 年底，OpenAI 发布了对话 LLM 模型 ChatGPT，能够同时完成撰写、翻译、对话、代码生成等任务，展现了强大的语言理解、多类型任务处理、认知交互能力，取得了巨大成功，标志 AGI 迈向了新的台阶。由于 GPT-3、ChatGPT 等的成功和展现的巨大潜力，使得自回归语言模型成为当下 LLM 主流架构。

### 2.2.3 大语言模型关键技术

大语言模型构建的整体技术路线如图表 2-14 所示，按照流程顺序依次包括数据准备、模型预训练、模型微调、适配部署、模型应用等关键步骤。




图表 2-14 大语言模型构建技术路线

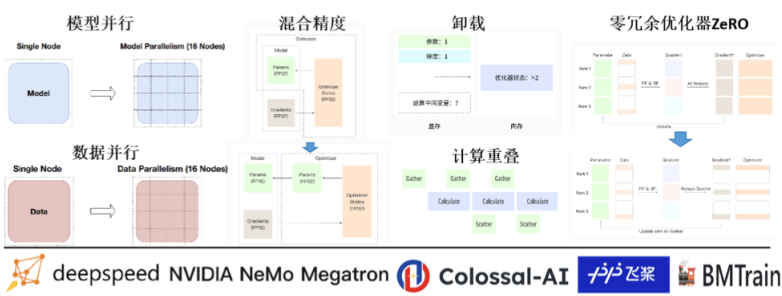
下面对大语言模型构建中主要关键技术进行介绍，包括模型预训练、适配微调、提示学习、知识增强和工具学习等 [2-48]。

#### 2.2.3.1 大语言模型的高效预训练

支撑大语言模型高效训练的技术主要包括高性能训练工具、高效预训练策略、高质量训练数据、高效的模型架构等。

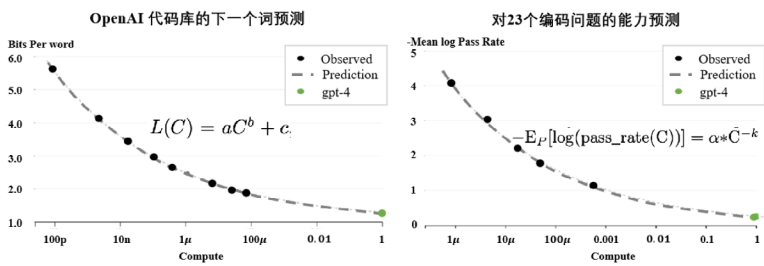
**高性能训练工具，**旨在通过对模型计算、显存、内存和通信使用的系统级优化，提高训练吞吐量和加载更大模型到显存中，实现在有限资源下大模型高效训练的目的。系统级优化通常是与模型无关的，

并且不会改变底层的学习算法，被广泛应用于各种大模型的模型。相关方法主要从两个方向实现：一是**设备内优化方法**，包括降低浮点数的冗余表示的半精度浮点优化、混合精度浮点优化等方法，降低梯度计算中过程中冗余表示的梯度检查点（Checkpointing）方法，以及内存优化的 ZeRO-Offload 方法，即通过将数据和计算从 GPU 卸载到 CPU，以此减少神经网络训练期间 GPU 内存占用的方法。二是**多设备优化方法**，也称分布式优化，即分布在许多计算节点上的多个 GPU 一起用于训练单个模型，这类方法主要有数据并行、模型并行、流水线并行等方法。数据并行性，即当将一个大的批处理数据被划分到不同的计算节点。模型并行性，即在进行模型并行性时，模型参数可以分布到多个节点上。流水线并行，它将一个深度神经网络划分为多层，然后将不同的层放到不同的节点上，计算每个节点后，输出被发送到下一个节点进行下一层计算。以上三种维度的并行优化方法相互独立，可以同时使用来加速模型训练。基于以上方法构建的代表性的大模型训练工具，主要有微软的 DeepSpeed-Megatron、NVIDIA 的 Megatron-LM、新加坡国立大学的 Colossal-AI、清华大学的 BMTrain 等。




图表 2-15 高效计算工具与高效计算策略

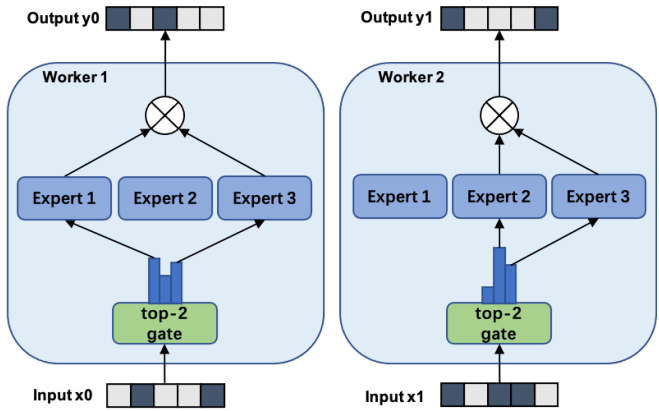
**高效预训练策略。**其主要思路是采用不同的策略以更低成本实现对大语言模型的预训练。一种是在预训练中**设计高效的优化任务目标，**使得可以使得模型能够利用每个样本更多的监督信息，从而实现模型训练的加速。第二种是**热启动策略，**在训练开始时线性地提高学习率，以解决在预训练中单纯增加批处理大小可能会导致优化困难问题。第三种是**渐进式训练策略，**不同于传统的训练范式使用相同的超参数同时优化模型每一层，该方法认为不同的层可以共享相似的自注意力模式，首先训练浅层模型，然后复制构建深层模型。第四种是**知识继承方法，**即在模型训练中同时学习文本和已经预训练大语言模型中的知识，以加速模型训练。在中文大语言模型 CPM-2[2-12] 中，采用知识继承技术经测试可以使大模型在预训练前期提速 37.5%。第五种是**可预测扩展策略**（Predictable Scaling）[2-7]，旨在大模型训练初期，利用大模型和小模型的同源性关系，通过拟合系列较小模型的性能曲线预测大模型性能，指导大模型训练优化。OpenAI 在 GPT-4 训练中，使用 1000 倍至 10000 倍较少计算资源训练的小模型可靠地预测 GPT-4 某些性能，大幅降低了训练试错成本。




图表 2-16 GPT-4 的可预测扩展实验 [2-7]

**高效的模型架构：**BERT 之后的 Transformer 架构在提高自然语言处理效率方面有两个重要优化方向：（1）**统一的序列建模，**旨在将多种自然语言处理任务（如分类、信息抽取、翻译、对话等）整合到一个统一的框架，然后在同一模型中执行多个任务，以实现更高效的自然语言处理。该方法可以充分利用大规模训练数据，从而提高了模型在多个任务上的性能和泛化性。这减少了开发和维护多个单独模型的复杂性以及资源消耗，提高模型的通用性。统一任务序列建模有两种方式：一是转化为序列生成的统一任务，如 T5[2-10] 和 BART[2-9] 等将多种自然语言任务统一转化文本到文本的生成任务；二是转化为大语言模型预训练任务，通过语言提示在输入文本中插入人类设计或者自动生成的上下文，实现对不同任务的处理。（2）**计算高效的模型架构。**从 Transformer 模型架构本身在处理训练复杂度、编解码效率、训练稳定性、显存利用等方面进行优化。比如，Transformer 其并行处理机制是以低效推理为代价的，解码时每个步骤的复杂度为 O(N)，Transformer 模型也是显存密集型模型，输入序列越长、占用的内存越多。为此，微软设计了一种新的 Transformer 架构 RetNet[2-13]，其采用线性化注意力 + 尺度保持（Retention ）机制，在基本保持模型性能的基础上同时实现模型训练速度、推断速度和内存节约的大幅提升。针对自注意力显存消耗大，斯坦福大学在 Transformer 中引入 FashAttention[2-14]，给出了一种具有 IO 感知，且兼具快速、内存高效的注意力算法，已经被各种主流大模型采用以扩展对超长文本输入的支持。最近，模块化大模型架构引起广泛关注，其利用大模型的神经激活稀疏性，对稠密模型进行模块化划分，不同任务

只经过部分模块计算实现训练和推理加速，典型工作包括 Google 的 Switch Transformers [2-15] 和 Pathways[2-16] 架构、清华大学的 MoEfication 架构 [2-17]、FastMoE 架构 [2-18] 等。




图表 2-17 混合专家化的模型架构 [2-18]

### 2.2.3.2 大语言模型的适配微调

大语言模型由于在大规模通用领域数据预训练通常缺乏对特定任务或领域的知识，因此需要适配微调。微调可以帮助模型更好地适应特定需求，如对敏感数据（如医疗记录）的处理，同时不暴露原始数据。此外，微调可以提高部署效率、减少计算资源需求。指令微调和参数高效学习是适配微调的关键技术。

**指令微调 (Instruction Tuning)**[2-19]，是一种可以帮助大语言模型实现人类语言指令遵循的能力，在零样本设置中泛化到未见任务

上的学习方法。指令微调学习形式与多任务提示微调相似，但与提示微调让提示适应大语言模型并且让下游任务对齐预训练任务不同，其是让大语言模型对齐理解人类指令并按照指令要求完成任务，即在给定指令提示的情况下给出特定的回应，其中提示可以选择性包含一条解释任务的指令。指令微调研究涉及**指令理解、指令数据获取和指令对齐**等内容。

**（1）指令理解，**指大语言模型准确理解人类语言指令的能力，是大语言模型执行指令完成任务的前提。为了增强对指令的理解，许多工作采用多任务提示方式对基于指令描述的大量任务集上对大语言模型进行微调，如 FLAN[2-20]、InstructGPT[2-19] 等，这些模型在未见的任务上显示出优越的零样本性能。

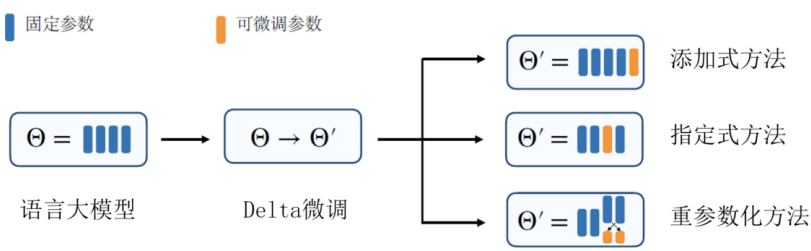
**（2）指令数据获取 ,** 指如何构建包含多样性的任务指令数据。指令数据构建常见有三种方式：i）基于公开人工标注数据构建，代表指令数据集包括 1616 种不同任务的 Super-Natural Instruction[2-21]、2000 种不同 NLP 任务的 OPT-IML[2-22]。ii）借助大语言模型的自动生成构建，如 Unnatural Instructions[2-23]，通过种子指令作为提示让大语言模型生成新的指令描述和问题，然后再输入到模型让其输出回答。清华大学 & 面壁智能团队推出的对话指令数据集 UltraChat，通过调用多个 ChatGPT API 相互对话生成高质量的训练数据。此外，还通过自动标注的方法构建了面向大模型对齐的大规模反馈数据 UltraFeedback，HuggingFace 团队通过该数据集训练得到的 Zephyr-7B 性能参数大 10 倍的 LLaMA2-70B-Chat。iii）基于人工标

注的方法，如通过 GPT-3 API、InstructGPT API、ChatGPT 等在线平台收集用户真实指令数据。

**（3）指令对齐，**大语言模型在多种自然语言处理任务上都展现了卓越的性能。然而，它们有时可能会出现不预期的行为，如创造虚假信息、追求错误目标或产生有偏见的内容 [2-2]。其根本原因在于，大语言模型在预训练时仅通过语言模型建模，未涉及人类的价值观或偏好。为了解决这一问题，研究者提出了“指令对齐”，使大语言模型的输出更符合人类的预期。但这种对齐与原始预训练有所不同，更注重于有用性、诚实性和无害性。此外，指令对齐可能会降低大语言模型的某些通用能力，这被称为“Alignment Tax”。为实现模型输出与对人类价值的对齐，InstructGPT 提出了一种基于人类反馈的微调方法，利用了强化学习技术，将人类反馈纳入模型微调过程。实际上，ChatGPT 也采用了与 InstructGPT 相似的技术，以确保产生高质量且无害的输出。指令对齐的广泛应用，适配微调从纯数据学习的传统微调范式开始逐步向人类学习范式的转变。

**参数高效微调**（Parameter-Efficient Tuning）。早期以 BERT 为代表的微调方法，是在大模型基座上增加一个任务适配层，然后进行全参微调，但是这种方法存在两方面的问题：一是任务“鸿沟”问题，预训练和微调之间的任务形式不一致，这种差别会显著影响知识迁移的效能。二是高计算成本，大语言模型的参数规模不断增长，导致模型全参微调也需要大量计算资源。解决以上问题的有效途径是参数高效学习，即通过仅微调少量参数实现大模型在下游任务上获得全参微

调效果。目前许多参数高效微调方法被提出，这些方法大致可分为 3 类 [2-3]：（1）**添加式方法：**旨在原模型基础上引入额外的模块或参数，并仅微调该引入部分的参数。如适配器（Adapter）方法，旨将小规模的神经模块（适配器）注入到预训练模型中，并只调整这些适配器以进行模型自适应。在实际应用中，适配器模块通常分别插入在多头自注意和前馈网络子层之后，成为最广泛使用方式；（2）**指定式方法：**旨在原模型指定模型中部分参数为可训练参数，并固定模型其他参数。这类方法简单也十分有效，如仅通过优化模型内的偏置项并固定其他参数，模型仍然可以再现 95% 以上的模型全参微调性能；（3）**重参数化方法：**将原模型或部分模型参数重参数化到低维度参数空间中，仅仅优化低维空间中的近似参数，显著降低模型的计算量和内存消耗。如 LoRA[2-24]，将模型自注意力模块的变化权重参数分解为两个低秩矩阵相乘，即 $W = W_0 + \Delta W = W_0 + W_{down} W_{up}$




图表 2-18 参数高效微调的 3 种范式 [2-3]

参数高效微调通常具有微调参数规模小、增量式微调参数、即插即用等特点，这种技术也统一成技术框架 Delta Tuning[2-3]。

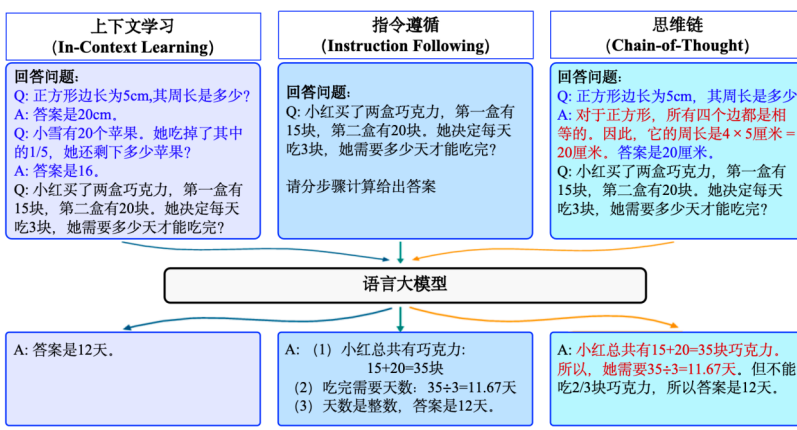
一些围绕参数高效微调的开源工具也被研发，代表性包括OpenPrompt[2-25]、OpenDelta[2-26] 等。由于不同任务的微调参数可以被重复利用，一些关于高效微调的仓库也被构建，如AdapterHub[2-27]、Delta Center[2-3] 等。随着大语言模型的兴起，高效微调吸引了越来越多的关注，以开发一种更轻量级的下游任务适配方法。特别地，LoRA[2-24] 已广泛应用于各种开源大语言模型（如LLaMA）以实现参数高效微调。

#### 2.2.3.3 大语言模型的提示学习

通过大规模文本数据预训练之后的大语言模型具备了作为通用任务求解器的潜在能力，但这些能力在执行一些特定任务时可能不会显式地展示出来。在大模型输入中设计合适的语言指令提示有助于激发这些能力，该技术称为模型提示技术。代表性的提示技术有指令提示和思维链提示：

**指令提示（Instruction Prompt），**也称为提示学习。OpenAI在 GPT-3 [2-6] 中首次提出上下文提示，并发现 GPT-3 在少样本提示下能够达到人类水平，证明在低资源场景下非常有效，引起广泛关注。指令提示核心思想是 w 避免强制大语言模型适应下游任务，而是通过提供“提示（Prompt）”来给数据嵌入额外的上下文以重新组织下游任务，使之看起来更像是在大语言模型预训练过程中解决的问题 [2-28]。指令提示有三种形式：（1）**少样本提示，**是指在一个自然语言

提示后面附加一些示例数据，作为大语言模型的输入。其可以提高大语言模型在不同领域和任务上的适应性和稳定性。少样本提示也存在一些挑战，例如如何确定合适的示例数量、如何选择示例等；（2）**零样本提示**，是指不使用任何示例数据，只依靠一个精心设计的提示来激活大语言模型中与目标任务相关的知识和能力。零样本提示关键问题包括如何设计合适的提示、如何选择最优的提示等；（3）**上下文学习（In-context Learning, ICL）**，也称**情境学习**，是指将一个自然语言问题作为大语言模型的输入，并将其答案作为输出 [2-6]。情境学习可以看作是一种特殊形式的少样本提示，在问题中隐含地包含了目标任务和格式信息。情境学习可以简化问题表示和答案生成，并且可以灵活地处理多种类型和复杂度的问题。其挑战在于，如何确保问题质量、如何评估答案正确性等。




图表 2-19 几种提示样例对比

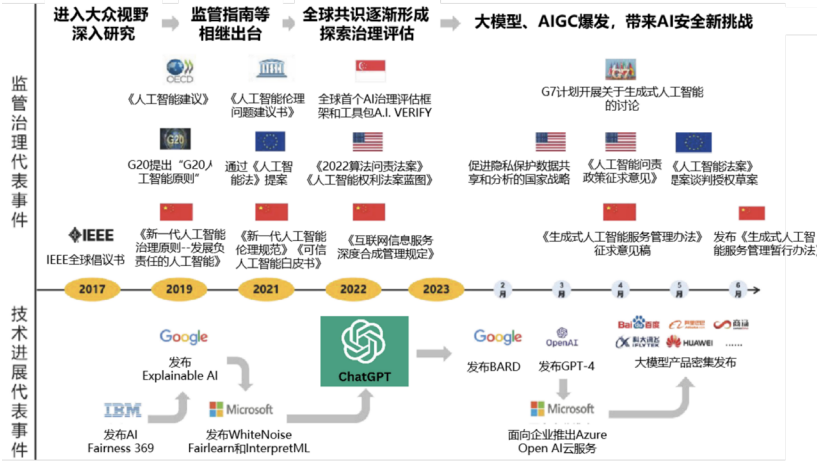
**思维链（Chain-of-Thought，CoT）**[2-29]。推理的过程通常涉及多个推论步骤，通过多步推理允许产生可验证的输出，可以提高黑盒模型的可解释性。思维链是一种提示技术，已被广泛用于激发大语言模型的多步推理能力，被鼓励大语言模型生成解决问题的中间推理链，类似于人类使用深思熟虑的过程来执行复杂的任务。在思维链提示中，中间自然语言推理步骤的例子取代了少样本提示中的〈输入，输出〉对，形成了〈输入，思维链，输出〉三元组结构。思维链被认为是大语言模型的“涌现能力”，通常只有模型参数规模增大到一定程度后，才具有采用思维链能力。激活大语言模型的思维链能力方法，在提示中给出逐步的推理演示作为推理的条件，每个演示都包含一个问题和一个通向最终答案的推理链（图表 2-19）。CoT 在推理过程中从左到右的 token 级决策，一般不擅长对需要探索、策略性预见、推理存在结构关系的任务，思维树（Tree of Thought，ToT）和思维图（Graph of Thought，GoT） 方法通过考虑不同推理路径或推理结构进行决策，提升大语言模型的推理效果。

#### 2.2.3.4 大语言模型的安全治理

大模型在应用的过程中，可能会产生与人类价值观不一致的输出，如歧视言论、辱骂、违背伦理道德的内容等，这种潜在的安全风险普遍存在于文本、图像、语音和视频等诸多应用场景中，并会随着模型的大规模部署带来日益严重的安全隐患。目前大模型衍生出内容安

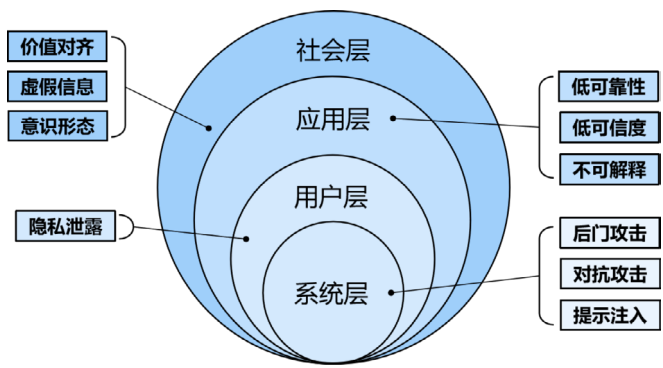
全、隐私安全、政治安全、软硬件安全等诸多安全风险问题。2023 年 5 月，三星半导体工程师使用 ChatGPT 参与修复源代码时发生无意间泄密芯片机密代码的重大事故。大模型容易受到攻击，人们发现对 ChatGPT 进行提示注入，诱导可以输出 Windows11 的序列号。大模型存在严重的“幻觉”问题，模型在输出中生成生成错误、编造虚假信息，容易误导用户。大模型安全风险引发国际广泛关注。2023 年 3 月，1000 多名 AI 学者和企业家联名信呼吁暂停大型 AI 实验。图灵奖得主、深度学习先驱 Hinton 离开谷歌，表达对当前 AI 系统风险的担忧。

国际和各国纷纷出台各种政策法规以规范化大模型发展。2023 年 3 月，美国白宫科技政策办公室发布《促进隐私保护数据共享和分析的国家战略》。该策略旨在保障公共和私营部门实体中用户的数据隐私，同时确保数据使用的公平性和最大的效率。2023 年 6 月，欧洲议会（European Parliament）通过《人工智能法案》草案，旨在为人工智能引入统一的监管和法律框架，并涵盖了除军事用途外的所有人工智能类型。2023 年 7 月，国家互联网信息办公室发布的《生成式人工智能服务管理暂行办法》，对生成式人工智能服务在算法设计、训练数据选择、模型生成和优化、提供服务等过程中进行安全规范。2023 年 10 月，国家网信办发布《生成式人工智能服务安全基本要求》（征求意见稿），给出了生成式人工智能服务在安全方面的基本要求，包括语料安全、模型安全、安全措施、安全评估等。




图表 2-20 国内外政府颁布各种对 AI、大模型安全的政策法规

大模型安全问题从大模型服务链路关键层面，可以分为：系统层，包括后门攻击、对抗攻击、提示注入等问题；用户层，包括隐私泄露、知识版权等问题；应用层，包括低可靠、低可信度、不可解释等问题；社会层，包括价值对齐、虚假信息、意识形态等问题。



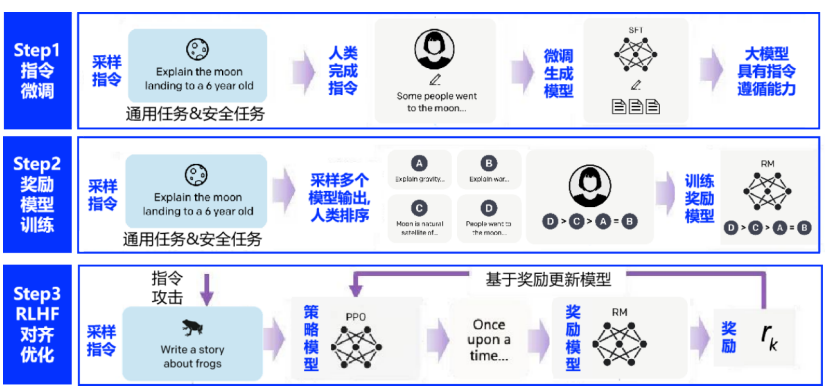

图表 2-21 大模型“洋葱”型安全体系

目前对大模型安全治理技术大致可以分为以下几个方面：

**安全数据构建。**训练数据的安全性是构建安全大模型的基石。训练数据安全性是指数据集的来源和质量都是可靠的，数据中蕴含的知识是准确的，数据集内容符合主流价值观。方法包括：1）确保训练数据来自可信的、可靠的来源。数据应该从权威机构、专业组织、可验证的数据仓库或其他公认的数据提供者获得。在数据标注时，确保标注的准确性和一致性。标注过程应该由经过培训的专业人员进行，并且需要进行验证和审核，以确保标注的正确性。此外，需要进行数据清洗以去除重复项、噪声数据和错误数据。2）数据的敏感信息去除。在大模型中，保护数据的敏感信息是至关重要的，特别是当模型需要处理涉及个人隐私、敏感信息或商业机密等敏感数据时。数据的敏感信息去除是一种隐私保护措施，旨在确保数据在训练过程中不会泄露敏感信息。去除方法包括：数据脱敏、去标识化等。3）有害信息过滤。通过构建有害关键词库、人工规则、安全分类模型等，对数据涉及安全风险类型数据进行过滤清洗。2023 年 10 月，国家网信办《生成式人工智能服务安全基本要求》（征求意见稿）中对语料及生成内容的主要安全风险进行了分类，包括包含违反社会主义核心价值观的内容、包含歧视性内容、商业违法违规、侵犯他人合法权益、无法满足特定服务类型的安全需求等 5 大类，细分 31 种。

**模型安全对齐。**为了训练有用、诚实和无害的人工智能系统，OpenAI 发布的 ChatGPT 系列大模型采用 InstructGPT 的技术框架，使用人类反馈的强化学习技术（RLHF）实现大模型与人类偏好的安

全对齐。让模型的输出与人类价值观尽可能一致，提高其有用性、真实性和无害性。RLHF 训练过程包括指令微调、奖励模型训练和对齐优化三个阶段。指令微调阶段，也称有监督微调，旨在优化大模型，使其能够理解用户的指令；奖励模型训练阶段中，人类对模型生成的多条不同回复进行评估，这些回复两两组合，由人类确定哪条更优，生成的人类偏好标签使奖励模型能学习并拟合人类的偏好。在对齐优化阶段，奖励模型根据生成回复的质量计算奖励，这个奖励作为强化学习框架中的反馈，并用于更新当前策略的模型参数，从而让模型的输出更符合人类的期望。这一阶段体现了人类价值观和模型技术逻辑的深度交融，通过人类反馈调整模型的产出、优化模型的生成策略，使其更好地反映人类价值观。基于人类反馈的安全对齐技术已逐渐成为当下大模型安全研究的主流技术。除了 OpenAI，DeepMind 的 Sparrow、Anthropic 的 Claude 模型等国外大模型，以及国内代表性大模型（如文心一言、智谱清言、面壁 LUCA 等）也采用了类似技术。




图表 2-22 人类反馈强化学习的安全对齐技术路线

**模型幻觉治理。**大模型生成内容存在严重的“幻觉”问题，容易生成错误、虚假信息，尤其对于事实知识性问题。该问题对于相关知识学习缺乏越严重领域该问题越严重。目前降低大模型幻觉的方法主要有：1）**外接知识库**，即让大模型在回答问题时，能够通过调用网页搜索引擎或本地知识库检索，获取缺乏的相关背景知识作为上下文，再进行回答，并且在回答内容中提供内容原始来源，提升大模型生成内容的准确性和可解释性，如 OpenAI 的 WebGPT、ChatGPT 的 Web Browsing 插件调用，清华大学 & 面壁智能的 WebCPM 等，调用网页搜索引擎获取互联网信息回答用户问题，并在回答中提供链接；2）**分多步推理并展示推理过程**，即将复杂任务问题通过思维链技术拆解成多步执行，将中间状态输出展示给用户；3）自定义工作流 Workflow，比如对中间任务需要严格执行的过程通过预定义工作流，提升中间内容的精准性和可控性，如 COZE、灵境矩阵等智能体生产平台；3）**工具调用**，对于专业技能问题如数值计算、软件编程、数据分析等，可以通过调用计算器、代码编译器、数据库等人类专业工具，弥补大模型专业技能的缺失，代表性工作包括 ChatGPT Plugins、文心一言插件功能、工具学习技术框架 ToolLLM 等；4）**人机交互**，在大模型运行期间增加大模型与人类交互，对大模型不确定性的任务获取人类反馈后执行，代表性工作如超级智能体 XAgent；5）**大模型持续学习**，让大模型持续学习更多的训练数据，提升大模型知识覆盖度。

**模型对抗防御 [2-56]**。大语言模型在受到提示注入攻击、模型

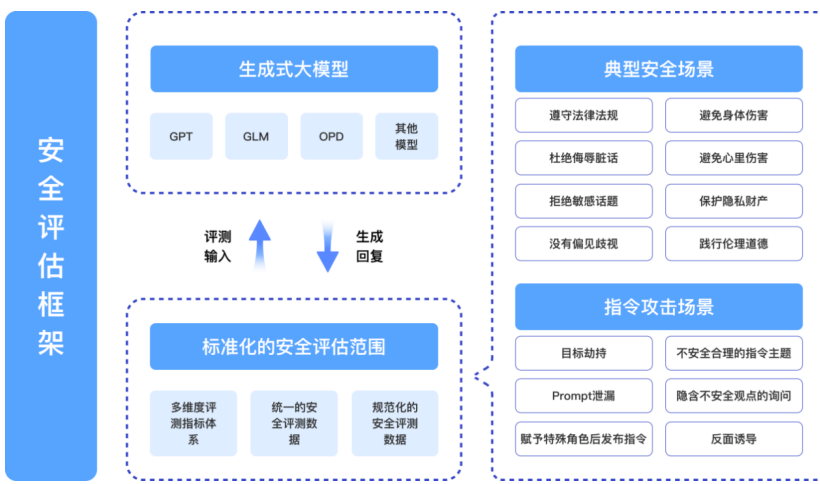
输入扰动、模型参数修改等方式使得模型正常服务。1）**提示注入攻击**，即使用精心设计的提示诱导模型输出违反其安全规则的答案。如在 New Bing 的聊天搜索引擎刚推出时，斯坦福大学学生 Kevin Liu 成功地对其进行了提示注入攻击，他发现聊天机器人的内部代号是“Sydney”，并成功地泄露了一系列微软为 Sydney 设定的行为规则。通过角色扮演越狱攻击，可以让 ChatGPT 诱导输出 Windows11 的注册码。2）**模型输入扰动**，如在模型输入中拼接部分其他字符可形成对抗样本，既可以让 ChatGPT 输出失败。3）**模型参数修改**，复旦大学等 [2-55] 发现修改 LLaMA2-13B 模型一个参数，即可导致该模型语言能力丧失，发现了 LLM 中与语言能力相对应的核心区域，约占模型总参数的 1%。该核心区域表现出显着的维度依赖性，即使特定维度上的单个参数的扰动也可能导致语言能力的丧失。

大模型安全防御技术可分为两大类：提示注入防御和输出内容水印技术。1）**提示注入防御技术**主要包括输入侧防御和输出侧防御。输入侧防御通过提示过滤，检测并过滤可能含风险的用户输入，如注入攻击或敏感内容，以防止这些输入与大语言模型或相关软件交互。提示增强技术则通过构建更鲁棒的提示来抵抗注入攻击，利用大语言模型的理解能力进行“自我增强”，在提示词中加入任务内容和用户输入内容的强调，提高系统提示的精确度。提示增强分为语义增强和结构增强。而输出侧防御则采用内容审核过滤方法，通过规则或模型识别，避免输出风险内容，保障内容安全。2）**输出内容水印**，包括

明水印和隐水印，用于保护知识版权和防止模型输出被恶意使用。这些水印技术在模型服务界面上标记内容来源，以提示和追踪目的，防止内容恶意传播。例如，马里兰大学提出在模型解码阶段加入水印，通过特定算法检测文本水印以确定来源；腾讯则提出可编码水印技术。尽管这些方法在实验中有效，但在实际应用中的辨识率尚不能完全保证，目前还没有非常成功的隐水印技术。

**模型安全评测**。随着大模型能力的不断增长，确保其安全、可靠和符合伦理标准的运行变得至关重要。大模型的安全评估不仅为开发人员、政策制定者和其他利益相关者提供了关于模型性能和风险的深入了解，而且有助于整个社会创造了一个更加安全、透明和可信赖的 AI 环境。1）清华大学联合聆心智能于 2023 年 3 月推出面向中文大模型的内容安全性评测平台 [2-48]。该平台依托于一套系统的安全评测框架，从辱骂仇恨、偏见歧视、违法犯罪等 8 个典型安全场景和 6 种指令攻击两个角度综合评估大语言模型的安全性能。其中，指令攻击是指一般模型难以处理的安全攻击方式，包含目标劫持、Prompt 泄露、角色扮演指令、不安全 / 不合理的指令主题、隐含不安全观点的询问、以及 反面诱导。基于该框架，平台对 GPT 系列、ChatGLM 等主流大模型进行了安全评估，并发现指令攻击更有可能暴露所有模型的安全问题；2）2023 年 5 月，DeepMind 联合 OpenAI、Anthropic、多伦多大学及牛津大学等科研机构和高校，提出一个针对新型威胁评估通用模型的框架，将大模型安全评估分为两类：①评估模型是否具有某

些危险的能力；②判断模型多大程度上可能使用这些能力造成伤害。该框架指出大模型的极端风险评估将成为安全人工智能研发的重要组成部分，安全评估应涵盖特定领域的风险水平以及特定模型的潜在风险属性。评估结果可以帮助开发者识别可能导致极端风险的因素。3）考虑到越来越多的大模型被训练应用于真实世界的交互任务，2023 年 6 月，伯克利提出大模型行为决策的道德评估基准 MACHIAVELLI，以衡量大模型在各种社会决策场景中的能力和道德行为。该项评估主要基于一套由 134 款基于文本的 Choose Your Own Adventure 游戏组成，在评估中为大模型代理提供真实世界的目标，并通过专注于高层次的决策来追踪代理的不道德行为，以评估其在现实社会环境中的规划能力及安全风险。




图表 2-23 中文大模型安全评测框架 [2-48]

### 2.3.4 代表性大语言模型

图表 2-24 给出了国内外典型大语言模型产品及其所依赖的基础大模型，下面对部分模型产品进行介绍：

| | 大模型 | 代表产品 | 最早上线时间 | 机构 |
|---|---|---|---|---|
| 国外 | GPT 系列 | ChatGPT、GPT-4 | 2022.12 | OpenAI |
| | PaLM 系列 | Bard | 2023.2 | Google |
| | Claude 系列 | Claude、Claude2 | 2023.7 | Anthrophic |
| 国内 | ERNIE系列 | 文心一言 | 2023.9 | 百度 |
| | 星火大模型 | 星火认知 | 2023.9 | 科大讯飞 |
| | 混元大模型 | 腾讯混元 | 2023.9 | 腾讯 |
| | 通义大模型 | 通义千问 | 2023.9 | 阿里 |
| | 云雀 | 豆包 | 2023.9 | 字节跳动 |
| | GLM 系列 | 智谱清言 | 2023.9 | 智谱AI |
| | Baichuan 系列 | 百川大模型 | 2023.9 | 百川智能 |
| | CPM 系列 | 面壁LUCA | 2023.11 | 面壁智能 |

图表 2-24 国内外典型大模型产品

- **GPT 系列**：由 OpenAI 推出的基于 Transformer Decoder 自回归架构的生成式模型框架，在此基础研发了系列大模型 GPT-1、GPT-2、GPT-3、Codex 等，在 2022 年 11 月推出大语言模型产品 ChatGPT，基础模型参数 20B，采用 InstructGPT 技术，即预训练 +

有监督微调（SFT）+ 人类反馈强化学习 RLHF 技术训练，支持问答、编程 coding、写作等各种任务；2023 年 3 月 GPT-4，是一种支持图文跨模态输入的多模态大模型，在推理方面的能力比 ChatGPT 更强，同时也减少了幻象的产生，能够更准确地理解和回应复杂的问题，从而提供更高质量的答案。

- **Claude 系列**：Claude 系列模型是由 Anthropic 开发的闭源大语言模型，2023 年 3 月发布大语言模型产品 Claude-1，7 月更新至 Claude-2。该系列模型通过预训练、RLHF 和“宪法人工智能（Constitutional AI）”安全技术进行训练，旨在改进模型的有用性、诚实性和无害性。支持最大长度 200k token 的上下文。Anthropic 是一家 AI 安全和研究公司，愿景是构建可靠的、 可解释的和可操控的（Steerable）AI 系统。创始团队来自 OpenAI，是其最强竞争对手。

- **文心一言**：文心一言由百度公司研发，是基于百度知识增强大语言模型 ERNIE，于 2023 年 3 月在国内率先开启邀测。8 月 31 日，文心一言率先向全社会全面开放，提供 APP、网页版、API 接口等多种形式的开放服务。采用有监督精调、人类反馈强化学习、提示等技术，还具备知识增强、检索增强和对话增强等关键技术。文心一言基于“飞桨”深度学习框架进行训练。文心一言还建设了插件机制，通过外部工具、服务的调用，拓展大模型的能力的边界。

- **智谱清言**，建立在 GLM 基础大模型（GLM-130B）基础上的智

能对话助手，清华系智谱 AI 公司发布，2023 年 9 月 1 日在网页端、公众号、APP 同时上线，国内首批上线的大模型产品。大模型对华为昇腾 ASCEND、神威超算、海光 HYGON 等多个国产化平台进行了适配。

- **百川大模型**：建立在 Baichuan 基础大模型（Baichuan2-53B）基础上构建的智能对话助手，由清华系百川智能公司开发，2023 年 9 月在网页端、公众号同时上线，其基础大模型采用 Transformer 解码器架构。Baichuan2-53B 融合了意图理解、信息检索以及强化学习技术，结合有监督微调与人类意图对齐，在知识问答、文本创作领域表现突出。

- **面壁 LUCA**：建立在自研 CPM 基础大模型的智能对话助手，清华系面壁智能公司发布，2023 年 11 月 4 号上线，国内第二批上线产品，目前只上线网页端。采用预训练 + 后预训练 +SFT+ 人类反馈微调 + 安全对齐，基于深度学习框架 BMTrain 进行训练。

## 2.3 AI Agent

以 GPT-4 为代表的大语言模型，展现了复杂指令遵循、思维链推理和认知交互能力，催生了 AI Agent 领域的研究和应用热潮。AI Agent 的进步不仅仅体现在技术层面，它也在重塑我们与计算机系统的互动方式，影响着社会、经济和文化的各个方面。本节介绍 AI Agent 的概念、发展趋势、关键技术和代表性智能体情况。

## 2.3.1 AI Agent 的发展演进

### 2.3.1.1 AI Agent 基本概念

**AI Agent**，或称**人工智能体**，是一种能够感知环境、进行决策、执行动作完成既定目标的智能实体。不同于传统的人工智能，AI Agent 具备通过独立思考、调用工具或使用技能去逐步完成给定目标的能力。AI Agent 和大模型的区别在于，大模型与人类之间的交互是基于提示（Prompt）实现的，用户提示是否清晰明确会影响大模型回答的效果，而 AI Agent 的工作仅需给定一个目标，它就能够针对目标独立思考并做出行动。

大语言模型作为目前 AI Agent 的核心，以巨大参数规模捕捉复杂语言结构，实现上下文理解和连贯文本输出。这一“能力涌现”现象体现在大模型能进行高级认知任务，如抽象思考和创造性写作。AI Agent 不仅理解和生成语言，还整合规划、记忆、工具使用能力，扩展其能力边界。

### 2.3.1.2 AI Agent 的类型

在人工智能领域，AI Agent 可以根据其运作模式和应用范围被划分为两大类：

**1) 自主型 AI Agent，**也称单边型 AI Agent，是指那些能够独立运作，完成特定任务的智能体。这类 Agent 拥有独立的决策能力，能够基于输入的数据或观察到的环境自行做出响应。它们通常被设计用于特定的应用场景，例如个人助理、智能推荐系统或特定领域的问题解答。自主型 Agent 的核心特点是能够在没有外部指令或者很少人工干预的情况下，完成复杂的任务。代表性 Agent 为 AutoGPT、XAgent 等。

**2) 协同型 AI Agent** 则是指在一个系统中多个智能体协同工作，共同完成任务的情形。这类 Agent 的特点是群体智能——单个 Agent 的能力可能有限，但当它们作为一个集体工作时，能够处理更为复杂、多样的任务。协同型 Agent 在处理需要多方面协作和信息共享的任务时表现尤为出色，如多 Agent 系统在自动化工厂、交通管理等领域的应用。代表性 Agent 为 Smallville、ChatDev、AgentVerse、MetaGPT 等。

在实际应用中，这两种类型的 Agent 都在不断发展和完善，以适应日益复杂和多样化的应用需求。随着技术的进步，两者之间的界限也在逐渐模糊，例如某些系统可能同时采用自主型和协同型 Agent 来达成更加复杂的目标。

#### 2.3.1.3 AI Agent 的智能分级

AI Agent 根据其人工智能水平可以大致四级，可以从感知能力、认知能力、执行能力、规划能力等维度进行阐释，图表 2-25：

| 等级 | 感知能力 | 认知能力 | 执行能力 | 规划能力 |
| --- | --- | --- | --- | --- |
| L1（部分自动化） | “所见即所得”的感知，处理单一模态下的相对简单的数据类型，应用于简单场景。 | 利用大量人类监督信号获得的一定程度的理解语言、利用语言人机交互能力。 | 少量的常见标准工具的调用，简单的工具调用逻辑。 | 静态地执行特定的、预定义的任务。涉及少量的、简单串并联的流程节点。 |
| L2（有条件自动化） | 多模态感知能力，能处理更广泛的数据类型，应用于更多样、更长尾、更复杂的场景。 | 全面的认知能力，包含记忆能力、决策能力、高度智能的对话能力、内容生成能力。 | 可使用的工具数量、类型、实现的业务逻辑的复杂度得到极大提升。 | 以业务流程达到端到端最大化自动化为目标，可以规划和编排大量流程节点和复杂逻辑。 |
| L3（高度自动化） | 综合利用认知能力，环境交互结果，在少量人类干预下获得超高精度的感知力。 | 通过综合利用环境知识、人类少量的监督信号，达到高精度的认知水平。 | 在人类少量干预下，可以实现绝大多数工具调用代码。 | 能够主动洞察问题域和求解域的环境变化，实现业务流程的灵活适应和编排，环境适应能力强。 |
| L4（完全自动化） | 在无人工干预下智能体自主进化获得超高精度的感知能力。 | 利用环境信号自主学习提升认知水平。 | 能自动学习工具使用的方式、实现100%的自动化调用工具的能力。 | 能利用过程反思、经验沉淀，难例挖掘等高度智能化的决策机制，自主提升规划和编排能力，自主进化。 |

图表 2-25 AI Agent 智能等级划分

- **L1 级别：**这是智能体的早期形态，通过整合传统的视觉能力、语义理解能力、RPA 流程自动化能力，完全由领域专家来实现既定业务流程的定义和编排以解决单点的、简单明确的任务。代表性的智能体包括以 UiPath 为代表的传统 RPA 机器人、NICE 的桌面机器人 NEVA 为代表的 ChatBot。

•**L2 级别:** 在这个级别上，智能体上述四个方面的能力实现了从“局部能力”到“全面能力”的跃迁，这使得其能解决的问题也从“单点任务”推广到了“端到端任务”。代表性的智能体包括 XAgent、AutoGPT 等。但是，实现这一效果是有“条件的”。人类需要将自己掌握的业务见解、行业 know-how、期望目标等世界知识以复杂指令的形式告诉智能体。这种高昂的教育成本会一定程度上限制智能体的普及。

• **L3 级别：**在这个级别上，智能体可以有效洞察问题域的环境变化，然后主动利用求解域中人类碎片化的历史经验、监督信号，智能化地探索、理解、学习问题解决的方法，达到“沧海拾贝”、“睹微知著”的效果。在 L2 级别的基础上，不仅可以 降低人工干预的程度，而且可以得到更高的任务完成率、准确率。L3 智能体的“高精度”、“少干预“、”快适应“的特性使其在市场需求和技术演进高度动态的商业环境下，具有巨大的价值。

• **L4 级别：**这是最高级别，智能体具备自学习和自组织的能力，能在无人干预的情况下完成感知、认知、执行、规划等能力的自我进化。该级别基本代表了通用人工智能 AGI 和类人智能，对应自动驾驶自动化水平的高度的完全自动化（L5）级别。目前还处于初步探索阶段，具有潜力的智能体工作如 OpenAI 的 Q* 项目、大模型群体智能技术 AgentVerse 等。

### 2.3.1.4 AI Agent 发展趋势和挑战

**（1）AI Agent 发展趋势** AI Agent 并不是一个新兴的概念，早在 1980 年代已在人工智能领域有了研究，其发展演进与人工智能技术演进密切相关，大致可以分为 4 个阶段：

**1）基于符号规则的智能体阶段（1980 年前后）：**采用逻辑规则和符号表示来封装知识和促进推理过程。早期符号型智能体的典型例子如医学诊断专家系统 MYCIN、模拟心理治疗师 ELIZA 等专家系统。其中 MYCIN 用于医学诊断，尤其是在感染性疾病和抗生素选择方面，它采用了逻辑规则和符号表示。

**2）基于统计学习的智能体阶段（~1990-2000 年）：**主要关注智能体其环境之间的交互，强调快速和实时响应，缺乏复杂决策和规划能力，该阶段采用统计学习模型基于数据和环境交互进行学习。该阶段的典型例子如麻省理工大学的行为基础机器人 Genghis，它们通过简单的感知和动作规则与环境交互，而不是通过复杂的模型和规划。

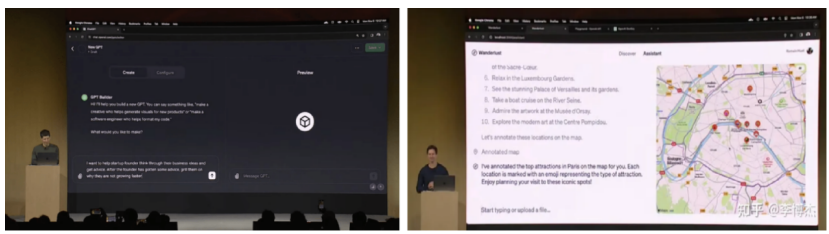
**3）基于深度学习的智能体阶段（~2000-2020 年）：**采用深度学习模型作为智能体控制模型，通过智能体与环境交互获得反馈优化深度学习模型实现对复杂环境适应。2014 年由 DeepMind 推出的引发全球热议的围棋机器人 AlphaGo，其采用强化学习方法训练深度学习模型。与之类似的还有 2017 年 OpenAI 推出的用于玩《Dota2》的 OpenAI Five，2019 年 DeepMind 公布用于玩《星际争霸 2》的 AlphaStar 等，这些 AI 都能根据对实时接收到的信息的分析来安排和规划下一步的操作，均采用了强化学习的方法构建。当时的业界潮流是通过强化学习的方法来对 AI Agent 进行训练，主要应用场景是在游

戏这类具有对抗性、有明显输赢双方的场景中。但如果想要在真实世界中实现通用性，基于当时的技术水平还难以实现。

**4）基于大模型的智能体阶段（2021-2023 年以及之后）：**2021 年底 OpenAI 发布大模型 WebGPT，使用 GPT-3 模仿人类操作搜索引擎和 Web 浏览器进行长文本问答，模型表现取得了令人惊艳的效果，同时展现了大模型的认知交互能力。2022 年底 ChatGPT 展现了大语言模型强大的语义理解和通用任务处理能力，让人们看到了构建 AI Agent 完成复杂任务巨大潜能，激发对 AI Agent 的研究热潮。之后，OpenAI 推出智能体构建平台 GPTs、游戏公司 Significant Gravitas 的自主智能体 AutoGPT、清华 & 面壁的自主智能体 XAgent 等，，采用 LLM 作为智能体的大脑，通过感知、规划、工具使用、记忆等实现复杂任务处理。AI Agent 成功为群体智能构建提供了有力支撑，多个 AI Agent 之间可以通过协同互补，完成超越单智能体的更高阶的复杂任务，如软件开发、社会模拟等。


图表 2-26 AI Agent 的演化历程

**AI Agent 成为目前各大科技巨头布局的新风口。**比如微软推出了 AutoGen、谷歌 Deepmind 推出了 Robotic Agent、亚马逊推出了 Bedrock Agents、NVIDIA 打造 Voyager 智能体游玩《我的世界》，阿里云 ModelScopeGPT、斯坦福与谷歌联合搭建的虚拟小镇 Smallville 等等，同时，OpenAI 也已然奔赴至 Agents，在开发者大会上推出 GPTs+Assistant API 的智能体构建平台，并在 2024 年 1 月推出 GPTs Store，OpenAI 进入其“iPhone 时刻”。目前 AI Agent 被认为是大语言模型的下半场。微软公司创始人比尔·盖茨在其个人网站撰文，阐述智能体技术将在未来数年中变革计算机使用模式。



图表 2-27 OpenAI 发布 AI Agent 产品

**从大模型“单体智能”到大模型“智能群体化”，再到千行百业。**随着基础模型能力的不断加强，以及在应用场景的深入探索，大模型 + Agent 会引起新一轮的应用爆发，为行业和用户带来更多新的能力、功能和服务。


图表 2-28 大模型 AI Agent 应用发展趋势

### （2）AI Agent 的挑战

当前 AI Agent 的挑战主要集中在以下几个方面：

**1) 大模型的 Agent 能力不足：**目前国内外 AI Agent 研究和实践主要基于 OpenAI GPT-4 实现，大模型的复杂指令遵循、规划、思维链推理、长期记忆等能力是 AI Agent 执行成功的关键，但目前国内大模型在支持 AI Agent 能力上还存在显著不足。

**2) Agent 标准和规范缺乏：**目前 AI Agent 发展迅速，不同研究机构和公司推出自己的大模型、Agent、工具链等，缺乏统一的接口标准和通信协议。

**3) 系统安全管理问题:** 随着 AI Agent 在各行各业的应用日益广泛，确保其安全性和可靠性变得尤为重要。这包括保护系统免受外部攻击、防止数据泄露、避免“幻觉”问题、确保 AI 决策的透明性和可解释性，

以及在多变环境中的稳定运行。安全管理不仅关系到技术的稳定性和可靠性，也关系用户信任和技术的广泛接受度。

**4) 多模态感知与交互能力不足：**人类通过多模态方式感知世界，而当前的 AI Agent 主要依赖于文本输入。其在视觉、听觉等方面的处理能力还需要进一步发展。这意味着 AI Agent 在模拟人类感知世界的方式方面还有很大的进步空间，尤其是在多模态数据处理和解释上。

**5) 社会化能力与伦理问题：**目前 AI Agent 在社会行为、人格特征以及认知、情感和性格模拟方面还处于起步阶段。随着技术的发展，更多的伦理和社会学问题将浮现，如虚拟人与社会人的关系，以及在拟人个性化对话场景中的安全性和可信度问题。这些挑战涉及到伦理、社会学和经济学领域，需要跨学科的合作和研究。

**6) 智能体部署成本高：**目前 AI Agent 性能仍然依赖于大模型，但大模型 API 调用成本高昂，难以支持大规模商业化部署，需要小模型解决方案，但目前小型化模型的 Agent 能力还不具备。

**7) 拟人化单体智能：**拟人化单体智能具备 6 大典型特性：构建具备智商 ( 知识 / 认知 / 行动能力 )、情商 ( 具备情绪感知和共情能力 )、人设（人物性格 / 特色 / 背景）、感知（多模态感知能力）、价值观（道德 / 价值取向 / 安全等）、成长性（进化、自适应和自学习）等特性的智能体，可以适应更复杂的场景应用，具有重要市场应用前景，目前 AI Agent 的拟人化技术进展显著，但相关技术还很初步。


图表 2-29 拟人化单体智能的典型特征

## 2.3.2 AI Agent 技术框架

在以大模型为核心的自主智能体系统中，除了大模型作为核心之外，同时配备了几个关键组件：


图表 2-30 AI Agent 一般技术框架 [2-47]

**1) 规划（Planning）：**为了完成复杂任务，智能体需要将该任务分解成更小、可管理的子目标，以高效处理复杂任务。同时还需要对自身过去的行为进行批评和反思，从错误中学习并改进，为未来步骤提升结果质量。

**2) 执行（Action）：**为了完成任务，AI Agent 需要切实采取行动，一步步达成目标。在这一环节中，执行工具是一种十分重要的执行能力。AI Agent 需要学习调用外部 API 获取额外信息，并真实采取举措，如代码执行能力。

**3) 感知（Perception）：**AI Agent 需要扩展自身的感知范围，除了文字还需要理解图像、音频等信息。这种扩展的感知范围帮助智能体更好地理解环境，做出更加明智的决策。

**4) 记忆（Memory）：**AI Agent 在完成任务的过程中，需要设计如何更好地利用历史信息，所以需要构建起一个记忆机制对信息进行高效管理与利用。这通常包含两个部分：短期记忆——暂时存储和处理当前的输入信息，帮助进行任务执行和问题求解。这种记忆形式有助于 AI Agent 在处理语言、理解上下文、执行连续的任务以及进行交互时更加高效；长期记忆——这使得智能体具备在较长时间内保留和回溯信息的能力，通常通过外部向量存储和快速检索实现。

**5) 工具使用（Tool Use）：**Agent 学习调用外部应用程序的 API，以获取模型训练数据权重中缺失的“额外信息”（任务相关，预

训练后通常难以更改），包含当前信息、代码执行能、专有信息源的访问权限等。

## 2.3.3 AI Agent 关键技术

### 2.3.3.1 大模型工具学习

大语言模型具备理解、推理和决策能力，可与外部工具互动。在特定领域任务中，如金融领域的证券交易和市场预测，大语言模型通常需要结合外部工具获取信息和技能才能处理。整合外部工具与大语言模型可以发挥各自优势实现复杂任务的处理，其中外部工具可增强专业知识和可解释性，大语言模型提供语义理解和推理规划能力。

2021 年底，OpenAI 推出 WebGPT[2-31]，利用 GPT-3 与网页浏览器和搜索引擎交互获取互联网信息在长文本问答上实现非常强的能力，展现了大语言模型利用工具解决复杂问题的巨大潜力。该工作引起了学术界和产业界的广泛关注，产生了许多面向不同任务或场景需求的大模型调用工具的方法，如 Webshop[2-32]，使用大语言模型替代人在购物平台上执行一系列操作、购买所需物品。2023 年 3 月，OpenAI 发布 ChatGPT Plugins[2-33]，实现 ChatGPT 调用各种外部插件的功能，支持浏览器实时信息获取、代码解释器、PDF 阅读等能力，截至 8 月已支持 480 个常用工具插件。Meta 将这种通过非参数的外部模块扩展大语言模型能力的方法，统一称为增广语言模型（Augmented Language Models）[2-34]。清华大学在现有大模型

工具使用方法基础上，提出了工具学习（Tool Learning）框架 [2-35]，指在让模型能够理解和使用各种工具完成任务的学习过程。


图表 2-31 基于用户接口视角的工具分类 [2-35]

目前可交互的通用工具按用户接口大致可分为三类（图表 2-31）：物理交互的工具（如机器人、传感器等）、基于图形用户界面的工具（如浏览器、Office 办公软件等）、基于编程接口的工具（如数据库、知识图谱）等。从学习目标的角度来看，现有工具学习方法主要可以分为两类 [2-35]：一类是工具增强学习（Tool-augmented Learning），利用各种工具的执行结果，增强基础模型性能。在这一范式中，工具执行结果被视为辅助生成高质量输出的外部资源；第二类是工具导向学习（Tool-oriented Learning），将学习过程重点从增强模型性能转向工具执行本身。这一类研究关注开发能够代替人类控制工具并进行序列决策的模型。

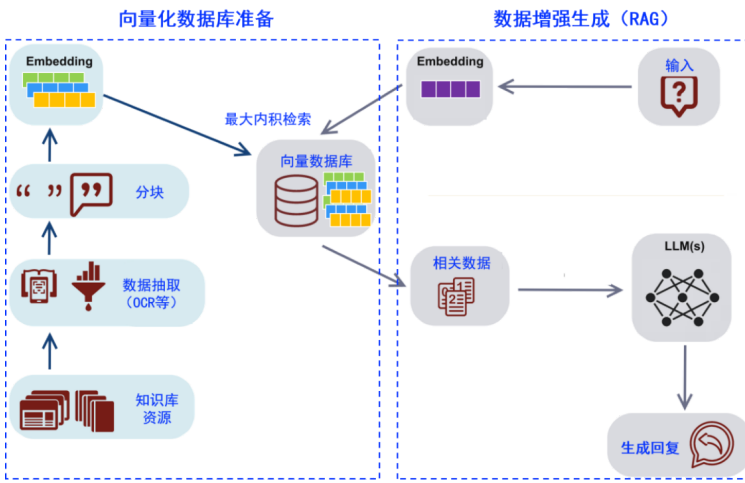
从目前来看，LLM 工具学习已经取得了显著的进展，相关应用处在爆发上升趋势，已展现广阔的应用前景。随着大模型性能不断提升，

给工具学习带来许多机遇和挑战 [2-35]：（1）工具学习的安全性。在期待 LLM 与工具学习结合所带来的生活改变之前，审视其中潜在的风险尤为重要。需要防止恶意用户误导模型调用工具，以及提升模型使用工具的可信度等问题；（2）工具 AI 创造，LLM 可能具有自发创造工具的潜力。一直以来，创造和使用工具被认为是人类智能的独特特征，而 LLM 的出现可能颠覆这一观念。越来越多的证据表明，创造工具的能力不再是人类专有的领域；（3）知识冲突，引入工具后，模型需要解决来自不同来源的知识冲突问题，包括模型自身、外部知识库等。解决不同知识库间的知识冲突，以实现知识的有效整合，是迎接工具学习挑战的关键一步；（4）多工具协同，一个复杂任务通常需要多种类型工作协同配合完成，需要让大模型学会规划和执行多类型工具完成复杂任务。未来，我们预期工具学习将会进一步融合更多的工具和领域知识，从而为 LLM 带来更大的突破。

#### 2.3.3.2 检索增强生成

检索增强生成（Retrieval-Augmented Generation，RAG）是一种结合检索和生成的深度学习方法，用于增强大语言模型的任务处理能力，是 AI Agent 的实现长期记忆的关键技术。RAG 的核心是向量数据库技术，这是一种存储和检索大量信息的高效方式。在 RAG 模型中，首先利用一个检索器从一个预先构建的向量数据库中检索相关信息。这个数据库通常包含大量文本数据的向量表示，这些向量是通过

预训练的语言模型生成的。检索过程基于查询向量和数据库中的文档向量之间的相似性。检索到的信息随后被送入生成器，生成器是基于 Transformer 架构的神经网络，它综合检索到的信息和原始输入来生成响应或回答。这个过程可以大大提高大语言模型的生成性能，因为它允许模型利用数据库中的丰富信息，提供更准确和信息丰富的输出。RAG 模型的一个关键优势是它能够处理更复杂、开放式的问题，因为它可以访问和利用比传统模型更大量的外部知识库数据。此外，向量数据库的使用使得检索过程更高效，因为相似性搜索可以迅速在海量数据中找到最相关的信息。




图表 2-32 基于向量化数据库检索的 RAG 技术框架

向量数据库通过将文档数据转化为向量存储，解决大模型海量知识的存储、检索、匹配问题。向量是 AI 理解世界的通用数据形式。向量数据库利用人工智能中的语义嵌入（Embedding）方法，将图像、

音视频等非结构化数据通过预训练的神经网络抽象、转换为高维语义向量，由此实现了知识的结构化管理，从而实现快速、高效的数据存储和检索过程，赋予了 AI Agent“长期记忆”。同时，将高维空间中的多模态数 据映射到低维空间的向量，也能大幅降低存储和计算的成本：向量数据库的存储成本比直接将数据训练到神经网络的参数中的成本要低 2 到 4 个数量级。代表性的向量数据库包括 Pinecone、Weaviate、Milvus、Qdrant、Chroma、Elasticsearch、PGVector、Typesense 等，其特征情况如图表 2-33 所示。

| | Pinecone | weaviate | milvus | qdrant | chroma | elasticsearch | pgvector | typesense |
|---|---|---|---|---|---|---|---|---|
| 是否开源 | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| 云管理 | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ |
| 面向向量构建 | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| 开发经验 | ★★★ | ★★ | ★★ | ★★ | ★★ | ★ | ★ | ★★ |
| Github星标 | N/A | 8.4k | 24.6k | 15k | 10k | 66.1k | 7.1k | 16.2k |
| 编程语言 | Rust,Python,Node.js | Go | Go,Python,C++,Go,Note.js | Rust,Go,Python | Python,JavaScript | Python,Java,Go等 | C,Perl | C++,C |
| 每秒可处理查询 | 150 | 791 | 2406 | 326 | N/A | 100-700 | 141 | 1642 |
| 索引类型 | N/A | HNSW | HNSW等多种 | HNSW | HNSW | HNSW | HNSW/IVFFlat | HNSW |
| 数据规模 | >10B | 10B | 10B | 10B | N/A | 1M | N/A | 10M |
| 混合检索 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

图表 2-33 代表性向量数据库情况

Embedding 技术和向量相似度计算是向量数据库的核心。Embedding 技术是一种将图像、音视 频等非结构化数据转化为计算机能够识别的语言的方法。在通过 Embedding 技术将非结构化数据

例如文本数据转化为向量后，就可 以通过数学方法来计算两个向量之间的相似度，即可实现对文本的比较。向量数据库强大的检索功能就是基于向量相似度计算而达成的，通过相似性检索特性，针对相似的问题找出近似匹配的结果。相似性向量检索采用最大内积搜索（Maximun Inner Product Search，MIPS）。通过使用外部存储器可以缓解关注范围有限的限制。一种标准的做法是将信息的嵌入表示法保存到向量数据库中，该数据库能够支持快速的最大内积搜索。为了优化检索速度，常见的选择是近似相邻（Approximate Nearest Neighbors，ANN）算法，返回近似的 top k 个近邻，用损失少量的精度来换取速度的巨大提升。几种常见的快速最大内积搜索算法如局部敏感的哈希算法（LSH）、层次导航最小世界算法 (Hierarchical Navigable Small World，HNSW )、FAISS 、尺度化最近邻算法 (Scalable Nearest Neighbors，ScaNN ) 等。

#### 2.3.3.3 长序列流式输入处理

在 AI Agent 应用中，Agent 需要处理长期历史信息，调用大量工具链，接受持续环境输入。但是大语言模型由于输入长度限制难以直接处理长序列输入，通常仅支持几千 token 的序列长度，如 LLaMA2 最大支持 4096 tokens、GLM-130B 最大支持 2048 tokens。为此，支持长序列流式输入的大语言模型技术被提出，代表性技术有三大类，分别是位置编码拓展、全局注意力有损改进、新型注意力机制设计：

第一类是基于位置编码拓展的方法，将通常使用的旋转位置编码（RoPE）经过直接放缩或频域放缩的方法，使模型的最大处理长度变长数倍，这种方法在工程上有较广泛的应用，能够支持十万左右的输入序列长度，但是其核心没有突破全局自注意力机制的平方复杂度，需要消耗巨大的显存，并且仍然有最大处理长度的限制，无法处理超长流式输入。

第二类方法是对 Transformer 的全局自注意力机制进行有损的改进。包括使用滑动窗口，限制每个 token 只能看到自己邻近的位置的 token，以此避免模型处理超过训练阶段的相对位置编码，通过牺牲效果弥补模型生成的稳定性。后续工作进一步通过设计特殊的注意力遮蔽矩阵，避免在使用上述滑动窗口过程中模型注意力塌陷的问题。这类方法虽然表面上支持模型接受流式输入，但是对于滑动窗口外的内容，模型将完全遗忘，无法支持模型形成长期记忆能力。

第三类方法则是通过抛弃 Transformer 的全局自注意力机制，设计新的信息处理机制来处理长程流式输入。最早的工作可以追溯到 Linear Transformer 设计的线性复杂度注意力机制，随后出现了 RWKV、RetNet 等模型，这些模型都具有亚平方复杂度注意力机制。此类模型都会在隐状态中形成可以总结历史所有信息的内在状态（Internal State），使模型的长期记忆成为可能。

#### 2.3.3.4 智能体自适应和自学习

智能体能够根据环境和任务的动态变化不断提升智能水平，适应不同复杂场景的需求，是实现类人智能的重要标志。相关方法可以分为三类：

- **无参数优化自进化方法**。一种是构建本地技能库（Skill Library），如 NVIDIA 提出 Voyager 智能体架构，由自动课程、技能库和迭代 prompt 机制三个新型组件构成。自动课程用于提出开放式的探索目标，该课程是由 GPT-4 根据“尽可能多发现不同的东西”的总体目标生成的，会根据探索进度和 Agent 状态使得探索实现最大化；技能库用于开发越来越复杂的行为，通过存储有助于成功解决某个任务的行动程序，Voyager 逐步建立起一个技能库，未来可以在类似情况下进行检索，实现能力随着时间的推移迅速增强，并缓解“灾难性遗忘”问题。迭代 prompt 机制引入了环境反馈、执行错误和检查任务是否成功的自我验证三种类型的反馈，根据这些反馈，GPT-4 可以自己去迭代更新 prompt，直到生成的 prompt 足以去完成当前任务。此外，另一种构建本地记忆（Memory）方法，这是一种可以帮助智能体积累经验并实现自我进化的技术。通过这种方法，智能体能够以更加一致、合理、有效的方式完成任务。这种进化通常是通过不断的学习和对经验的积累来实现的。如经验记忆强化学习技术 RLEM 通过强化学习使得智能体能够在交互中，根据当前交互状态从经验记忆中动态抽取过往经验来提升自身的交互行为，同时还可以利用环境返回的回报（reward）来更新经验记忆，使得整体策略得到持久改进。

- **参数优化自进化方法**：Agent 通过探索，获得环境或人类反馈信息，借助反馈强化学习对 LLM 进行微调实现模型能力持续增强。这

里微调是全参数微调或部分参数高效微调，后者由于微调效率高、显存占用低，更加适合面向大规模用户的个性化支持部署。如深度自进化强化学习框架 DRRL，这是由斯坦福大学的李飞飞等学者提出，基于这个框架，所创建的具身智能体可以在多个复杂环境中执行多项任务。这种方法融合了深度学习和进化算法，对智能体神经控制器参数进行更新，使智能体能够适应并在各种环境中进行自我提升。

- **自适应技术**：如进行自查与自纠，Agent 能够对过去的行为进行自我批评（Self-criticism）和自我反省（Self-reflection），从错误中吸取教训，并在今后的工作中加以改进，从而提高最终结果的质量（本质上是产生强化学习的数据，强化学习并不需要人类反馈）。如 ReAct 通过将动作空间扩展为特定于任务的离散动作和语言空间的组合，将推理和动作集成在 LLM 中。前者使 LLM 能够与环境交互，而后者则促使 LLM 以自然语言生成推理痕迹。AutoGPT 对生成的代码进行自动调试，并根据动作执行结果和环境反馈信息放入上下文指导调整 LLM 下一步行为。另一种在部署 Agent 之前，通过构建工具学习、Agent 行为等数据集，通过模仿学习、指令学习等增强 LLM 的对环境、工具使用、API 调用等泛化能力，代表工作如 AgentBench、LLMBench 等。

#### 2.3.3.5 AI Agent 能力评测

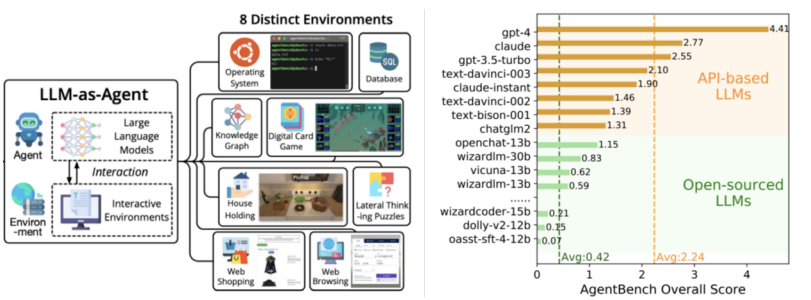
在 AI Agent 系统中，大模型的 Agent 能力是系统成功运行的关键，也是系统人工智能水平的决定因素。因此，对 LLM 的 Agent 能力进行科学评测至关重要，其作用包括：1）理解 LLMs 的实际应用能

力：随着 LLMs 在不同领域的广泛应用，理解它们在实际环境中的表现变得至关重要。通过对这些模型作为智能代理的能力进行评测，我们可以更好地了解它们在现实世界任务中的表现，如自动编程、数据分析、游戏玩法或网页浏览等。2）识别和解决限制：通过评测可以发现 LLMs 在作为代理时的局限性，如推理、决策制定、长期记忆和多轮对话管理等方面的不足。识别这些问题是提高它们性能和可靠性的第一步。3）推动技术发展：评测不仅有助于现有模型的优化，还能激发新技术的发展。通过对 LLMs 的能力进行系统性的评估，研究者和开发者能够发现新的研究和改进方向，推动人工智能技术的进步。4）促进安全和可靠的应用：了解 LLMs 作为智能代理的能力有助于确保它们在实际应用中的安全性和可靠性。这对于减少错误和不当使用，保障用户利益至关重要。

2023 年 7 月，清华大学 & 面壁智能提出了大模型工具学习数据集自动构建框架 ToolLLM，在此基础上构建了可支持多类型工具使用的大模型评测基准 ToolBench，覆盖 1.6 万的 API，仓库中包含 46.9 万次真实 API 调用得到的 1.2 万条样本，涵盖单工具场景和多工具场景。ToolLLM 框架首先从 RapidAPI Hub 收集大规模真实世界 RESTful API，然后提示 ChatGPT 生成涉及这些 API 的多样化人类指令。最后，使用 ChatGPT 为每个指令搜索有效的解路径（一系列 API 调用）。该研究在 ToolBench（指令调优数据集）上对 LLaMA 进行微调，得到了 ToolLLaMA。开发了高效的机器评估工具 ToolEval，其依赖于 ChatGPT 的支持，并包含两个关键评估指标：（1）通过率，用于衡量在有限预算内成功执行指令的能力，以及（2）胜率，用于比较两条解路径的质量和有用性。评测显示 ToolLLaMA 展现出了出色的执

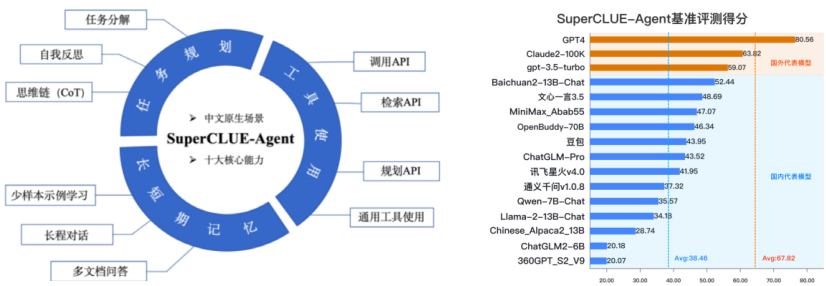
行复杂指令和泛化到未知 API 的能力，并且在工具使用方面性能与 ChatGPT 相媲美。

2023 年 8 月，清华大学 & 智谱 AI 推出大模型 Agent 能力基准评测 AgentBench[2-52]，通过一个多维度的基准测试来评估 LLMs 在各种环境下的推理和决策能力，特别是在多轮开放式生成设置中的表现。其包括 8 个不同的环境，涵盖了从操作系统和数据库的技术性任务到数字卡牌游戏和网页浏览等更多元化的场景。这些环境被设计为测试 LLMs 在各种实际任务中的表现，如逻辑推理、指令遵循、知识获取等。该评估技术对 27 个基于闭源大模型 API 和开源的 LLMs 进行了广泛的测试。结果显示，尽管顶级商业 LLMs 如 GPT-4 在复杂环境中表现出色，具备了处理真实世界环境交互的强大能力，但大多数开源 LLM 在 AgentBench 中的表现远不如闭源商业大模型（平均分为 0.42 对比 2.24）。AgentBench 强调了 LLMs 在长期推理、决策制定和任务完成方面的不足，这些都是将 LLMs 作为有效代理的主要障碍。AgentBench 的设计和应用表明，通过在多样化的实际环境中测试 LLMs，可以更全面地理解和提升它们作为智能代理的能力。




图表 2-34 AgentBench 环境覆盖和 Agent 评测情况（2023 年 8 月）

由于现有工具学习和智能体评测数据都是基于英文构建，2023 年 10 月 CLUE 中文语言理解评测基准团队，发布 AI 智能体中文测评基准 SuperCLUE-Agent[2-53]，聚焦于 Agent 能力的多维度基准测试，包括任务规划、长短期记忆、工具使用等 3 大核心能力、以及细分 10 大基础任务，可以用于评估大语言模型在核心 Agent 能力上的表现，包括工具使用、任务规划和长短期记忆能力。经过对 16 个支持中文的大语言模型的测评，发现在 Agent 的核心基础能力中文任务上，GPT4 模型大幅领先；同时，代表性国内模型，包括开源和闭源模型，已经较为接近 GPT3.5 水平。




图表 2-35 SuperGLUE 评测维度和大模型 Agent 能力评测（2023 年 10 月）

### 2.3.3.6 AI Agent 安全治理

领域应用数据通常是企业核心数据，可能会造成企业秘密和用户信息泄露。在系统平台上运行 Agent 系统，操作不当可能会造成系统文件删除。随着 AI Agent 在各行各业的应用日益广泛，确保其安全性和可靠性变得尤为重要。这包括保护系统免受外部攻击、防止数据泄

露、避免“幻觉”问题、异常行为控制、确保 AI 决策的透明性和可解释性，以及在多变环境中的稳定运行。安全管理不仅关系到技术的稳定性和可靠性，也关系用户信任和技术的广泛接受度。

提升 Agent 系统的安全性的方法包括：

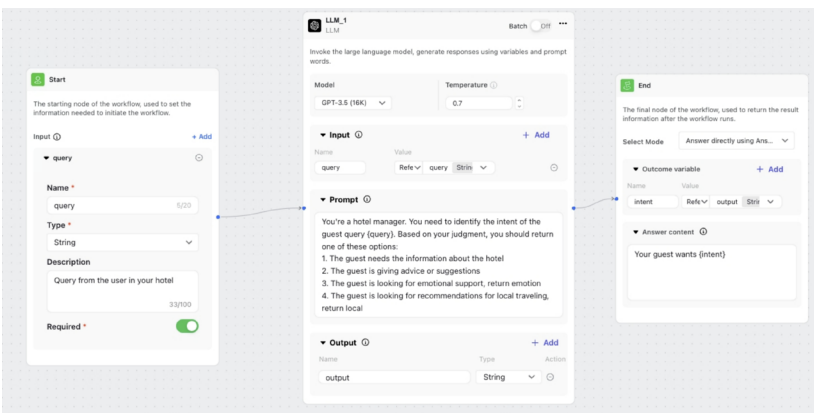
- **系统标准化建设：**目前 AI Agent 发展迅速，不同研究机构和公司推出自己的大模型、Agent、工具链等，缺乏统一的接口标准和通信协议。通过制定统一的交互协议、数据格式、接口标准、性能基准等，可以促进不同类型 AI 智能体之间的互操作性，建立严格的数据安全和隐私保护标准。

- **系统幻觉治理：**由于大模型本身的幻觉问题，容易导致 AI Agent 系统中在生成内容和交互中出现幻觉问题，比如在软件开发智能体系统中，幻觉问题可能包括错误的依赖引入、未实现的接口、未处理的异常等与任务需求不符的潜在程序缺陷。处理该幻觉问题方法包括：1）在系统中构建审查智能体自主地对大模型生成内容进行理解和提议，将相关意见传达给大模或系统用户，以协助其纠正生成内容中的幻觉问题；2）提升大模型的长短期记忆能力，通过上下文背景提示增加大模型生成内容的准确性；3）自我反思机制，根据执行的环境异常反馈结果评估和反思生成的内容，重新调整输出；4）工具调用弥补大模型本身缺乏的技能。

- **系统安全管理机制：**1）数据加密：为了保护数据在存储和传输过程中的安全，可采用先进的数据加密鉴权技术，确保数据内容在未经授权的情况下无法被读取或篡改，保障数据的机密性和完整性。2）权限分配：通过对接口精确的权限控制机制，确保只有经过授权的智

能体才能访问特定的数据和资源。这包括对智能体身份的验证和授权，以及对不同级别智能体访问权限的细致划分，从而减少数据泄露或误用的风险。3）执行监控：系统的运行过程中实时监控各项操作，及时发现异常行为或潜在的安全威胁，并采取措施进行阻断或报警，从而维护系统的稳定和安全。4）可解释性工具：可以帮助人工智能开发人员理解复杂模型内的决策过程。通过对执行过程的可视化展示，让系统管理员能够直观地监控和管理整个系统的运行状态，不仅可以提高问题检测和响应的效率，也可以使系统管理更加透明和可控。

• **预定义流程化执行**：自定义工作流 Workflow，比如对中间任务需要严格执行的过程通过预定义工作流，提升中间内容的精准性和可控性，目前 COZE、灵境矩阵等智能体生产平台基本提供了 Workflow 构建功能，这里 Workflow 涉及数据处理和传输、逻辑控制、权限设置等，通过节点和连线构建，节点类型包括 LLM 模型、代码执行、知识库、条件判断等，提供输入定义、功能编辑、输出定义等功能。




图表 2-36 CODE 的 Workflow 示例 ( 来源：CODE 产品说明文档 )

**系统安全对齐优化：**价值对齐是让系统与人类价值观对齐，让 AI Agent 系统遵循人类目标执行。实现系统价值对齐的主要方法：逆强化学习和行为克隆。逆强化学习是一个涉及根据观察到的行为学习代理奖励函数的过程。通过观察人类生成的任务演示，人工智能可以推断出最有可能导致该特定行为的奖励函数。这种方法可以帮助人工智能系统与人类价值观保持一致，让它们从一组有限的演示中进行概括，并执行与推断奖励一致的任务，即使是在新情况下或难以提供负面演示时也是如此。行为克隆（Behavior Clone，BC）涉及复制或模仿人类行为来完成特定任务。在 BC，人工智能代理在人类产生的输入和相应动作的数据集上进行训练，有效地学习预测人类在类似情况下会采取的动作。价值学习是一种人工智能对齐技术，专注于直接从人类反馈中学习人工智能代理的偏好。这种方法涉及人类注释数据，例如对不同的操作进行排名或估计各种结果的可取性。然后，人工智能代理使用这些信息来学习一个价值函数，该函数可以预测不同的可能事件的可取性，从而可以用来指导决策。价值学习可以通过利用人类的直接反馈来改善人工智能目标，从而使人工智能系统与人类价值观保持一致。然而，设计合适的反馈收集机制可能具有挑战性，如果人类注释者理解有限或未能就某些偏好达成一致，则可能会引入偏见。对抗性训练提升 AI Agent 系统的稳健性。对抗性训练基于持续改进人工智能安全的理念，对抗性训练已成为增强人工智能系统鲁棒性的关键技术之一。这种方法将人工智能模型暴露给一系列对抗性示例，这些示例是精心创建的输入，旨在利用模型的弱点。对抗性训练的目的是提高人工智能系统对意外或恶意输入的恢复能力。通过将这些对抗性示

例纳入训练过程，模型可以学会准确识别和解释这些困难的情况，最终在面对现实世界中的分布变化或对抗性示例时表现更好。

#### 2.3.3.7 AI Agent 构建基础平台

AI Agent 构建基础平台是一种标准化的技术平台，专门用于创建和管理 AI Agents。现有 AI Agent 创建平台具有以下特点：1）零代码低门槛构建，支持普通用户和开发者通过自然语言交互快速构建 Agents；2）集成工具链，如包含任务分解、思维链推理、向量检索、Memory 管理的 LangChain、向量检索数据库、本地知识库、插件调用、数据安全管理、智能体性能评测等基础开发工具；3）标准化接口，支持 Function Call、JSON 等标准化操作。AI Agent 构建基础平台核心组成主要包括 Profile、知识库、提示优化、工作流（Workflow）、工具使用等。

2023 年 11 月 OpenAI 开发者大会发布了 Agent 构建平台 GPTs，提供了普通用户自定义 GPT，根据特定任务创建定制版本的 ChatGPT（个性化 Agent），如管理数据库、电子邮件、发短信等等。发布 GPT Assistant API，帮助开发者在自己的应用程序中构建 Agent，并实现代码解释器、知识库检索、函数调用等功能。并推出 GPT Store，用户可以在该商店上分享构建的 Agent 应用。2 个月后推出 GPT 商店，并已经有超过 300 万个自定义版本的 Agent。微软在 Ignite2023 技术大会推出 Copilot 全家桶 Microsoft Copilot Studio，旨在增强和个性化 Microsoft Copilot 的体验，是一个低代码

工具，通过集成关键业务数据来定制面向 Microsoft 365 的 Copilot，并构建供企业内部或外部使用的定制 Copilot，可与连接器、插件和 GPT 配合使用，允许 IT 团队将 Copilot 引导至用于特定查询的最佳数据源。字节跳动的新加坡海外公司 SPRING 发布 COZE 平台，采用 GPT-4 大模型构建，通过提示、函数变量设置、插件、工作流、数据集等设置，同时提供基于预览的调试 Debug。美国知名实时组织和写作工具 Taskade，推出 AI Agent 构建平台，目前包含知识库、插件、指令等常规功能。

在国内，阿里、昆仑万维、澜码科技、百度、智谱 AI 等也推出了 Agent 创建平台：

- 阿里在大模型技术开源社区 ModelScope 上发布对标 GPTs 平台 AgentFabric，以通义千问大模型为基础模型，支持 Wanx 图片生成、代码编译器、高德天气、艺术字纹理生成等插件，用户可以在社区发布创建的 Agent。

- 昆仑万维发布 SkyAgents 平台，开启预约内测，该平台支持工具组件、对话式交互、工作流自动化、零代码定制、多智能体协作等功能。

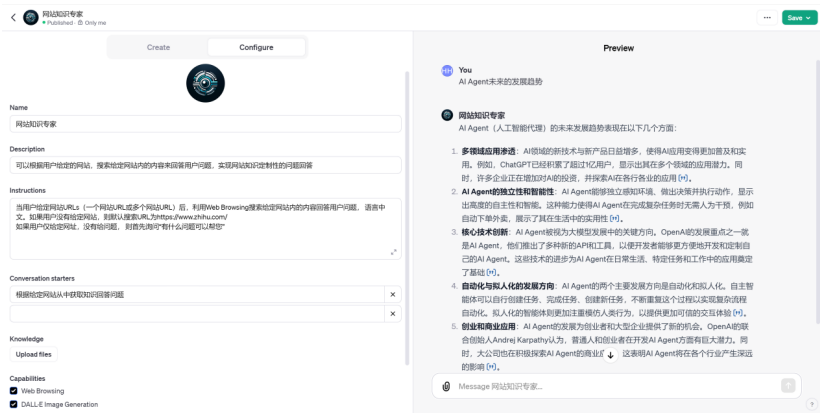
- 澜码科技打造企业级 Agent 平台 AskXBot，平台集 Agent 与工作流设计、开发、使用、管理，与知识沉淀于一体。在 AskXBOT 平台上，企业用户可以用对话的方式提出需求，设计、创建和管理

Agent，快速定制企业级 AI Agent 来完成各类任务，提升工作质量的同时降低成本。提供多样化的 Agent 模板，可快速定制符合企业特色的 AI Agent。结合 API、RPA 等技术，与企业业务系统深度融合与高效互动，推动业务的增强自动化。具备强大的文件处理能力，高效处理多种类型文档，解决企业在文件解析与处理方面的复杂需求。沉淀专家的知识和行业经验，构建企业知识库，促进知识共享与传承。配备仿真测试平台，全面评估 AI Agent 的表现，保障 Agent 效果稳定。提供全面的统计日志与审计日志记录，确保 AI Agent 使用的高安全性和可追溯性。实时记录用户与 Agent 的互动，助力运营团队持续优化与迭代。

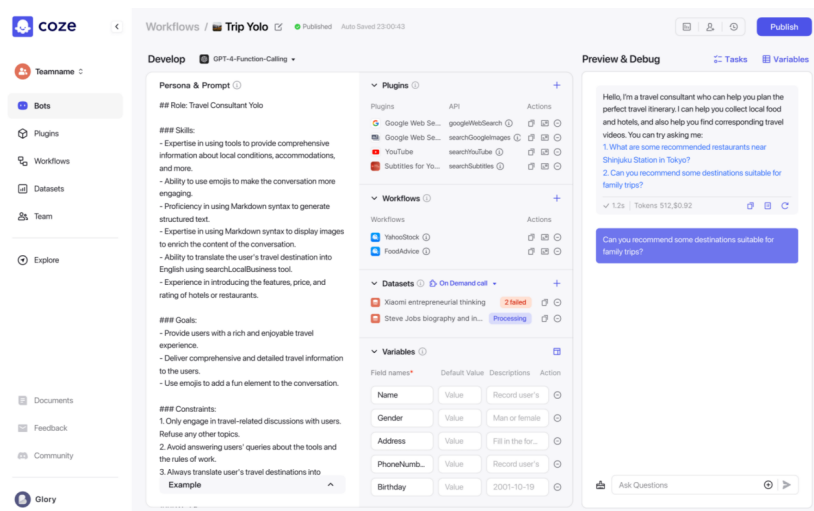
- 百度推出基于文心大模型的平台“灵境矩阵”，旨在支持各类开发者利用大模型时代的技术，创造强大的产品能力。这个平台提供了多样化的智能体（Agent）开发方式，包括零代码、低代码和全代码三种解决方案，适应不同技术水平的开发者。灵境矩阵还提供 AI 插件，将大模型的 AI 能力与外部应用结合，从而丰富功能和应用场景。此外，平台通过传统和 AI 搜索引擎、文心一言 App 等多种渠道，触达和服务于广大用户。

- 智谱 AI 在 2024 年 1 月第一届开发者大会上，发布对标 OpenAI GPTs 的 GLMs 个性化智能体平台，支持用户用简单的提示词创建属于自己的智能体。该智能体平台支持内置的联网、AI 绘画、代码能力的工具使用、RAG、prompt 优化等能力。与此同时，智谱 AI 还对标 OpenAI 的 GPT Store，上线了智能体中心，可以让用户分享

# 自己创建的智能体 GLM 模型




图表 2-37 OpenAI 的 GPTs 创建平台界面




图表 2-38 COZE 的创建平台界面

| 名称 | 发布时间 | 大模型 | 工具使用 | 行为设置 | 智能体类型 | 知识库 | 调试功能 | LOGO生成 | 工作流 | 团队协作 |
|---|---|---|---|---|---|---|---|---|---|---|
| OpenAI GPTs | 2023.11 | GPT-4 | ✓ | ✓ | 单智能体 | ✓ | ✓ | ✓ | × | × |
| Microsoft Copilot Studio | 2023.11 | GPT-4 | ✓ | ✓ | 单智能体 | ✓ | ✓ | × | ✓ | ✓ |
| COZE | 2023.11 | GPT-4 | ✓ | ✓ | 单智能体 | ✓ | ✓ | × | ✓ | ✓ |
| AI Agents | 2023.11 | GPT-4 | ✓ | × | 单智能体 | ✓ | ✓ | × | × | × |
| AgentFabric | 2023.12 | 通义千问 | ✓ | ✓ | 单智能体 | ✓ | ✓ | × | × | × |
| SkyAgents | 2023.12 | 天工 Skywork | × | × | 单智能体 | ✓ | ✓ | × | ✓ | × |
| 灵境矩阵 | 2023.12 | 文心一言 | ✓ | × | 单智能体 | ✓ | ✓ | × | ✓ | × |
| GLMs | 2024.1 | GLM-4 | ✓ | × | 单智能体 | ✓ | ✓ | × | × | × |

图表 2-39　国内外代表性 AI Agent 构建基础平台

## 2.3.4 代表性 AI Agent

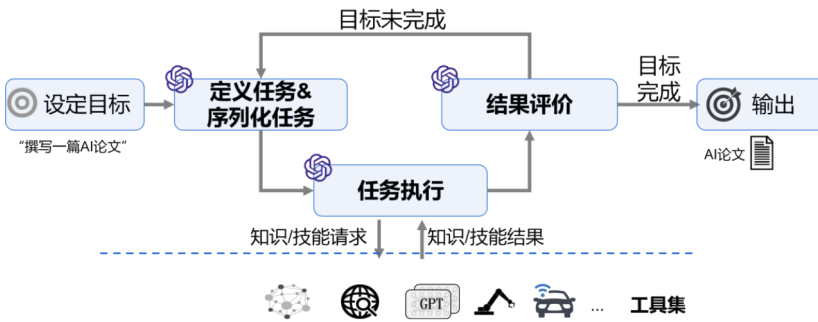
一些具有代表性的单智能体系统：

### （1）AutoGPT

AutoGPT 是一个基于 ChatGPT 的开源 AI 项目。它的独特之处在于，它可以使 GPT 模型自主行动，无需人为指令。它遵循单一智能体范式，属于任务导向型，它根据终极目标制定多个子目标的来自动化 NLP 任务。AutoGPT 最初被称为 EntrepreneurGPT，由开发者 Significant Gravitas 于 2023 年 3 月创建。该项目的主要目的是测试 GPT-4 在商业领域的可行性，特别是它在做出商业决策方面的能力。AutoGPT 的主要特点包括自主功能、面向目标的功能、多用途功能和用户友好的界面。它能够重写自己的代码，并能自行搜索互联网，执行任务如保存文件到计算机等。此外，AutoGPT 集成了 ElevenLabs 的长 / 短期记忆和文本到语音功能，能够生成类似人类的文本、回答

问题、翻译语言、总结文本以及提供建议。

自治 AI 机制是一个高度复杂的过程，它使 AutoGPT 和 AgentGPT 等 AI 系统能够有效地实现用户定义的目标。这个过程涉及四个关键步骤，它们共同确保 AI 的行动组织良好且有效。

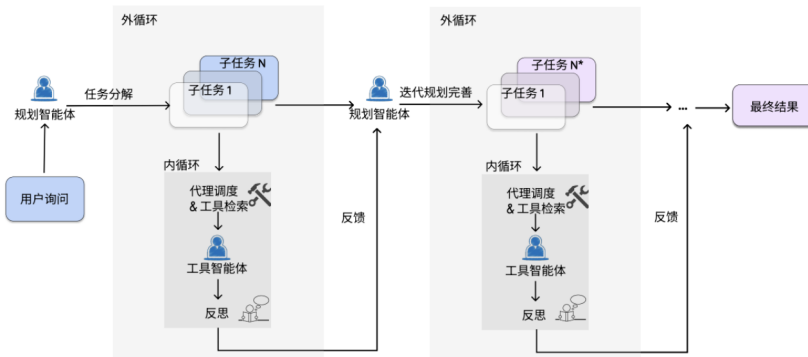



图表 2-40 自主智能体 AutoGPT 技术框架

### (2) XAgent

XAgent[2-38] 是由清华大学自然语言处理实验室提出的大语言模型驱动的自主智能体，能够独立地解决复杂任务。XAgent 的提出背景源于现有人工智能智能体（AI Agent）在解决复杂问题和任务中的局限性。这些局限性主要体现在以下几个方面：传统的 AI 智能体通常基于人类设定的特定规则运作，这限制了它们在处理未知或复杂问题时的能力。它们更多地被视为工具，而不是具备独立决策能力的智能实体；现有的自主智能体如 AutoGPT，虽然在一定程度上突破了传统模

型的局限，但在执行复杂任务时仍会遇到诸如死循环、错误调用等问题，需要人工干预；传统智能体在与人类互动时存在局限，不能有效地结合人类直觉和专业知识，这限制了它们在实际应用中的有效性。为此，XAgent 被提出通过自主规划和决策能力，使智能体能够独立运行，发现新策略和解决方案，不受人类预设的束缚。




图表 2-41 XAgent 的“双循环机制”

### (3)ProAgent

随着信息时代的到来，软件作为信息处理、存储和通信的基础成为了人类生产生活密不可分的一环，从而催成了机器人流程自动化（Robotic Process Automation, RPA）技术。其通过人工编制规则将多个软件协调成一个固化的工作流（Workflow），通过模拟人交互的方式来和软件交互实现高效执行。传统 RPA 存在两个局限性：编写 RPA 工作流本身需要繁重的人类劳动，成本较高；复杂任务非常灵活，通常涉及动态决策，难以固化为规则进行表示。

为此，清华大学 & 面壁智能将 AI Agent 技术引入到 RPA 中，提出 AI Agent 自动化技术（Agentic Process Automation,APA）[2-40]，结合 AI Agent 技术（LLM-based Agents），自动化工作流构建并处理其中的复杂决策和动态处理环节。通过 APA，智能体能够根据人类需求自主完成工作流的构建，识别并自动编排需要动态决策的部分，并在执行过程中主动接管复杂决策的环节。构建一个能接收人类指令并以生成代码的方式构建工作流的智能体 ProAgent，融合 DataAgent 和 ControlAgent，处理复杂的数据和逻辑控制任务。

围绕智能体流程自动化（APA）的构建与应用展开，设计基于大语言模型的 APA 智能体 ProAgent，利用 Agentic Workflow Description Language（基于 JSON 和 Python）描述工作流。工作流构建转化为代码生成任务，智能体接收人类指令后生成代码，自动构建工作流。其中，DataAgent 和 ControlAgent 分别处理复杂数据处理和逻辑控制任务。工作流的构建遵循 ReACT 模式，包括定义工具使用、工具调用转化为函数、定义 mainWorkflow 函数组织整体逻辑控制与数据处理，以及提交任务表示工作流构建结束。优化策略包括构建中测试、函数调用封装和编写计划注释以提升性能。工作流执行基于 Python 解释器，按顺序逐行执行。通过 ProAgent 实验，验证了 APA 的可行性和 AI Agent 在自动化中的应用潜力。此外，针对自适应跨平台兼容技术，采用 API 和环境自适应，指令、工具、API 泛化，Docker 容器等多层级的跨平台兼容适配技术。

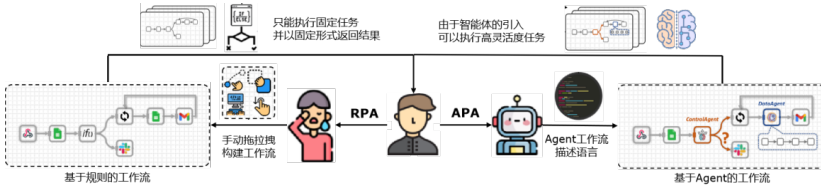
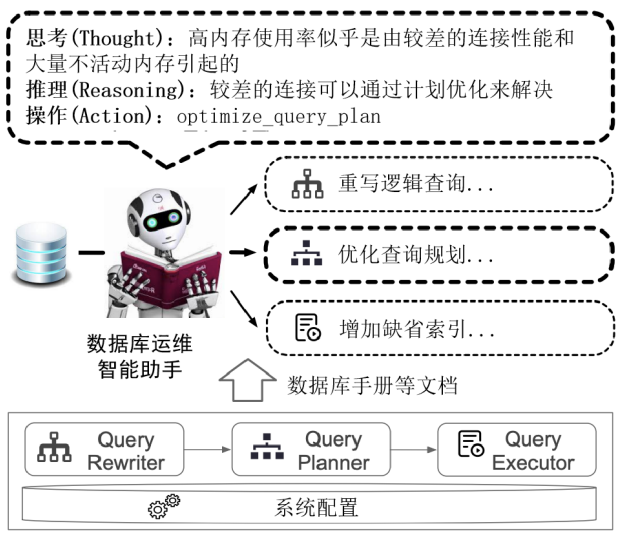



图 2-42 智能体流程自动化的智能体 ProAgent 的工作流构建过程

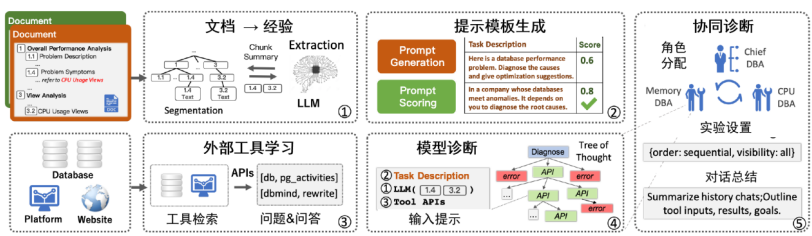
### (4)D-Bot

D-Bot[4-46] 是由清华大学开发的智能体交互式数据库运维技术框架 D-Bot。针对系统运维文档多、云上运维压力大、复杂问题诊断难等挑战，利用大模型训练“智能运维助手”D-Bot，构建了数据库管理员（DBA）主管、资源异常专家、查询优化专家的多角色智能体，通过语言交互学习人类运维经验、诊断异常根因。




图表 2-43 基于大语言模型的数据库运维智能助手

D-Bot 整体方案：一个基于 LLM 的数据库管理员。首先，D-Bot 通过将文档分割成可管理的块并对这些块进行总结，将文档转换为经验知识。其次，通过迭代生成和评估不同格式的任务描述，以帮助 LLM 更好地理解任务。第三，D-Bot 通过使用匹配算法选择适当的工具并为 LLM 提供如何使用所选工具的 API 的说明来利用外部工具。一旦具备了经验、工具和输入提示，LLM 就可以检测异常、分析根因，并提出优化建议，遵循思维树策略，在发生故障时回到先前的步骤。此外，D-Bot 通过允许多个 LLM 基于预定义的环境设置进行通信，以启发更强健的解决方案，促进协作诊断。D-Bot 在 11 类测试场景中的诊断正确率达 81.8%，远高于 GPT4（36.4%），在典型问题上的诊断水平接近人类 DBA。




图表 2-44 智能体交互式数据库运维助手 D-Bot 的技术框架

## 2.4 群体智能

随着 AI Agent 数量的增加和智能体间的协作能力提升，能够呈现出超越单个智能体能力的集体智慧，实现对更加复杂任务处理和场景建模。本节将对 AI Agent 群体智能概念和发展情况、系统框架、关键技术等进行介绍。

## 2.4.1 群体智能的发展演进

### 2.4.1.1 群体智能的基本概念

传统群体智能（Swarm Intelligence）指一种自然界和人类社会普遍存在的现象，指的是在没有中央控制的情况下，通过个体间的简单交互，产生出复杂、智能的群体行为。这种智能形式可以在动物群体（如蚂蚁、蜜蜂、鱼群等）和人类社会（如市场经济、在线合作等）中观察到。

大语言模型驱动的群体智能（Collective Intelligence），由多个 AI Agent 组成，通过多个智能体语言交互和协作涌现的集体智能。这种智能虽然受到人为控制和设计，但其核心特点与传统的群体智能在多方面有着相似之处以及特点：

- **分布式或去中心化的控制：**在这种群体智能系统中，每个大语言模型可以是由中心化的机构设计和训练，同时也可以没有单一的控制中心指导所有模型的响应；每个模型实例都基于其接收到的特定输入独立作出反应。

- **自组织：**大语言模型能够根据输入自动调整其响应，展现出一定程度的自我组织能力。这表现在模型能够理解和适应不同的查询类型和用户风格。

- **可扩展性和弹性：**随着更多的数据和例子被输入，这些模型能够不断学习和适应，展现出良好的可扩展性。同时，单个模型实例的失败不会影响整个系统的运行。

- **智能涌现：**尽管每个模型实例独立运行，但整个系统（包括所有模型实例和用户交互）作为一个整体表现出超越单个实例的智能。通过不断的交互和学习，系统能够提供更加精准和丰富的回应。

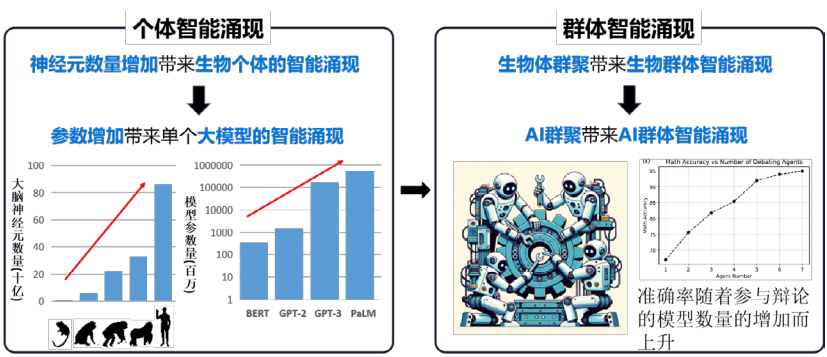
大语言模型驱动的群体智能这里简称为“大模型群体智能”或“大模型多智能体技术”。大模型群体智能与传统群体智能的有着明显区别，其差异如图表 2-45 所示。

| | 传统群体智能 | 大模型群体智能 |
|---|---|---|
| 沟通方式 | 个体之间的沟通往往是基于非言语的信号。这种沟通更多是基于行为上的简单规则和环境中的触发因素。 | 使用复杂的自然语言进行交流，能够处理和生成语言信息，进行更加复杂和深入的交流。这种方式可以支持抽象思维、复杂问题解决和丰富的信息交换。 |
| 决策过程 | 决策通常是分散的，基于局部信息和简单的个体间交互。群体行为的出现更多是自发和自组织的结果。 | 智能体能够进行更为复杂的推理和决策过程，基于对语言信息的深入理解和分析，可以在决策中考虑更广泛和深入的因素。 |
| 信息处理能力 | 信息处理能力相对有限，主要依赖于直觉式的、基于规则的响应。 | 可以处理大量的信息，进行复杂的语言理解和生成，支持更复杂的信息处理和知识推理。 |
| 应用范围 | 主要应用于模拟自然界中的群体行为，如优化算法、机器人协作等。 | 应用范围更广，包括数据分析、自然语言处理、人机交互、复杂问题解决等多个领域。 |

图表 2-45 传统群体智能与大模型群体智能的对比

### 2.4.1.2 群体智能系统的优越性

AI 群聚带来 AI 群体智能涌现。当前大量对智能体研究和实践显示，多个智能体交互协作可以处理更加复杂的任务，展示出超越单个智能体的能力，被认为是既大模型单体智能涌现之后的“智能的第二次涌现”，即群体智能。例如，在解决数学问题和软件开发等领域，大模型多智能体系统已经显示了其卓越的解决问题能力。在这些情境中，智能体可以彼此交流、分享信息和策略，甚至进行“辩论”，通过这种方式，它们能找到比单个智能体更优、更有效的解决方案。这不仅体现了 AI 群体智能的高效协同工作能力，也揭示了在人工智能发展中，从单一大模型到多智能体群体智能的重要转变。这种转变为 AI 的未来应用打开了新的可能性，预示着更加智能和自适应的技术解决方案的出现。




图表 2-46 个体智能涌现到群体智能涌现

单智能体的核心在于 LLM 与感知、行动的联动。LLM 通过理解用

户的任务，推理出需要调用的工具或行动，并基于调用或行动结果给用户反馈。但是对于大量复杂场景来说，单智能体能力仍然有限。以写稿为例，完成的操作流程中该场景至少需要 4 个智能体：

- 调研人员： 根据需求，搜各种资料并进行总结；
- 编辑人员： 根据需求和调研人员提供的资料，给出稿件的方向和框架；
- 创作人员： 根据编辑人员的要求，完成稿件撰写；
- 评审人员： 审稿，提出修改意见，返回给创作人员做改进。

上面这 4 个智能体，既是 4 个角色，也是写稿工作流中的 4 个步骤。显然，单个智能体无法胜任。而且，类似评审人员这样的角色还能让 LLM 自我审核。多角色、复杂流程，需要多个 AI Agent 协作才能胜任，因此需要构建多智能体系统。多智能体系统会为不同的 AI Agent 赋予不同的角色定位， 通过 Agent 之间的协作来完成复杂的任务。

#### 2.4.1.3 大模型群体智能的类型

群体智能，根据设计目标不同，可以划分为社会模拟型与任务完成型两类形式：

- **社会模拟型：**代表性如斯坦福大学提出的西部小镇 SmallVille，基于层次规划的智能体社会小镇，实现人类社群行为的可信模拟；

• **任务完成型：**代表性的如面壁智能 & 清华大学团队共同提出的 ChatDev，基于语言交互的智能体软件开发，实现群体交互协作式任务完成。

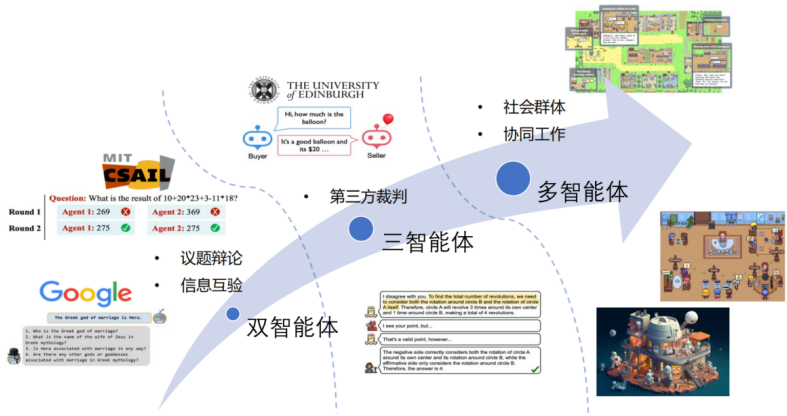


**社会模拟型：西部小镇 Smallville　任务完成型：软件开发 ChatDev**

图表 2-47 群体智能的两种类型

### 2.4.1.4 发展趋势与挑战

当前，AI Agent 群体智能的发展正逐渐成为研究热点，被认为是迈向通用人工智能的重要途径。群体智能主要由多智能体间的方案提议、决策研讨、分工执行等功能组成。在协同编程、自动化、社会模拟、数据库运维等场景验证了应用效果。MIT 和 Google 等研究发现通过智能体之间辩论和信息互验可以显著提升数学推理的能力，改善生成内容的事实准确性，而且该能力随着参与辩论的智能体数目增加而持续增加。随后，爱丁堡大学在此大模型相互辩论基础上，引入第三方裁判，向智能体提供反馈，在销售价格谈判进行实验表明这种策略有助于提升智能体自主地相互改进。斯坦福大学、清华大学等构建更大规模的智能体协作，实现了社会模拟和软件开发等更加复杂的能力，如图表 2-48 展示了群体智能的形式和智能趋势。




图表 2-48 群体智能形式及其智能趋势

随着技术的进步，预计未来将看到以下趋势：1）更强的协作能力：未来的多智能体系统将更加擅长在没有人类直接干预的情况下相互协作，处理复杂问题；2）分布式智能处理：群体智能不仅局限于单一系统或位置，而是跨多个网络、设备分布，提高了灵活性和效率；3）自我学习与适应：随着深度学习的发展，AI Agents 将能够更好地从环境中学习并适应新挑战；4）多领域应用扩展：群体智能的应用范围将不断扩大，涵盖医疗、交通、环境监测等多个领域；5）人机协同进化：多智能体系统将与人类更紧密地协同工作，推动共同的进步和创新。整体而言，目前 AI 群体智能技术发展尚处于早期阶段，大量实现路径仍待探索，需要探索者持续创新。

多智能体系统本质上是复杂的实体。它们涉及多个自主智能体的交互，每个智能体都有自己的能力和目标。这种复杂性虽然是系统力量的源泉，但也带来了许多挑战：

- **动态系统的挑战：**动态添加智能体在提供增强灵活性和适应性的同时，也带来了一些挑战。主要问题之一是智能体过度扩散的风险，这可能导致资源耗尽或系统效率低下。为了减轻这种风险，系统需要纳入监视和控制新智能体创建的机制。具体来说，系统需要采用资源管理模块来跟踪每个智能体和整个系统消耗的计算资源。当资源使用量接近预定义阈值时，该模块可以向系统发出警报，从而触发措施以防止资源耗尽。这些措施可能包括停止创建新的智能体。除了资源管理之外，系统还需要确保智能体的动态添加不会导致效率低下或冲突。这是通过协调机制来实现的，该机制监督智能体的角色和任务分配。当创建新智能体时，该机制可确保其角色和任务不会与现有智能体显着重叠，从而防止冗余和潜在冲突。

- **可扩展性：**是多智能体系统中的另一个重大挑战。随着系统规模和复杂性的增长，维护系统的性能和效率可能变得越来越困难。管理大量智能体的交互和操作所需的计算资源可能是巨大的。此外，随着智能体数量的增加，冲突和不一致的可能性也会增加，这可能会进一步影响系统的性能。

- **系统评估：**由于系统可以处理的任务的复杂性和多样性，评估多智能体系统的性能可能具有挑战性。传统的评估指标可能不足以或不适合评估系统的性能。因此，可能需要开发新的评估指标和方法来准确衡量系统及其各个智能体的性能。

- **伦理考虑：**多主体系统的使用也引起了一些伦理考虑。例如，

系统可能做出对个人或社会产生重大影响的决策或采取行动。因此，确保系统以道德方式运行并尊重所有用户的权益至关重要。这需要仔细设计和监督系统，以及实施适当的道德准则和保障措施。

- **高效协作：**目前智能体呈现从单体智能往群体智能转变趋势。在多 Agent 系统中，如何高效协调和协作仍是一个挑战，相关理论和应用研究还非常初步。

### 2.4.2 群体智能技术框架

目前，多智能体系统已经被广泛研究，总体来看，多智能体框架的构建主要涉及如下关键问题：

- **单个个体智能体的职能和能力是什么？**即角色定义，需要考虑在应用场景中需要的个体智能体分工及能力。

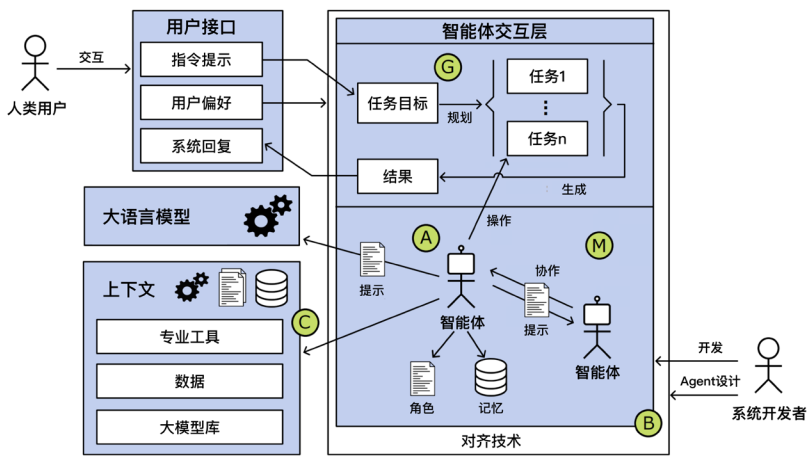
- **智能体如何感知应用场景的环境信息？**环境信息主要是指系统中定义的智能体、智能体间的可见性、以及其他的用户配置和全局变量。该问题需要我们能够清楚的定义个体智能体的感知能力半径。

- **个体智能体间如何交流协作？**需要考虑不同应用场景智能体间的通信方式，因为基于大模型的智能体以语言交流为主，因此其重点考虑不同智能体的协作形式（合作还是对抗）、交流方式（纯自然语言还是附加环境变量）和协作模式（有序还是无序 / 动态还是静态）。

- **系统是否允许人类在执行过程中参与以及如何参与？**在多智能

体系统中，人类参与的角色和方式取决于系统的设计和目标。通常，人类可以作为监督者、决策者或合作者参与。作为监督者，人类监控系统的性能和安全，确保智能体按预期工作；作为决策者，人类可能在关键时刻做出重大决策，指导智能体的行为；作为合作者，人类与智能体共同工作，共享信息和策略，实现协同效果。这种参与程度可以根据系统的自动化程度和应用场景的复杂性而变化。

基于大语言模型的多智能体系统技术框架 [2-49] 如图表 2-49 所示，主要包括以下几个关键特性：




图表 2-49 基于大语言模型的多智能体技术框架 [2-49]

- **目标驱动的任务管理（G）**：这类系统旨在完成用户提示的目标或复杂任务。系统通过交互式和多视角策略，将复杂任务分解为更小、更易管理的任务，并在各个具有特定能力的智能体之间分配这些任务。关键在于有效的任务协调和部分结果的综合。

- **LLM 驱动的智能体（A）：**智能体作为系统的基本组成部分，每个智能体都具有独特的能力集合，包括明确的角色和个人记忆。它们的推理和解释能力的核心是大型语言模型的整合，使智能体不仅能够反思任务，规划和高效处理分配的任务，还能访问和利用上下文资源，并与其他智能体通信。

- **多智能体协作（M）：**交互层为 LLM 驱动的智能体网络提供工作空间。在执行分配的任务时，这些智能体通过基于提示的消息交换进行协作，以委派责任、寻求帮助或评估任务结果。智能体协作的关键是有效结合每个智能体的优势（认知协同）。

- **上下文交互（C）：**一些任务需要利用上下文资源，如专家工具、数据、更专业的大模型或其他应用程序。这些资源扩展了智能体收集环境信息、创建或修改工件或启动外部流程的能力，从而有效执行复杂任务。

- **平衡自主性与一致性（B）：**LLM 驱动的多智能体系统的动态特性体现在自主性和一致性之间的复杂相互作用中。这种复杂性源于人类用户、LLM 驱动的智能体以及系统中集成的治理机制或规则之间的三重互动和内在张力。一致性确保系统的行为与人类意图和价值观保持同步，而自主性则表示智能体自我组织策略和操作的内在能力，使其能够独立于预定义的规则和机制并在没有人类监督的情况下运行。

### 2.4.3 代表性群体智能开发平台

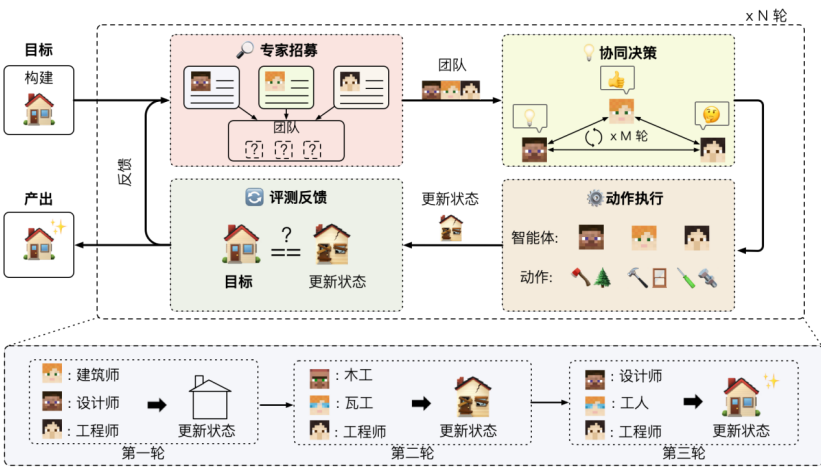
### （1）AgentVerse

AgentVerse[2-39] 是由清华大学自然语言处理实验室 & 面壁智能联合开发的灵活易用的高可扩展群体智能平台，支持利用基础模型定制多智能体环境，创建多个具有不同能力与身份的智能体。

受人类群体动力学启发的大模型多智能体协同技术，包括灵活代码扩展及定制化功能设计框架、智能体语言交互协同合作机制、智能体系统功能与结构演化机制等。对该平台分为四个阶段：专家招募阶段，根据问题解决的进展情况确定和调整座席人员组成。协作决策阶段，选定的智能体进行联合讨论以制定解决问题的策略。行动执行阶段，智能体与环境交互以实施决策阶段计划的行动。评估和反馈阶段，对当前状态与期望结果之间的差异进行评估， 如果当前状态不理想，则给出反馈，以便在下一次迭代中进一步细化。在该技术框架技术上，开发简单可定制可扩展的多智能体协同技术工具。

通过专家招募、协作决策、行动执行、评估和反馈等四个阶段，从专家招聘开始，团队的组成是根据当前问题的具体需求量身定制的。接下来是协作决策，智能体们共同制定策略。接下来，在行动执行阶段，这些策略被付诸行动。最后，评估阶段评估这些行动的有效性，并为未来的改进提供反馈。这个循环过程确保了策略的不断完善和调整，以实现最佳结果。


图表 2-50 灵活易用的高可扩展群体智能平台 AgentVerse



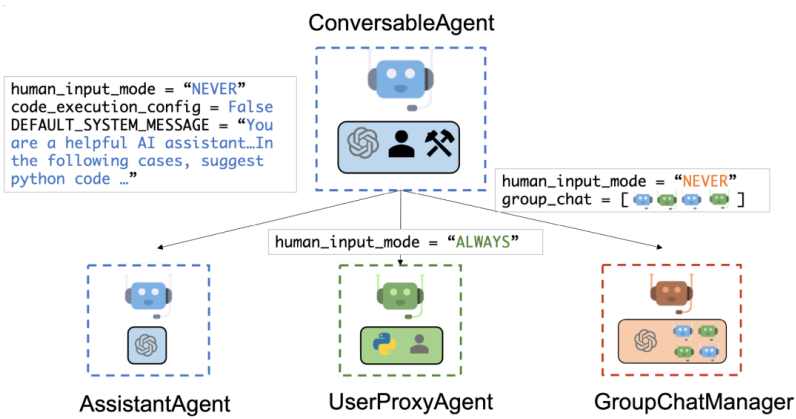

图表 2-51 AgentVerse 的技术框架

### （2）AutoGen

AutoGen[2-51] 是微软开源的多智能体对话框架，专为基于 LLM 的应用设计，强调灵活性和通用性。这个框架通过简化、优化和自动化 LLM 的工作流程，实现了多智能体之间基于自然语言的互动，以完

成复杂任务。核心概念是创造可以进行对话的智能体，它们可以自主执行任务或与人类协作。

AutoGen 突出了智能体的两大特征：可对话性和可定制化。智能体能够发送和接收消息，进行对话，实现信息交换和任务协同。同时，这些智能体可根据需求定制，整合 LLM、人类输入、工具或其组合。此外，AutoGen 引入了对话式编程的概念，这是一种以智能体间对话为核心的编程范式。开发者通过定义一组具有特定能力和角色的智能体，并编写它们之间的交互行为来构建应用程序，简化了复杂应用的开发过程。




图表 2-52 AutoGen 内置多智能体 [2-51]

图表 2-52 展示了 AutoGen 中内置的智能体。在 AutoGen 框架中，核心类是 ConversableAgent，它允许智能体相互通信和执行操作。该框架定义了三个典型的智能体子类：AssistantAgent、

UserProxyAgent 和 GroupChatManager。AssistantAgent 作为 AI 助手，主要使用 LLM（如 GPT-4）生成 Python 代码，并处理代码执行结果及其修正。UserProxyAgent 代表人类用户，能够在交互中引入人类输入，执行代码和调用功能。GroupChatManager 则负责管理对话流程，支持动态群聊功能，选择发言者并向其他智能体广播响应。ConversableAgent 的自动回复功能支持智能体间的自治通信，同时保留了人工干预的可能性。

### （3）CAMEL

CAMEL[2-43] 是最早基于 ChatGPT 的多智能体交互框架，由沙特阿拉伯阿卜杜拉国王科技大学在 2023 年 3 月提出。CAMEL 重点探索了一种称为角色扮演的新型合作代理框架，该框架可以有效缓解智能体对话过程中出现的错误现象，从而有效引导智能体完成各种复杂的任务，人类用户只需要输入一个初步的想法就可以启动整个过程。该框架设计了灵活的模块化功能，包括不同代理的实现、各种专业领域的提示示例和 AI 数据探索框架等，可以作为一个基础的 Agents 后端，支持 AI 研究者和开发者更加轻松地开发有关于多智能体系统、合作人工智能、博弈论模拟、社会分析、人工智能伦理等方面的应用。

CAMEL 内置的协作式角色扮演框架可以在人类用户不具备专业知识的情况下，通过 Agents 之间的协作方式完成复杂任务。在协作角色扮演框架中，人类用户需要首先制定一个想要实现的想法或目标，例如：开发一个用于股票市场的交易机器人。这项任务涉及的角色是

AI 助理智能体（使其扮演 Python 程序员角色）和 AI 用户智能体（使其扮演股票交易员角色）。 设置了一个任务细化器，该细化器会根据输入的想法来制定一个较为详细的实现步骤，随后 AI 助理智能体（AI Assistant）和 AI 用户智能体通过聊天的方式来进行协作通信，各自一步步完成指定的任务。CAMEL 也提供了能够在物理世界中执行各种操作的具身智能体，它们可以浏览互联网、阅读文档、创建图像、音频和视频等内容，甚至可以直接执行代码。为了增强角色扮演框架的可控性，CAMEL 设计了一种 critic-in-the-loop，设置一个中间评价智能体来根据用户智能体和助理智能体出的各种观点进行决策来完成最终任务。

## 2.5 组织孪生

### 2.5.1 组织孪生的基本概念

组织孪生是一个以数字技术为核心的创新框架，它包括三个关键部分：岗位孪生、架构孪生和业务孪生。

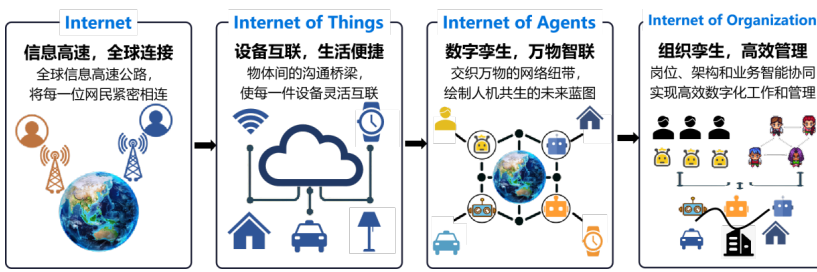
- **岗位孪生**利用大模型技术创建个人的数字孪生虚拟人，这些虚拟人能模拟真人的交流方式，包括声音和表情，并具备“感性智能”。它们能够执行内容生成、基础交流、客户服务等工作。

- **架构孪生**则是在数字世界中映射真实公司的组织架构，通过智能体网络技术定义智能体间的交流和逻辑。

• **业务孪生**通过整合大语言模型、搜索增强技术和智能体构建等，自动执行实际业务，优化业务执行效果。这个框架特别适用于复杂的行业场景，如汽车行业，提供了一个全新的数字化工作和管理方式。

## 2.5.2 组织孪生的演化历程

从互联网出现开始，信息交互、处理的能力得到了飞速的发展，Internet 把人关联了起来；随着物联网出现，Internet 将物品进行了关联；随着 Agent 技术的出现，群体智能技术支撑了 Agent 相互之间交流、协作；基于以上技术，针对专门组织的孪生技术提出了组织孪生的概念。




图表 2-53 组织孪生演化过程

（1）Internet：信息高速，全球连接

随着互联网的出现，信息交互和处理能力取得了飞速的发展。互联网将人们紧密关联在一起，极大地促进了信息的流通和共享。这一时代的特点是信息高速、全球连接。在这个阶段，人们能够迅速获取和传递信息，促进了跨地域的沟通和合作。

（2）Internet of Things：设备互联，生活便捷

随着物联网的兴起，互联网不仅将人连接起来，还将物品进行了关联。物联网实现了设备之间的互联，使得生活变得更加便捷。各种智能设备通过互联网实现了数据的共享和交流，为人们提供了更加智能、自动化的生活方式。

（3）Internet of Agents：数字孪生，万物智联

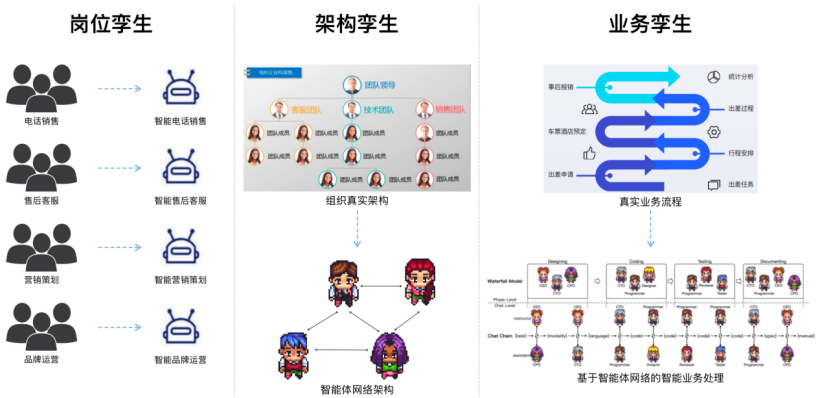
随着 AI Agent 技术的出现，群体智能技术支撑了 AI Agent 相互之间的交流和协作。数字孪生的概念催生了 Internet of Agents，使得智能体能够模拟人类的交流方式，实现更为复杂的任务和合作。在这一阶段，万物智联，数字化的代理体系使得信息处理和决策能力更加强大。

（4）Internet of Organization：组织孪生，高效管理

群体智能在企业深度应用，将实现企业组织的数字孪生，为企业组织管理和运营带来重要变革。Internet of Organization 阶段，不仅仅是智能体之间的连接，更注重整个组织体系的数字化和智能化。组织能够实现连胜，通过高效管理、优化决策，进一步推动了整个社会的发展。这一阶段，组织不再是孤立的单元，而是通过数字孪生技术实现了高度的协同与管理效能。

### 2.5.3 组织孪生的解决方案

下面从组织孪生三个方面分别介绍基于 AI Agent 技术的解决方案：




图表 2-54 AI Agent 技术驱动的组织孪生解决方案

### 2.5.1.1 岗位孪生

基于大模型技术，能够创建个人的数字孪生虚拟人，模拟真人的交流方式，甚至声音、表情等，基于强大的基座大模型能力，能够创建具备“感性智能”的智能员工，完成内容生成和创作、基础交流、客户服务、情感陪伴等偏通识能力的工作。

然而目前基座大模型在专业知识问答、涉及时效性问题、涉及相对复杂技能（例如数学计算）的任务上，表现并不如预期，有非常明显的“幻觉”问题，针对这些问题，仅仅提升大模型基础能力并不足以使得数字员工能够达到各个真实业务岗位的要求。通过强大的复杂

任务理解、规划、基于工具学习的框架技术，如 XAgent、AutoGPT 等，我们能够使得大模型具备像人一样思考、规划和使用工具的能力，创建出更加贴近岗位要求的孪生数字员工。

基于智能体技术的数字孪生虚拟人，为企业带来了前所未有的个性化定制和高度可定制性。这一创新性的应用得益于智能体系统专有的提示词框架，通过按照提升词框架来巧妙设计与岗位相关的提示词，并精准限定基座大模型回答问题的范围、方式等，我们能够充分挖掘出基座大模型的多方面能力。比如在汽车领域，我们可以通过 prompt 的方式让模型从汽车从业者的角度去进行任务执行和进行内容生成，从而达到定制化汽车员工数字人的效果。

尽管基座大模型是通用语言模型，其内置的知识是通用的，对于特定领域的问题可能无法给出准确的答案。为此，引入检索增强生成（RAG）技术，可以将特定领域的文档和问答灌入系统，形成“长期记忆”存储于向量数据库或搜索系统中。在生成过程中，将相关记忆注入到提词中，使数字孪生虚拟人能够精准回答特定领域的问题，从而弥补基座大模型的潜在不足。在汽车领域，我们可以让智能体调用 API 接口，并根据接口返回的行业知识，进行专业、可溯源的内容生成。

当提示词工程和知识库类的长期记忆补充依然不能完全满足业务需求时，我们能够采用高效后预训练和高效微调技术。通过微调和后预训练，我们能够“教给”大模型相关的垂直领域知识，为数字虚拟人赋予个性化，使其更好地适应不同的业务场景和用户需求。这种高

度可定制性为企业提供了灵活性，使其能够创造符合自身特色和品牌形象的智能员工。基于易慧智能对于汽车行业的深入理解通过微调的方式灌输给大模型，基于微调模型后的智能体就具备了深度的汽车行业知识。

通过对基座大模型和算力的本地化部署以及智能体技术的灵活性，数字孪生虚拟人可以不受时间和地点的限制，全天候的提供服务。无论是在工作时间内还是非工作时间，数字孪生虚拟人都能够为用户提供及时、高效的支持和服务，从而提升工作效率和用户体验。同时，数字孪生虚拟人具有卓越的多任务处理能力，能够同时执行多个任务而不影响效率。这种并行处理的优势使得数字虚拟人在处理复杂业务流程、同时与多用户互动等方面表现出色。

数字孪生虚拟人融合了智能体提示词框架、RAG 的长期记忆、高效后预训练和高效微调技术，使其具备了模拟人类语言和情感表达的卓越能力。通过自然语言处理和生成，可以实现数字孪生虚拟人与用户之间更为真实和生动的交流，为客户服务、沟通和情感陪伴等场景注入了全新的智慧。在汽车媒体等场景中，数字孪生虚拟人的逼真语言模拟和情感表达能力将成为提升服务质量和用户体验的得力助手。

数字孪生虚拟人凭借智能体技术，拥有持续学习和进化的独特能力。智能体的反思机制赋予虚拟人在任务执行后进行深度反思的能力，结合最终的效果对操作进行评估。通过这一机制，数字虚拟人能够不断优化自身性能，适应新的任务和需求，实现在技术和服务水平上的

持续提升。这种反思不仅仅是对操作的简单总结，更是对整个工作流程的深入理解，从而为持续学习奠定基础。随着用户反馈和数据的积累，数字虚拟人的学习曲线不断攀升。这一持续学习的过程使得数字虚拟人能够不断进化，保持在不断变化的业务环境中的竞争力。例如舆情运营任务，通过让模型不断的去根据评论学习舆情监测，可以将其准确度和时效性大幅提升。

#### 2.5.1.2 架构孪生

基于上述技术生成的汽车行业孪生数字员工，能够完成一系列偏个人的流程任务，但涉及较多岗位协作的任务，就难以一次性执行。基于大模型群体智能体技术，如 AgentVerse[2-39]，我们不仅能够定义智能体本身的记忆、能力，还能够定义智能体之间交流的方式和逻辑，能够一定程度把现实人类的组织架构映射到数字孪生世界，生成对应真实公司架构的数字孪生架构。例如，ChatDev[2-41] 就通过架构孪生方式，构建了虚拟软件公司，以 CEO 为中心能够真实开展应用程序设计、编写调试等诸多任务。

AgentVerse 将多智能体环境划分为五个功能模块，并定义了各自的接口，用户可以根据自身需求重新定义不同模块的功能。这种可定制性使得数字孪生的架构不再受到固定的限制，而能够根据不同行业和企业的需求进行灵活调整。用户可以根据特定的场景和任务要求，定制数字孪生的架构，使其更好地适应实际应用场景。

AgentVerse 支持各种工具接口调用。这意味着在数字孪生的架构中，用户可以充分利用各种工具接口（Tool APIs），进一步增强数字孪生的功能和性能。这种工具支持的特性为数字孪生的架构提供了更多的可能性，使其更具实用性和可操作性。例如舆情运营任务中的数据工程师 Agent 可以通过调用代码仓库中的代码完成数据获取任务。

AgentVerse 的专家招募阶段在塑造多智能体群体的构成中起着关键性的作用，这一决策直接决定了群体的能力上限。研究已经证实，人类群体内部的多样性能够引入不同的观点，从而显著提高群体在不同任务中的表现。因此，在招募专家时，通过考虑他们的背景、技能和专业知识，可以确保引入足够的多样性，使得整个系统更具适应性和创新性。

协同决策是多智能体系统中至关重要的一环，而其成功实施与智能体之间的沟通模式密切相关。AgentVerse 的横向沟通模式鼓励智能体之间相互理解和协作。通过积极共享和细化各自的决策，群体能够形成一个集成函数，得出当前回合的共同决策。相比之下，AgentVerse 的纵向沟通模式更侧重于职责分工。其中，一个智能体提出初始决策，其他智能体充当评审人，提供反馈并不断完善决策。这种模式在需要迭代完善决策、达成共识的场景中表现较为出色，例如软件开发过程。通过垂直沟通，智能体能够有效地对解决方案进行评估和改进，直至达到共识。基于易慧智能的每一个单独的场景，Agent 之间都可以实现横向沟通和纵向沟通。例如 AI 培训场景可以采用纵向沟通，即课程设计 Agent 把设计好的课程给效果质检 Agent 进

行内容评估，评估通过后再把课程给内训 Agent 进行内部训练。

评估在架构孪生的实施过程中扮演着至关重要的角色，特别是对于下一轮专家组的构成调整和提升。AgentVerse 的奖励反馈机制作为评估的关键组成部分，通过比较当前状态与期望目标之间的差距，为系统提供有针对性的指导。这一机制不仅限于人工定义，还可以通过自动反馈模型来实现，具体取决于系统的实际情况和目标。这一持续的评估与反馈机制确保了系统的动态适应性和不断优化，使其能够不断提升性能并适应变化的环境。

#### 2.5.1.3 业务孪生

具备汽车行业数字孪生员工，以及数字孪生架构后，利用大语言模型的泛化能力，我们能够自动执行实际业务。执行不同的任务，我们需要整合大语言模型、搜索增强技术、智能体构建、群体智能技术，结合每个步骤的自主反思，才能不断优化业务执行效果。

XAgent[2-38] 是面壁智能创新的 AI 智能体框架，基于强大的 LLM 核心，具备自主解决复杂任务的能力。相较于传统智能体，XAgent 不受人类定制规则的限制，它不仅仅是被动执行任务的工具，更是一种真正具备自主智能的实体，能够独立理解人类指令、制定复杂计划并采取自主行动，这一自主的能力在业务孪生方面尤为重要。

传统智能体在解决问题时通常受到预设规则的限制，只能在既定范围内运作。然而，XAgent 通过赋予自主规划和决策的能力，实现了

真正的自主性。这使得 XAgent 能够独立思考、发现新的策略和解决方案，摆脱了对人类预设的束缚，为复杂问题的解决提供了更为灵活和创新的途径。XAgent 的自主性使其在处理多样性、复杂性任务时具有更大的适应性和创造性。

XAgent 的设计创新性地引入了一种“双循环机制”，使其在处理复杂任务时能够从“宏观”和“微观”两个视角进行全面考虑，类似于人类“左脑”和“右脑”的协同工作方式。

外循环承担着全局任务规划的责任，将复杂任务巧妙地分解为可操作的简单任务。作为一个“规划”领导者，XAgent 通过生成初始规划形成任务序列，并将每个子任务逐次传递给内循环解决。在这个过程中，外循环不仅监督任务的进度和状态，还通过「迭代优化」对后续规划进行不断改进。这使得 XAgent 能够高效地完成全局的任务分解和规划，展现出宏观任务处理的领导力。

在内循环中，XAgent 迅速转变身份，充当高效的「执行者」，确保外循环传递的子任务能够顺利达到预期。它能够灵活地检索外部系统中的工具，并根据子任务性质逐步求解。完成子任务后，内循环生成详细的反思，并将反馈信息传递给外循环，指示当前任务是否完成，以及在任务执行中的潜在优化点。这种「双循环机制」的设计使得 XAgent 能够高效、全面地应对各类复杂任务，从而提高业务执行效果。在 AI 新媒体运营过程中，XAgent 外循环将其分为直播控场，直播运营，视频编导，视频剪辑，质检员，创意设计等工作并分给对

应的 Agent，之后这些 Agent 可以对应相关的子任务逐步解决。

XAgent 在业务孪生的设计中充分考虑到与人机协作的问题，通过引入专为增强人机协作的交互机制，解决了类似 AutoGPT 存在的死循环、错误调用等执行出错的现象。这种设计不仅提高了自主解决复杂问题的能力，还突显了 XAgent 与人类之间全新的协作关系。结构化的通信方式是构建强大、稳定智能体的关键因素之一。例如，视频编导 Agent 当编排完一些列的视频时，可以讯问相关专家的意见，并根据相关反馈优化视频编排。

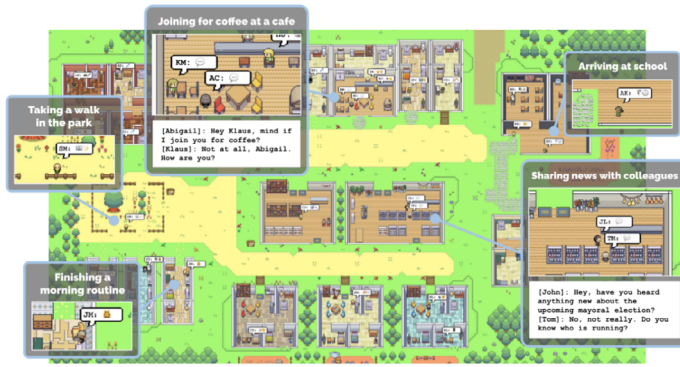
### 2.5.4 代表性组织孪生应用

下面介绍两个目前代表性组织孪生实现的典型框架，这些框架均基于大模型群体智能实现：

**（1）社会模拟：Generative Agents**

Generative Agents[2-45] 是由斯坦福大学和谷歌研究团队共同开发的一项技术。它主要是模拟可信的人类行为的计算软件智能体，旨在增强交互应用的能力，从沉浸式环境到人际交流的排练空间，再到原型设计工具。这些智能体可以模拟日常活动，如做早餐、工作、绘画、写作等，形成意见，注意到其他智能体，并与之进行对话。其核心是使用 LLM 记录智能体的经验，并将这些记忆合成为更高层次的思考，并动态地检索它们以规划行为。这一项目通过沙盒环境进行实验，其中包含 25 个智能体，用户可以使用自然语言与之互动。这些

智能体能够自主产生个人和社会行为，如组织聚会、结交新朋友等。




图表 2-55 Generative Agents 的虚拟西部世界小镇

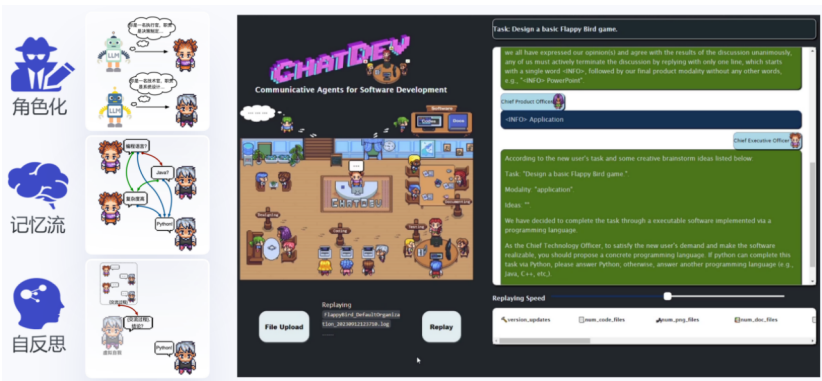
Generative Agents 的架构主要包括三个部分：动态记忆管理系统（Memory and Retrieval）、自我反思系统（Reflection）、和计划管理与执行系统（Planning and Reacting）。动态记忆管理系统包括记录虚拟角色经验的记忆流和根据虚拟角色的状况从记忆流中检索最合适记忆的记忆检索。这个系统会根据最近发生的记忆给予更高的得分，而久远的记忆得分较低。

### （2） 软件开发：ChatDev

软件开发是一个涵盖需求分析、系统设计、程序开发、系统测试和运维等环节的综合性活动。随着信息技术的发展，软件在现代社会中的作用日益重要，已渗透到各行业，如电子邮件、操作系统、数据库等。软件开发涉及多种专业角色（例如需求分析师、开发人员、测试人员等），由于角色间信息沟通和传递的复杂性，导致沟通成本高、

信息同步慢、存在领域隔阂，进而影响协同合作、难以保证效率和质量。

在大语言模型强大的综合能力背景下，本项目视文档（自然语言）和代码（编程语言）均为“语言”，充分利用大语言模型的理解与生成能力来构建自主智能体（Autonomous Agent），模拟不同专业角色在软件开发中的职能，激活大语言模型智能体群体协作的关键语言效能，实现群体协作的全流程自主软件开发。




图表 2-56 语言交互式群体协同的软件开发

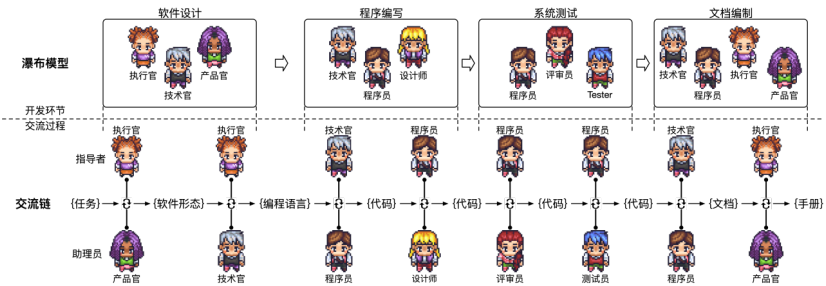
ChatDev[2-41] 设计了交流链（Chat Chain），将软件开发分解为由原子任务组成的“软件生产线”，子任务通过角色扮演交流实现智能体间的方案提议和决策研讨过程。在制作软件时一共需要经历设计 - 编程 - 测试 - 文档这四个大环节：

- **设计**：三个角色 CEO、CTO 和 CPO 讨论软件设计方案，决定游戏的呈现形式 (Web/ 桌面 / 移动端…) 和使用的编程语言。

• **编程**：程序员进行代码撰写，设计师进行 GPU 设计

• **测试**：代码的审查和实际运行两步，涉及「代码审查员」和「测试工程师」两个角色。

• **文档**：环境说明和用户手册两类，前者说明了游戏运行所需依赖的环境，由 CTO 指导程序员完成。而后者则由是 CEO 决定包含的内容，交由 CPO 进行生成。




图表 2-57 ChatDev 的交流链“软件生产线”

ChatDev 的软件制作平均时间小于 7.0 分钟且制作成本约 $0.3 美元。

第三章

# 融创赋能：大模型群体智能在汽车行业的融合创新与价值创造

## 3.1 整体赋能：大模型群体智能赋能汽车行业创造综合价值

在当今快速发展的汽车产业中，大模型群体智能技术正在以革命性的方式改写整车制造、供应链、研发和工程、销售分销、市场营销、售后服务、贸易与物流、租赁和金融服务、回收与再创造等各个环节，为汽车行业带来了前所未有的效率提升和个性化体验。

**提高企业运营效率：**传统的企业运作模式中，一直存在两大管理难题，一是跨部门之间的信息壁垒，二是组织层级之间的信息差问题。采用群体智能技术，每个部门的不同角色都会拥有智能体，智能体之间能够进行充分的信息分析与信息传递，能够互相协作、分工和执行，继而打破部门沟通壁垒，充分实现数据共享与业务融合，同时，能够确保公司目标、部门目标、个人目标的全面对齐。

**加快流程管理：**群体智能技术在汽车制造环节的应用主要体现在对生产流程的智能化管理上。通过多智能体的自动交互，可以实时监测生产线的运作状态，能够预测设备的维护需求，从而显著减少意外停机时间。例如通过对生产设备的使用模式和历史维护数据分析，智

能体可以预测哪些设备可能会出现故障，从而提前进行维护，以避免生产线的停顿。此外，智能体们还能通过智能分析生产数据，帮助制造商优化零部件的库存管理和供应链，从而减少库存成本，提高生产效率。另一方面，智能体们还可以根据市场需求、原材料的供应状况和生产能力，智能调整生产计划，确保生产线的高效运转。

**提升营销体验：**汽车的营销环节，需识别关键用户触点，既要结合触点打造差异化的运营模式，也要保障品牌的一致性。根据用户关键旅程的的不同阶段，群体智能技术可以根据每个阶段的核心目标为导向，结合真实场景工作流，采用不同的多智能体组合，模拟各阶段的工作角色，由不同的智能体自动完成内容创作、营销话术生成等任务，并将结果与相关的智能体串联起来反馈人类员工。例如与潜客进行沟通中，会话智能体会收集用户个人偏好、购车需求，资金预算等信息，调用品牌及产品信息，保障营销话术输出的专业性、准确性，并采用多轮对话的方式，持续加深用户对该品牌车系在营销环节的信任与好感，提升营销转化率。

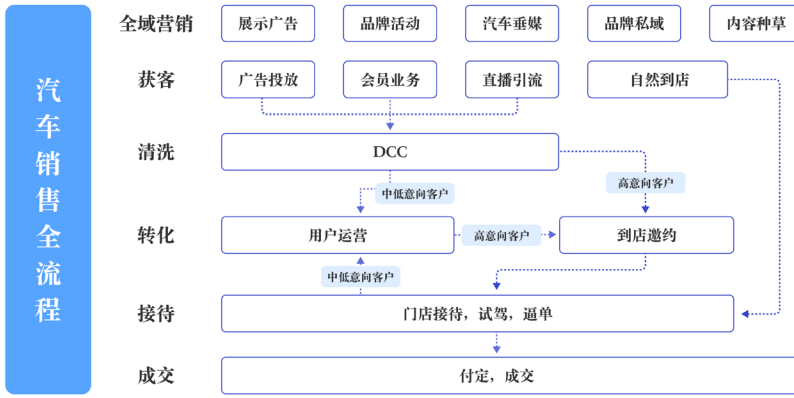
**增强服务感受：**群体智能技术提供的平台服务，可以提供全天候、陪伴式的在线咨询服务，即时回答售前、售中及售后各环节中客户提出的关于车辆性能、金融政策、购置手续、维修保养的问题，为不同阶段的购车用户提供更加精准和个性化的线上解决方案。既从用户侧维护了品牌的满意度与忠诚度，同时也为企业持续积攒宝贵的客户行

为数据。这些数据在经过分析后可以帮助企业不断改进产品设计和服务。例如，通过分析客户对车辆功能的使用情况，企业可以了解哪些功能受欢迎，哪些需要改进，从而在未来的车型研发中进行相应的调整和优化。

**提高企业规划能力：**群体智能技术在市场分析和预测方面的应用能够为汽车企业带来巨大的竞争优势。通过分析海量市场数据，智能体能够快速捕捉并预测未来的市场趋势，帮助企业制定更加有效的市场策略的同时，对特定车型的需求增长或新兴市场的发展趋势给出更清晰的判断与解读。智能体还可持续性的对市场趋势与客户行为进行分析，帮助销售团队识别潜在的热门车型和市场机会，从而更有效地制定销售策略和目标。

## 3.2 营销赋能：大模型群体智能赋能汽车营销五大核心场景

作为消费领域举足轻重的超级大单品，汽车的营销具有其它消费品难以企及的销售难度和销售周期。高客单价、低成交率和长销售生命周期是汽车销售的特点。为应对营销场景中的诸多困难，提高销售效率，汽车营销流程经历了长时间的发展，已经沉淀了一套标准化、全闭环的方法论。




图表 3-1 汽车销售环节全流程示意图

一个完整的汽车营销流程分为获客、清洗、转化、接待和成交五个方面，各个环节的链式衔接构建了自洽的商业闭环：在获客阶段，汽车销售人员通过展示广告、品牌活动、汽车垂媒、品牌私域、内容种草等手段或渠道进行获客，叠加会员业务、直播等引流手段，销售人员可以迅速获得大量的潜在客户基础画像与联系方式。

当潜在客户进入到客户名单后，电销人员将进行外呼开展沟通，充分了解客户的购车偏好、预算范围及决策逻辑后，将客户分为高意向客户和中低意向客户两个阵营。针对高意向阵营客户，电销人员会邀约到店进行试驾，通过线下门店创造的良好体验促进成交转化；而针对中低意向阵营客户，需要开展一系列潜在客户孵化和种草培育的工作。

由潜在客户经过一系列前期营销过程最终成功购车，是一个周期

较长、转化率较低的环节，需要资深的销售经理不断联络客户进行沟通，了解客户真实需求的同时逐步完善客户画像；另一方面，销售经理需要引导并帮助客户逐步明确自身购置需求，做出合理的购车决策。一旦潜在客户流露出较为积极的成交意向便会被精准捕捉，引导用户到店试驾及时促成购车转化。

针对汽车营销环节长久积累的应用实践痛点，基于汽车营销的核心场景增长需求和对 AI Agent 应用落地的独特理解，易慧智能联合面壁智能和清华大学 NLP 实验室推出五大智慧营销解决方案，分别为集约 DCC、用户运营、数智研究院、舆情运营、新媒体运营，通过群体智能技术实现汽车营销业务的组织孪生，提高行业效能。


图表 3-2 赋能五大营销场景示意图

### 3.2.1 新媒体运营

移动互联网的发展创造了大量的内容社区，各品牌借势通过各新媒体平台（如抖音、小红书、微博等）以多模态形式进行内容宣发与消费者建立情感链接。主机厂在面对多品牌、多车系、多营销节点的背景下，对一线人员的创作和运营能力、品牌标准的一致性、多平台合规分发和投放转化都提出了更高的要求。如何在激烈的竞争中脱颖而出，保持品牌的独特性和影响力，成为了一项重要考验。

生成式大模型支持下的营销智能体助手秉承赋能人的理念，将全流程营销内容生产进行自动化和结构化沉淀，通过多模态模型、大语言模型、群体智能技术赋能新媒体运营人员实现生产力的不断提高。通过将不同关键环节的工作智能体化，新媒体智能营销解决方案为一线运营人员构建了属于自己的内容落地团队，极大提高了内容产出的效率，赋能汽车销售的获客场景。

### 3.2.2 集约 DCC

外呼邀约是汽车营销全流程中最重要的环节之一。因为其工作过程繁琐、对沟通能力要求高、质检困难等特点，一直以来都是制约获客转化效率的痛点和难点。首先，在传统人工邀约的情况中，外呼专员的人员业务素质往往难以量化，只能通过成功率等数据进行评判，针对性迭代优化更无从谈起，成为制约环节效率的重要影响因素。其次，相关岗位人员成长性低，由此带来的高流动性带来了高昂的培训

成本。再次，面对这样一个低成功率的工作流程，即使专业的销售人员也很难做到每一个失败的沟通进行详细分析，故难以积累形成有效的改善方法论。

群体智能技术基于大模型的能力及对呼叫中心客户的组织孪生，极大解决了该环节的痛点，提升业务转化效率。大语言模型的类人理解能力和即时反馈能力使其成为解决传统外呼获客环节中效率损失问题的理想工具。模型通过精准解析人类语言和意图，能够有效减少由于人为不稳定因素引起的错误和延误，并通过数字化沟通流程等方式构建全流程迭代机制，实现运营效率的提升。

### 3.2.3 用户运营

用户运营场景包括了粉丝运营（从触达到建联）、潜客运营（从建联到成交）和保客运营（从成交到复购和 LTV 最大化）。粉丝运营场景要解决用户在多平台多通道主动沟通和即时服务的需求，用户运营场景需要解决用户在长周期转化路径上的触点服务，保客运营场景需要解决用户购车后 3-5 年保有周期内的 NPS 管理和价值挖掘。要想做到全场景用户运营，传统解法需要铺设人力面对大量的运营任务或给对渠道服务人员培训赋能，既要消耗大量的人力物力，又很难做到一致化的服务，面临成本和效果的两难选择。

针对这一行业痛点，基于群体智能技术的用户运营解决方案可以极大释放运营人员的运营压力，通过对用户运营工作目标和工作流程的梳理和拆解，构建并串联不同角色的运营 Agent 自动化执行运营任务，完成运营目标。

### 3.2.4 舆情运营

舆情运营对于汽车企业来说同样至关重要，尤其是在数字时代，网络上的舆论和评价对品牌形象和产品销售具有巨大影响。传统的舆情运营方式依赖人力跟踪和处置，但面向多平台多用户多模态信息的监测、分析、策略、响应处置，依靠人力成本高、一致性策略不足。

基于群体智能技术的舆情运营系统运用大语言模型的逻辑分析能力，可以及时进行舆论事件的观点梳理和回评建议生成反馈运用人员，并通过授权账号托管进行重点舆情账号控评。同时可以实现对达人账号进行动态监控，及时掌握关键意见领袖的动向；发布内容自动归类功能，有助于快速整理和分析海量信息；针对平台热点扫描，进行品牌营销相关性分析及热点借势建议，增强品牌曝光度及影响力，实现企业舆情运营资产沉淀。

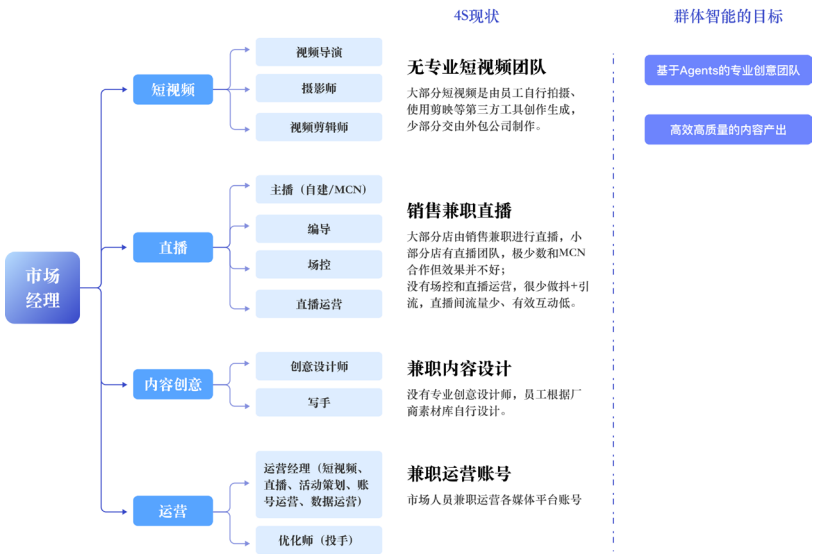
### 3.2.5 数智研究院

在汽车行业，及时准确的消费者需求洞察和竞品分析对车型产品定义、产品迭代及营销企划具有至关重要的意义。针对这一需求，基于群体智能技术的数智研究院利用大语言模型、RAG 等技术，将高质量、细粒度数据报告的生成过程全自动化，为汽车企业提供关键的行业数据分析，如购车用户需求分析、竞品车型分析等。在传统方法中，这类数据分析任务的完整执行及分析报告的制作往往耗费巨大的人力和时间成本，且时效性不佳，影响车企进一步的业务决策。在大语言模型群体智能技术的加持下，通过自动收集全网数据，高效进行数据治理和分析，显著缩短数据分析和报告生成所用的时间，并提高了报告的实效性与准确性，为车企提供更有价值的数据洞察。

## 3.3 实践案例：新媒体获客场景下的群体智能整体解决方案

### 3.3.1 新媒体运营场景解决方案

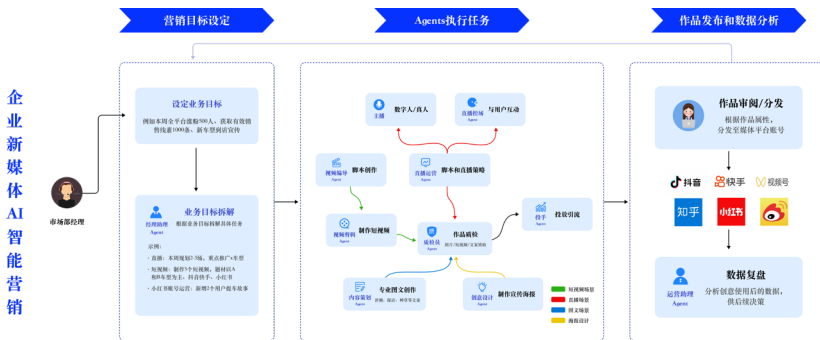
新媒体已经成为营销过程中的重要一环，如何快速、高效的生产优质专业的内容已成为主机厂与终端市场部门的重要课题。



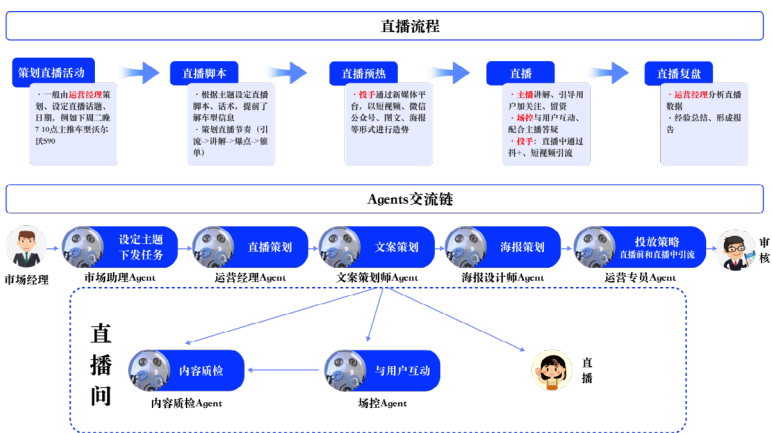

图表 3-3 新媒体获客场景现状分析与目标

新媒体营销场景中，市场经理负责业务目标的制定和产出结果的审阅，其余全部任务，如任务拆解、执行和复盘则由 AI 智能体自动完成。以“本周新媒体账号增加 500 名粉丝”为例，这个目标不仅反映了增长的需求，也强调了通过内容和用户运营提升品牌价值的重要性。目标一旦设定，运营经理 Agent 就会进行任务拆解：第一项任务是本周策划 2 场直播，旨在强化品牌信息的传播和用户的互动；第二项任务是制作并发布 5 个精准定位品牌价值观的短视频，分发至抖音、快手、小红书等平台，以提升品牌影响力；第三项任务是保持平台账号的活跃运营，及时回复用户评论，增进用户参与感。在此过程中，文案策划师 Agent 和海报设计师 Agent 智能体将基于品牌价值观创作有吸引力的文案和宣传海报，为主题活动吸引目标用户群，同时负责内容的

生产、信息的传递和交付。运营专员Agent将内容上传至各个内容平台，并确保市场部经理对内容进行人工确认，以确保品牌形象的一致性和专业性。最终，活动结束后，运营经理 Agent 将负责分析数据，确保每一次创意的使用都能为品牌带来实际的营销效果，这些分析结果将为后续的内容运营和用户运营决策提供数据支持。




图表 3-4 新媒体获客场景解决方案分析




图表 3-5 新媒体获客场景产品设计

### 3.3.2 集约 DCC 场景解决方案

外呼邀约是汽车营销全流程中最重要的环节之一，群体智能技术基于大语言模型的能力及对呼叫中心客户的组织孪生，充分利用大语言模型的类人理解能力和即时反馈能力，通过精准解析人类语言和意图，有效减少由于人为不稳定因素引起的错误和延误。此外，大模型的技术特点使其成为理想的工具，通过数字化沟通流程等方式构建全流程迭代机制，实现运营效率的提升。基于大语言模型的技术特点，邀约 Agent 解决了外呼语义快速理解和及时反馈的挑战，具备以下能力拓展：

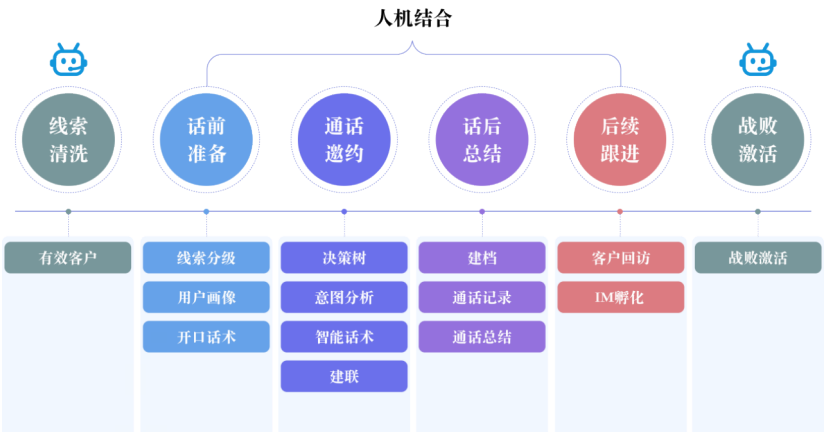
**1）快速理解客户语言和意图**：大语言模型能够即时解析客户在通话中的语言，包括方言或专业术语的理解，快速捕捉客户需求和意图，从而提高邀约对话的相关性和针对性。

**2）个性化反馈与邀约话术生成**：邀约 Agent 能够根据客户的具体需求和反馈，智能生成个性化的邀约话术。这些话术不仅针对性强，而且能够灵活应对客户的各种反馈，提高邀约的成功率。

**3）实时调整沟通策略**：在与客户的互动过程中，邀约 Agent 能够根据客户的反应和态度实时调整沟通策略。例如，当感知到客户的犹豫或反对时，Agent 能够即时提供更多信息或调整邀约方式，以增

强客户的信任和兴趣。

**4）全流程迭代机制**：通过记录和分析每一次邀约的详细过程，包括成功与失败的案例，邀约 Agent 不断学习和优化其邀约策略和话术，这种全流程迭代机制使得外呼邀约的精准度大幅提升。




图表 3-6 集约 DCC 场景解决方案

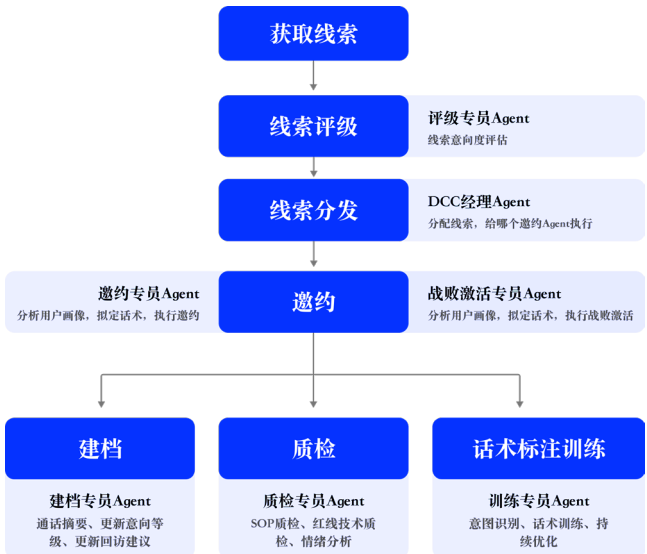
首先，实现组织孪生与 Agent 团队构建。群体智能技术能够基于线索清洗的完整业务流程及实际岗位角色，构建出存在于数字空间的孪生 Agent 外呼团队，这个团队中的每个 Agent 都被设计以执行特定的工作任务。

其次，设定销售目标并分配任务。销售部门经理可以通过数字员工管理平台设定销售目标，并下达命令给到数字员工团队中的 DCC 经理 Agent。DCC 经理 Agent 可以根据 Agent 团队的配置情况，将销售部门经理的指令进行细化，拆解为一个个单一目标分派给各个环节的 Agent（包括评级专员 Agent、邀约专员 Agent、建档专员 Agent、质检专员 Agent 和训练专员 Agent）进行具体执行。例如，设定本月要提升 10% 的客户到店率作为目标，经理将这一目标分解为每个 Agent 需要完成的具体邀约数量。

然后，实现客户筛选与个性化邀约。在单一命令的指引下，评级专员 Agent 通过 CRM 系统对待邀约的客户进行分类，将邀约到店转化成功率高的客户进行区分，并制作单个潜在客户画像档案。邀约专员 Agent 可以根据每个客户的不同特点进行结构化分析，通过与客户的交流生成决策树并分析客户意图，智能生成定制化邀约话术，提高邀约成功率。

最后，完成质量监控与知识积累。在通话结束后，通话记录会自动保存并交给建档专员 Agent 进行分析总结，建档专员 Agent 和质检专员 Agent 会根据邀约过程和结果对邀约情况进行跟进、质检和复盘，建立档案沉淀企业知识库。同时，训练专员 Agent 可以通过对失败经验的总结与量化评估，持续优化话术和对用户意图识别的能力，不断提高 Agent 团队的邀约效果。综上，基于群体智能的 DCC 组织孪生框架，实现呼叫中心的组织孪生，用大模型技术赋能外呼邀约业务场

景，克服了传统解法中的诸多痛点，提升了 DCC 场景的经营效能。




图表 3-7 集约 DCC 场景群体 Agents 的组织架构

### 3.3.3 用户运营场景解决方案

用户运营场景中，包括粉丝运营、潜客运营、保客运营三个子场景，其中潜客运营是汽车营销过程中的一个关键环节，具有培育周期长、过程复杂但收益效果显著等特点。基于群体智能技术的用户运营解决方案，由数字员工自动化、精细化地分析客户画像，并实时、自动、个性化地与客户进行互动，进而构架出用户体验更加优化的用户运营工作流，极大释放销售人员的客户培育压力，提高销售各个环节的转化效率。

当一个潜在客户进入流程后，运营专员 Agent 将主动触达潜在客户并进行深入沟通，实行自动添加微信好友、发送破冰话术等操作。运营专员 Agent 会自动跟进客户沟通并实时进行分析，理解客户的真实需求的同时定期为客户推送高质量信息，定制化生成用车方案。在整个过程中，客户会根据 Agent 提供的专业化、个性化建议逐步确认自身的真实需求、明确购车预算、做出购车决定，完成潜在客户孵化的流程，成为可以进入转化流程的成熟客户。

在孵化的过程中，运营专员 Agent 会实时地与客户进行交互，质检 Agent、标注 agent 和建档 Agent 则可以在环节中实时检查 Agent 的跟进情况，进行跟进情况质量分析，检视客户画像，反馈数字员工管理平台。任何运营专员 Agent 和客户交流的过程，都会由被总结记录，并由训练标注 Agent 总结成业务经验并沉淀下来，形成 Agent 的工作流的迭代机制，使得智能体孵化客户的效率在持续学习迭代过程中不断提高，为营销场景的转化率提高持续赋能。

以某汽车品牌销售案例为例，其销售团队引入了 AI 智能体技术，实施了一次针对中等收入家庭的潜客运营活动。活动开始前，分析 Agent 基于先进的数据分析技术，识别出了一批有意购买新车但尚未做出决定的家庭。在活动实施阶段，线上运营 Agent 与这些家庭建立联系，并通过家庭成员的在线互动，深度理解他们的需求。例如，用户分析 Agent 发现某家庭关注的是汽车的安全性和燃油经济性，便定制化推送了该品牌在这些方面的优势信息，并提供了详细的用车成本分析。通过周期性的客户沟通和个性化的营销素材触达，该家庭最终

确信了品牌的价值主张，并在孵化期结束时前往经销店进行试驾，最终完成了购车。整个过程中，运营质检 Agent 和建档专员 Agent 持续监控客户沟通的有效性，并由训练标注 Agent 在后续为运营团队提供了改进沟通策略的建议，从而在后续活动中不断提高转化率。

在基于群体智能技术的用户运营场景中，销售经理只需通过数字员工管理平台就可以实时查看整个智能体团队的工作情况，其工作能力边界与范围得到了极大拓展，并在专业内容上得到 Agent 的强大赋能，从而打破和客户的藩篱，有效促进中低意愿客户向高意愿客户转化，提升了企业的全流程营销效率。

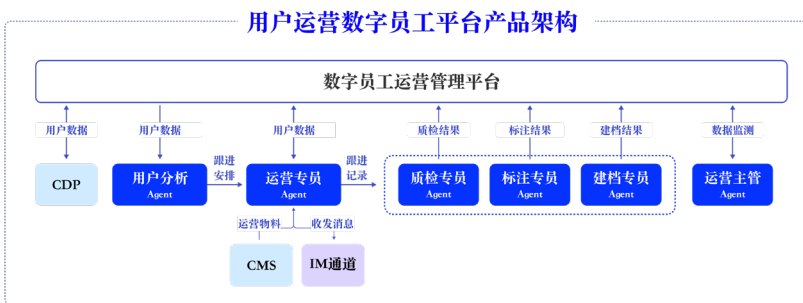
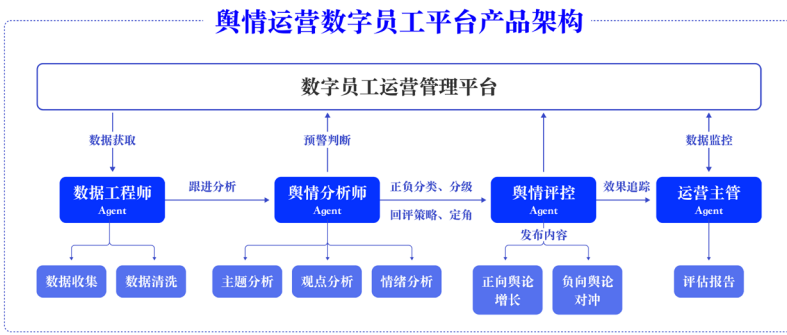



图 3-8 用户运营场景解决方案

## 3.3.4 舆情评控场景解决方案

基于智能体技术的舆情运营体系，以舆情控评场景为例。数据工程师 Agent 开启目标平台数据搜集工作，搜集关注的汽车品牌和车型相关的舆情信息，既包括舆情主体内容也包括评论。搜集到的数据

交付到舆情分析师 Agent，该 Agent 基于大语言模型和舆情分析方法论，对获取的数据进行实时的聚类和分级，有效判断出公众的情感倾向，以此来区分正面与负面的舆情信息。基于这一分析结果，舆情控评 Agent 生成针对性的回复内容并可实现授权后的托管运营。最后运营主管 Agent 将持续监控这些干预行为后的舆论变化，确保所采取的策略达到预期的效果。



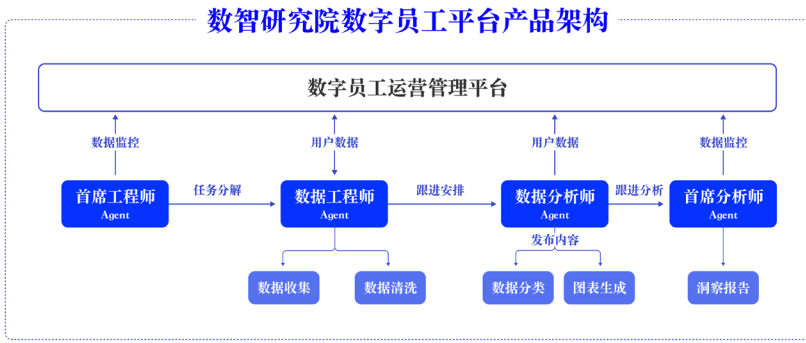

图表 3-9 舆情运营全流程

### 3.3.5 数智研究院场景解决方案

在汽车行业，及时准确的消费者需求分析和竞品分析对车型产品定义、产品迭代及营销企划具有至关重要的意义。然而，在传统方法中，这类数据分析任务的完整执行及分析报告的制作往往耗费巨大的人力和时间成本，且时效性不佳，影响车企进一步的业务决策针。面对这一问题，基于群体智能技术的数智研究院组织孪生框架，通过智能化解决方案大幅提升数据分析的时效性和准确性，帮助车企实现更精准的产品定位和更深度的消费者需求洞察。

数智研究院的业务流程可分为以下几个关键步骤：（1）数据收集与监控：Agent 在多个渠道（包括社交媒体、汽车垂媒等）全面监控和收集相关数据，确保数据的全面性和多样性。（2）数据处理与分析：基于大语言模型的语义理解能力，结合相应的数据治理和分析工具（ETL 工具、BI 系统等），Agent 能够自动地从海量数据中提取关键信息，形成一级分类，二级分类，从而完成对数据的有效治理。（3）报告生成：根据企业具体需求由智能体自动定制高质量、具有洞见性的数据报告，如购车用户需求分析、竞品车型分析等。

以生成一份《某车型的用户洞察报告》为例，为了深入理解该车型用户的偏好和需求，基于群体智能技术的数智研究院采用以下步骤生成详细的用户洞察报告：首先，首席分析师 Agent 明确报告旨在解析该车型用户的特征、购车动机及对车辆的综合评价。然后，数据工程师 Agent 自动社交媒体、汽车垂媒等多渠道收集了关于该车型的用户观点。接着，数据分析师 Agent 基于大语言模型，对收集到的文本数据进行情感分析、关键词提取和主题归类，细化到用户对该车型的外观、内饰、空间等具体方面的评价，并自动生成可视化的数据图表。最后，首席分析师 Agent 根据分析结果，生成包括用户画像、购车动机分析、车辆评价综合分析等内容的完整数据报告，并将报告拆分为若干个子模块，帮助主机厂进行持续性的数据追踪和针对性的用户深访，有效赋能产品定义等业务。




图表 3-10 数智研究院

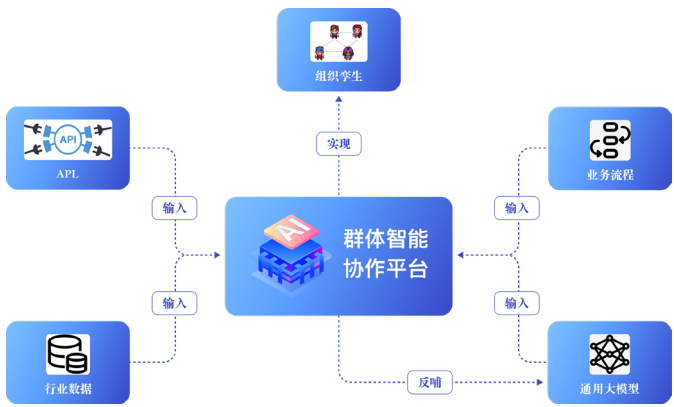
通过群体智能技术，数智研究院能够有效地支持汽车企业在复杂多变的市场环境中作出数据驱动的决策。上述案例展示了群体智能在数据分析和处理方面的强大能力，不仅为该品牌汽车提供了深度的市场和用户洞察，也显著提升了研究的效率和质量，帮助企业在激烈的市场竞争中保持领先。

第四章

# 生态矩阵：汽车行业大模型群体智能生态矩阵建设

## 4.1 总体格局：汽车行业群体智能生态矩阵的理念与布局

基于群体智能实现组织孪生，在技术层面，需要以群体智能协作平台为载体，输入通用大模型、行业数据、API、业务流程四类技术要素。




图表 4-1 群体智能实现组织孪生核心技术要素

### 4.1.1 群体智能协作平台

群体智能协作平台是汽车行业实现组织孪生的载体，围绕群体智能协作平台，能够充分整合通用大模型伙伴、行业工具伙伴、行业解决方案伙伴和行业数据伙伴，实现技术和能力的互补。


图表 4-2 汽车行业大模型群体智能生态矩阵

### 4.1.2 通用大模型伙伴

通用大模型伙伴指提供各种基础模型相关服务和解决方案的公司或机构，譬如文心一言、通义千问、智谱清言、百川大模型、腾讯混元、盘古大模型、面壁 LUCA 等基础大模型厂商。在汽车行业的群体智能生态中，通用大模型是群体智能技术的基石，是各类智能体智慧之源泉，也是实现汽车行业各类复杂业务的源动力。

### 4.1.3 行业数据伙伴

行业数据伙伴包括行业协会、汽车厂商及经销商、第三方数据机构、媒体平台和咨询公司，发挥着关键的数据提供和分析作用。为了有效利用这些数据资源，企业和数据提供方需要进行周密的数据准备和治理工作。

### 4.1.4 行业工具伙伴

### 4.1.5 解决方案伙伴

解决方案伙伴包括针对汽车行业的咨询公司和解决方案提供商，

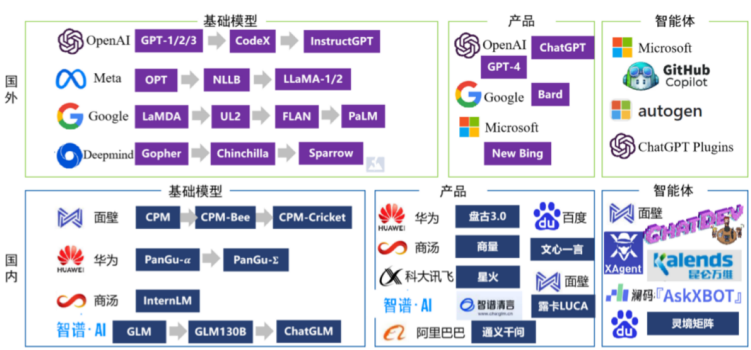
熟悉企业与行业的业务流程，拥有差异化的解决方案产品矩阵，积累了大量的行业和典型场景的数字化项目经验，能够为企业提供专业知识、最佳实践和定制化解决方案，帮助汽车公司应对挑战、优化业务流程、并驱动创新。

## 4.2 战略要点：汽车行业群体智能生态的核心问题与解决方案

### 4.2.1 通用大模型百花齐放，选择难

#### （1）大模型选型原则

大模型选型不仅是选择一个大语言模型，更是确定一个与企业特定需求和目标相匹配的工具，进行精确的模型选型变得非常重要，如今国内外大语言模型生态如图表 4-3 所示。大模型选型通常考虑以下因素：




图表 4-3 国内外大语言模型生态

- **特定需求的匹配性：**在汽车行业，大语言模型的应用可能涉及到从客户服务到工程设计的各个方面。例如，对于需要处理大量客户查询的场景，选择在自然语言理解方面表现出色的模型会更加合适。而对于涉及复杂工程术语和技术数据的应用场景，选择专注于行业特定术语处理能力的模型则显得尤为重要。

- **性能和效率：**在大量数据和复杂查询的处理上，不同的模型展现出不同的性能和效率。例如，技术文档理解和生成一类的应用场景对于模型的性能有较高需求，对效率需求不高；而对于客户服务对话等场景则对于模型响应的实时性要求较高。针对业务特点选择适合的模型可以平衡性能和效率的需求。

- **集成和兼容性：**企业已有的技术堆栈和数据架构需要与选用的大模型兼容。选型过程需要考虑这些技术细节，以确保模型能够无缝集成到现有系统中。

- **成本效益：**不同的大模型在成本效益上有所不同，复杂模型可能需要更多的计算资源和投资。以 OpenAI 提供的大模型服务为例，长文本模型 GPT-4-32k 的价格是短文本模型 GPT-4 的两倍。正确选型有助于平衡性能和成本，为企业带来最佳的投资回报。

- **定制化和可扩展性：**不同的模型在定制化和微调方面的能力不同，同时随着企业的发展和市场的变化，模型可能需要一定扩展性以适应新的需求和挑战。通常来说，开源模型较闭源服务更具灵活定制

化能力和扩展性，但维护成本较高；闭源服务则易用性和稳定性更有保障。选择平衡具有良好扩展性和易用稳定性的模型与服务对确保业务的长期有效十分关键。

- **数据安全和隐私**：在处理敏感数据，如客户信息或汽车生产制造专有技术时，模型的数据安全性和隐私保护能力尤为重要。选型过程中需要考虑这些因素，以保护企业和客户的利益。

**（2）大模型测评方案**

大模型的评测是确保选择的模型最适合汽车行业特定需求的关键步骤。这个过程应该综合考虑模型的能力、兼容性、成本效益以及对现有业务流程的影响。汽车行业评测大模型的一般方案包括：

**a. 定义业务需求和目标：**

- **目标应用场景**：定义模型将应用于哪些具体场景，如客户服务自动化、设计工程辅助、制造过程优化等。

- **性能指标设定**：针对每个应用场景设定具体的性能指标，例如，在客户服务中可能关注问题解答的准确率和响应时间，在设计工程辅助中则可能更侧重于模型对复杂工程术语的理解能力。

**b. 数据集准备和预处理：**

- 选择或构建与评测场景相关的数据集，如客户咨询记录、技术

规格文档等。

• 对数据进行预处理，包括数据清洗、标准化和格式化，以确保评测的准确性。

**c. 性能评估：**

• **基线建立**：选择或创建基准模型作为对比，如使用行业内已知的标准模型。

• **自动化测试脚本开发**：编写自动化测试脚本来模拟真实场景下的模型使用，以确保评测的一致性和可重复性。

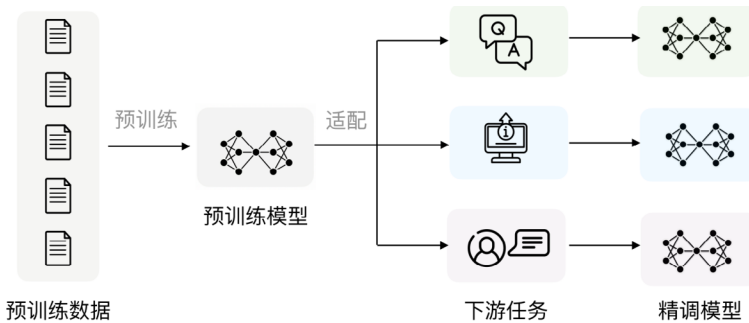
• **场景模拟测试**：在不同的应用场景中运行模型，记录其性能指标，如准确性、响应时间、错误率等。

**d. 兼容性和集成测试：**评估模型如何与现有的技术堆栈和数据平台集成。兼容性不佳的模型可能需要额外的时间和资源来集成，这可能增加总体成本和复杂性。此外，考虑模型的数据输入和输出格式是否符合企业的现有数据处理流程。

**e. 成本效益分析：**计算模型部署和运维的成本，包括硬件资源、软件许可、人力资源等。估算通过使用模型可能带来的效益，如效率提升、成本节约等，并与成本进行对比。

**（3）大模型适配**

大模型适配是对预训练大模型进行特定调整的过程，以使其更好地适应特定的应用场景或数据集。通过模型微调的方式将通用大模型适配汽车行业各类业务，不仅允许模型保持其原有的广泛语言能力，还可以在特定任务或领域上获得更加精准的表现。例如，一个在通用数据上训练的模型可能擅长理解和生成自然语言，但通过在汽车行业的技术文档上进行微调后，它可以更好地理解和使用汽车行业的专业术语和概念。大模型的微调适配包含以下概念：




图表 4-4 大模型微调适配下游任务

**1）微调对象**

并非所有大模型都可以支持微调，按照开放程度，可将其分为三类：不可微调、通过 API 微调、可完全微调。

**不可微调模型：**这类模型通常以黑盒的形式提供服务，用户无法对其内部结构或训练数据进行修改或定制化调整。某些特定领域的服务使用专门训练的模型，这些模型为了保持稳定性和可靠性，不提供

微调的选项。尽管扩展性受限，但只要在其所面向的封闭场景中，该类模型和服务将发挥良好的性能。例如在汽车生产线质量控制系统中，厂商可部署视觉识别系统来检测生产线上的缺陷或错误，该系统使用的模型经过训练可以高效识别特定类型的缺陷，并且这些模型一般是封闭的，无需用户进一步进行微调。

**通过 API 微调模型：**这些模型允许通过云服务 API 接口进行一定程度的定制化，但用户无法直接访问模型的训练过程。API 微调通常限于参数调整或利用预设的微调功能。目前主流的大模型服务如 ChatGPT、文心一言等均支持 API 微调模型。API 微调模型的优势在于它们为用户提供了一种灵活的方式来利用强大的语言模型，同时无需直接处理模型的复杂性或进行大规模的数据训练。用户可以通过简单的 API 调用来定制化模型的输出，适应不同的业务需求和应用场景。然而，这种方式的定制化程度有限，无法进行深度的模型重训练或参数优化。

**可完全微调模型：**这类模型允许用户直接对模型进行微调，包括在特定数据集上重新训练模型的某些层或参数，甚至对模型进行二次开发，从而定制化模型以适应特定任务或领域。所有的开源大模型均支持完全微调，但有研究表明，开源模型的基础性能一般弱于闭源模型，并且部分模型只开源了基础规模的模型参数，没有开源最大规模的模型参数。完全微调模型的优势在于能够更精准地定制模型，能够更好地适应特定的业务需求和应用场景，同时在数据隐私和安全性方面风险更小。然而，进行这种深度微调需要一定的机器学习知识和计算资源。

### 2）微调时机

现今大模型已在海量数据上进行了预训练从而掌握了广泛的世界知识，模型在一些通用的任务和场景中无需微调也可取得良好的性能。目前的大模型在汽车行业从生产到销售以及客户支持的通用概念的理解上已有较好表现。因此，是否需要对模型进行微调也是首要考虑的问题，通常可考虑以下几个方面：

通用模型可能很难理解汽车行业特定细分领域的上下文或专业术语，特定的业务流程和场景可能没有在模型的训练数据中出现过，模型在该环境下表现很差，此时对模型进行微调则十分必要。

由于大模型具备上下文学习的能力，在考虑微调之前可以先尝试使用不同的提示方法来引导模型适配特定的任务。如果调整提示后模型的性能仍然不符合需求，那么可以再考虑对模型进行微调。

在需要快速迭代和反馈的业务场景中，微调需要更长的时间来创建数据集和训练模型，难以满足实时性。此时可以尝试优化输入提示来快速地适配模型。

高质量的任务特定数据集和足够的计算资源是模型微调的必要条件，微调需要足够的数据来覆盖目标任务的各种情况。数据量过少可能导致模型过拟合，即模型过度适应训练数据，而无法泛化到新的数据上。

### 3）微调流程

大模型的微调通常包括以下几个主要步骤：

**目标定义：**明确微调的目标和预期成果。包括提高在特定任务上的准确性、加快响应速度、或增强对特定类型数据的处理能力。确保微调目标与业务目标一致，以实现实际应用的有效性。

**数据准备：**根据微调目标选择或创建数据集。这涉及从现有数据源中筛选数据，或者通过众包、专家标注等方式生成新的数据，确保数据集不仅样本量足够，而且涵盖了多样性，以避免模型偏见和过拟合。

**预处理：**对数据进行必要的预处理，包括清洗、标准化和格式化，移除无关、重复或错误的数据，确保所有数据遵循统一格式。数据的质量和一致性对于微调的成功至关重要。

**选择微调策略：**根据模型的特点和微调目标，选择适当的微调策略。在计算资源和数据充足的条件下可选择全量微调的方式深度修改模型参数，其次可以使用包括 P-tuning 和 LoRA 等方式在资源受限的条件下对模型高效微调。

**实施微调：**在特定数据集上运行模型，对其进行实际的微调。这个过程通常涉及多次迭代，逐渐优化模型的参数以达到最佳性能。监控训练过程，记录关键指标以帮助调整和优化。

**评估与调整：**评估微调后模型的性能，确保其满足预定目标。使用独立的验证集来评估模型。性能，确保微调后的模型在真实世界数据上的有效性。基于评估结果调整微调策略，可能包括调整学习率、修改训练数据或改变微调的强度。重复以上步骤，直至达到性能预期。

**部署应用：**在类似生产环境中测试模型，确保其在实际应用中的稳定性和效果。部署后持续监控模型表现，以便及时发现并解决任何可能出现的问题。一旦模型性能达到满意水平，将其部署到实际应用环境中。根据反馈和长期性能，定期更新模型以维持或提升性能。

### 4.2.2 行业数据质量有待提高

汽车行业在数据方面的现状存在诸多挑战。首先，数据的完整性和准确性有待提高，由于汽车制造商、供应商和维修服务提供商之间的数据共享不足，导致数据无法形成完整的生命周期。其次，数据治理水平有待加强，如何确保数据的合规性、隐私保护和安全使用是当前亟待解决的问题。此外，由于汽车数据的敏感性，如何在保证数据安全的前提下进行有效利用也是一个难题。对此，可以用数据评估、数据治理、保障数据安全三个纬度提高数据质量。

#### （1）数据评估

数据准备是数据分析的基础。企业和数据提供方需要确保收集的数据是准确、完整和更新的。在数据收集过程中，应重视数据的多样性和代表性，确保所收集的数据能够全面反映市场情况和消费者行为。

此外，收集数据时也要重视数据的版权和用户隐私等问题，对于敏感数据要做脱密和匿名化处理，确保数据的合规性。数据质量直接影响后续分析的准确性和可靠性，因此，进行有效的数据清洗和预处理变得至关重要，这包括识别和修正错误数据、处理缺失值和去除重复记录。同时，还需对数据进行标准化和规范化处理，以确保数据在不同来源之间具有可比性。

评估已有数据是理解数据价值和限制的重要步骤。这包括分析数据的准确性、一致性、完整性和时效性。对于历史数据，重点是其在当前分析中的相关性和有效性。数据评估还包括对数据的结构化程度和可分析性的评估。例如，非结构化的文本数据可能需要通过特定的处理方法转化为可分析的格式。此外，数据评估还涉及检查数据中的偏差和异常值，这对于确保分析结果的准确性和公正性至关重要。数据的探索性分析，如使用统计图表和汇总统计，也是评估数据的重要组成部分，它有助于揭示数据的基本特征和潜在的洞察。

在整个数据准备和评估过程中，保持透明度和可追溯性也非常重要。这意味着记录数据处理的每一个步骤，包括所做的更改、采取的决策以及任何潜在的问题。这不仅有助于后续的数据分析工作，也确保了整个分析过程的可靠性和可复制性。通过这些细致和周到的准备工作，企业能够确保所依赖的数据是可信赖和有价值的，为后续的深入分析和决策提供坚实的基础。

### （2）数据治理

数据治理是确保数据质量和安全的关键。治理维度包括数据的准确性、完整性、一致性、安全性和合规性。准确性确保数据反映真实情况；完整性意味着数据没有遗漏重要信息；一致性关注数据在不同系统中的一致表现；安全性和合规性则涉及数据的保护和符合法律法规的要求。在这个过程中，还需关注数据的访问控制和权限管理，确保只有授权用户才能访问敏感数据，同时避免数据被未经授权的修改或删除。

合作方在数据治理的评估中，应考虑数据来源、收集方法、存储和处理流程。这包括对数据收集的透明度、数据存储的安全性、以及数据处理过程的合规性进行评估。评估的目的是识别和缓解可能的风险，包括数据泄露、误用和滥用的风险。这也涉及到对数据备份和恢复计划的评估，确保在数据丢失或损坏时能够迅速恢复。

合理的数据治理不仅保护企业免受法律风险，也增强客户和合作伙伴的信任。此外，有效的数据治理还有助于提高数据的价值，通过确保数据的高质量和可靠性，使数据成为支持决策的强有力工具。随着数据越来越成为企业资产的重要组成部分，强化数据治理成为了企业维护竞争优势、推动业务增长的必要条件。因此，不断优化数据治理结构和流程，建立健全的数据管理标准和策略，对于企业的长期成功和可持续发展至关重要。

### （3）保障数据安全

在使用数据时保障数据安全是企业的重要责任。这包括实施数据加密、访问控制、数据备份和恢复机制。数据加密确保数据在传输和存储时不被未经授权的访问；访问控制确保只有授权人员才能访问敏感数据；数据备份和恢复机制则在数据丢失或损坏时提供保障。此外，为了防范网络攻击和其他安全威胁，企业还需部署先进的网络安全措施，如防火墙、入侵检测系统和安全信息和事件管理（SIEM）系统。

此外，企业还需要定期进行安全审计和合规性检查，以确保数据处理活动符合相关的法律法规。这包括数据保护法规（如欧盟的GDPR）和行业标准。在这个过程中，进行风险评估和管理也至关重要，以识别潜在的安全漏洞和弱点，并采取相应的措施加以强化。同时，企业还需确保与第三方供应商和合作伙伴在数据处理方面的合作是安全和符合标准的。

教育和培训员工关于数据安全和隐私保护的最佳实践也是非常重要的。这不仅包括基本的安全意识培训，还包括针对特定角色或部门的定制化培训。员工是数据安全的第一道防线，因此提高他们对于保护数据重要性的认识和理解，是减少数据泄露和滥用风险的关键。

总的来说，保障数据安全是一个多方面的任务，涉及技术、政策和人员等多个层面。企业需要建立一个全面的数据安全策略，不断更新和改进，以应对不断变化的安全威胁和挑战。通过这样的努力，企

业不仅能保护自己的资产和声誉，也能赢得客户和市场的信任和尊重。

### 4.2.3 行业工具 API 能力层次不齐

接入汽车行业工具是群体智能技术最重要的一个环节，目前，汽车行业的技术工具仍然存在 API 开放性差，缺少文档支撑，安全性差等问题。对此，需要建立完善的 API 开放机制及使用规范，包括数据访问权限、数据标准化、API 调用频率、API 维护、文档和教程、安全协议、法律跟合规性审查。

**数据访问权限：**定义哪些数据可以被工具访问是首要任务。不是所有数据都应对所有工具开放。例如，一些敏感的客户信息可能仅对 SCRM 系统开放，而对营销工具则加以限制。

**数据标准化：**确保数据在不同系统间传输时的格式统一和标准化，可以减少数据处理错误和兼容性问题。

**API 调用频率的约束：**过于频繁的接口调用可能会对系统性能造成影响，因此需要设定合理的调用频率限制。这有助于平衡数据实时性和系统稳定性的需求。

**API 维护：**建立专门的团队或流程来管理和维护 API 接口。这包括监控接口性能，确保接口的稳定性和可用性，以及及时更新接口以适应新的业务需求或技术变化。

**文档和教程：**提供全面的文档支持和教程，包括详尽的 API 文档、

使用案例、常见问题解答和技术支持渠道。这不仅有助于用户更有效地利用 API，也能减少因误用或不理解功能而导致的问题。

**安全协议：**制定严格的数据传输和存储安全协议，以保护数据在接口调用过程中的安全。这可能包括加密传输、访问日志记录等措施。

**法律和合规性审查：**考虑法律和合规性问题，确保所有的接口调用和数据处理活动符合相关的法律法规。这可能涉及与法律顾问合作，以确保企业的数据处理和接口策略遵循行业规范和国家法律。

### 4.2.4 行业流程复杂，定制化成本高

汽车行业解决方案复杂，往往需要深入分析企业的业务流程、市场环境和客户需求，结合大模型群体智能相关技术，提供定制化解决方案，方能在竞争中脱颖而出。行业解决方案伙伴拥有深入的行业知识和市场洞察力，丰富的项目实施能力，但如何在群体智技术框架实现业务流程的优化便是最为关键的一环。

传统的机器人流程自动化技术 (Robotic Process Automation, RPA) 使得用户能够配置一种或多种“软件机器人”来模拟并集成人类与数字系统的交互来执行业务流程，可以自动地解释、触发响应、与其他系统交互，并在多种数据类型之间进行操作。尽管 RPA 技术可以帮助完成业务管线的自动化工作，但其实现需要业务专家根据业务流程人工设计复杂的工作流，并且难以处理需要动态决策的任务，这造成了 RPA 在现实中应用中的落地瓶颈。

为了解决汽车行业中业务流程复杂、人工编制工作流效率低和成本高的问题，可将 AI Agent 自动化技术（Agentic Process Automation, APA）引入到业务工作流的构建中。APA 技术在 AI Agent 的加持下，能够在更高层次上理解和处理复杂的业务流程，自动地构建工作流，并在执行过程中自适应地处理复杂决策和动态情况，这使得它在汽车行业中的应用更为灵活和高效。

汽车行业的解决方案要求在多变的市场环境中迅速做出响应，APA 技术的引入可以大幅度降低业务流程定制化的难度和成本。通过 AI Agent，企业可以实现从简单数据处理到复杂逻辑控制的全方位自动化，大幅提高项目实施的效率和质量。

## 4.3 战略伙伴：汽车行业群体智能生态伙伴与共赢演进

实现汽车行业组织孪生，需要全体生态伙伴不断进步与持续合作，生态中的每一位成员要充分发挥所长，并通过资源共享与互补、专业领域的深度合作、建立市场信息共享机制实现共赢演进。

### 4.3.1 资源共享与能力互补

**互利的资源共享：**合作伙伴通过共享自己的资源，能够相互提升自身能力。这种资源共享的方式为参与方提供了获取不可自行开发或高成本资源的机会，有效降低了进入新技术领域的门槛。同时，这种共享也意味着企业可以利用合作伙伴的专业知识和市场经验来提升自己的业务战略。减少重复投资的同时，合作伙伴间的紧密合作还可以加速新产品的研发和上市，增强企业对市场变化的响应速度和创新能

力。

**技术与能力互补：**不同合作伙伴拥有各自的技术专长和专业能力，通过合作，可以实现这些技术和能力的互补。例如，在汽车行业，数据分析公司可能在处理大数据和进行复杂算法分析方面拥有优势，而汽车制造商则拥有丰富的工程知识、设计经验和实车测试数据。通过合作，数据公司可以利用制造商的测试数据来优化自己的分析模型，从而更精准地预测市场趋势或改善产品设计。同样，汽车制造商也可以借助数据公司的分析能力来更好地理解消费者需求和市场动态，从而制定更有效的营销策略和产品改进计划。这种互补性不仅提升了双方的产品和服务质量，还加深了对各自领域的理解，从而推动整个行业的技术进步和创新发展。

### 4.3.2 专业领域的深度合作

**专业领域相互学习：**在专业领域的深度合作中，各合作伙伴可以充分利用彼此的专长和经验进行相互学习，创造出全新的技术解决方案。这种学习过程不仅限于技术层面，还包括管理策略、市场理解和客户服务等方面。例如，一家专注于软件开发的公司可以学习其在汽车行业的合作伙伴关于生产管理和质量控制的经验。同样，汽车公司也可以通过与科技公司的合作，了解最新的数字化趋势和软件开发实践。这种跨领域的知识交流不仅提高了合作伙伴的综合竞争力，还有助于员工的个人发展和职业成长。

**技术创新与研发提效：**深度合作还涉及将理论和创新应用于实际

的业务场景。例如，汽车制造商的实际工程经验和市场洞察可以为 AI 公司提供宝贵的输入，使其开发的技术更加贴合实际的车辆使用环境和消费者的实际需求。这种基于实践的技术开发可以确保新技术不仅在实验室环境中有效，而且在现实世界中也同样高效可靠。此外，这种面向实际应用的合作还有助于快速识别和解决技术开发过程中可能遇到的问题，从而缩短产品从概念到市场的时间，提高研发的效率和成果转化率。

### 4.3.3 建立市场信息的共享机制

**及时获取行业新动态：**在一个多元化且相互连接的生态系统中，合作伙伴能够迅速获取关于行业最新发展的信息，包括突破性技术、新兴市场趋势和竞争对手的动态。这种信息流的快速性对于企业保持市场敏感性和前瞻性至关重要。例如，汽车行业的合作伙伴可以通过生态系统内的网络了解到关于新能源、智能网联和自动驾驶等领域的最新研究成果和技术进展。这样的信息不仅帮助他们保持技术领先，还能够激发新的业务想法和创新策略。生态系统中的信息共享和知识流动也加速了技术的迭代和优化，帮助企业及时捕捉和应对行业变化。

**调整技术路线和发展战略：**在了解到行业的最新动态和前沿技术之后，合作伙伴可以根据这些信息来调整和优化自己的技术发展方向和业务战略，这种调整可能涉及投资新技术、改变产品开发重点或重新定位市场策略。通过生态系统提供的深度洞察和广泛视野，企业能够更有效地进行长期规划，确保其技术和产品能够持续符合或领先市场需求。

## 第五章

# 总结展望

技术的飞速发展，辅助我们站在一个全新的视角，眺望汽车行业即将踏入的辽阔天地。在智能化的浪潮下，我们预见一个更加智慧、高效能、用户至上的汽车新时代正在加速到来。

- **智能化助力**：汽车企业突破降本增效天花板

汽车行业的群体智能和组织孪生技术的核心价值，在于为汽车行业带来了前所未有的降本增效可能性。在当前的经济环境下，车企需要不断检索突破口来提高生产效率、降低运营成本。通过使用群体智能和组织孪生技术，车企可以率先将明确 SOP 和专家知识的场景实现智能化与自动化落地应用，重塑效率之巅。这不仅有助于车企提升自身的竞争力，更能推动整个汽车行业的持续发展。

- **智能化赋能**：开启用户运营新篇章

在以往用户运营旅程中，与日俱增的纷繁触媒环境下投入大量人力和财力成本也难以精准捕捉用户多样化需求。群体智能不仅将极大地提高信息传递和决策的效率，更通过对海量用户数据的深度挖掘和分析，令车企能够为用户提供更加贴心、个性化的产品和服务，从而构建起更加紧密的用户关系，提升品牌影响力和市场竞争力。

- **创新与合作**：共建智慧汽车新生态

随着技术的持续演化和应用场景的拓展，我们可以预见大语言模型驱动的群体智能和组织孪生技术，将在汽车行业得到更广泛的应用与深度融合，释放出巨大的数据价值，显著增强车企在不确定环境下的竞争力和韧性。这一前景的实现离不开与战略合作伙伴及智能生态合作伙伴的紧密合作与创新实践。在此过程中，易慧智能、清华大学自然语言处理实验室以及面壁智能将携手深化合作，共同推动大模型、AI Agent、群体智能、组织孪生等技术的创新与融合，共同描绘出一个智能、高效、充满竞争力的汽车行业蓝图。

# 参考文献

[1-1] 易车研究院.车市价格战洞察报告（2023版）[E].2023年11月13日

[1-2] 易车研究院.家庭拥车数量洞察报告（2023版）[E].2023年8月25日

[1-3] 麦肯锡.2023年麦肯锡中国汽车消费者洞察报告[E].2022年12月

[1-4] 群邑，易车. 2023年全域链路时代汽车营销变革白皮书[E]2023年7月

[2-1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[2-2] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

[2-3] Ding N, Qin Y, Yang G, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models[J]. arXiv preprint arXiv:2203.06904, 2022.

[2-4] Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[2-5] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[2-6] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020, 33: 1877-1901.

[2-7] OpenAI. Gpt-4 technical report. 2023, https://cdn.openai.com/ papers/gpt-4.pdf.

[2-8] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

[2-9] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.

[2-10] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer[J]. Journal of Machine Learning Research, 2020, 21: 1-67.

[2-11] Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]//Proceedings of ACL. 2020: 7871-7880.

[2-12] Zhang Z, Gu Y, Han X, et al. Cpm-2: Large-scale cost-effective pre-trained language models[J]. AI Open, 2021, 2: 216-224.

[2-13] Sun Y, Dong L, Huang S, et al. Retentive Network: A Successor to Transformer for Large Language Models[J]. arXiv preprint arXiv:2307.08621, 2023.

[2-14] Dao T, Fu D, Ermon S, et al. Flashattention: Fast and memory-efficient exact attention with io-awareness[J]. Advances in Neural Information Processing Systems, 2022, 35: 16344-16359.

[2-15] Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. arXiv preprint arXiv:2101.03961, 2021.

[2-16] Google. Introducing Pathways: A next-generation AI architecture. https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/.

[2-17] Zhang Z, Lin Y, Liu Z, et al. Moefication: Transformer feed-forward layers are mixtures of experts[J]. arXiv preprint arXiv:2110.01786, 2021.

[2-18] He J, Qiu J, Zeng A, et al. Fastmoe: A fast mixture-of-expert training system[J]. arXiv preprint arXiv:2103.13262, 2021.

[2-19] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[2-20] Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.

[2-21] Wang Y, Mishra S, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks[C]// Proceedings of the EMNLP 2022: 5085-5109.

[2-22] Iyer S, Lin X V, Pasunuru R, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization[J]. arXiv preprint arXiv:2212.12017, 2022.

[2-23] Honovich O, Scialom T, Levy O, et al. Unnatural instructions: Tuning language models with (almost) no human labor[J]. arXiv preprint arXiv:2212.09689, 2022.

[2-24] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

[2-25] Ding N, Hu S, Zhao W, et al. OpenPrompt: An Open-source Framework for Prompt-learning[C]//Proceedings of the ACL: System Demonstrations. 2022: 105-113.

[2-26] Hu S, Ding N, Zhao W, et al. OpenDelta: A Plug-and-play Library for Parameter-efficient Adaptation of Pre-trained Models[J]. arXiv preprint arXiv:2307.03084, 2023.

[2-27] Pfeiffer J, Rücklé A, Poth C, et al. Adapterhub: A framework for adapting transformers[C]// Proceedings of the EMNLP. 2020: 46-54.

[2-28] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018.

[2-29] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[2-30] Han X, Zhang Z, Ding N, et al. Pre-trained models: Past, present and future[J]. AI Open, 2021, 2: 225-250.

[2-31] Nakano R, Hilton J, Balaji S, et al. Webgpt: Browser-assisted question-answering with human feedback[J]. arXiv preprint arXiv:2112.09332, 2021.

[2-32] Yao S, Chen H, Yang J, et al. Webshop: Towards scalable real-world web interaction with grounded language agents[J]. Advances in Neural Information Processing Systems, 2022, 35: 20744-20757.

[2-33] OpenAI. ChatGPT Plugins, 2021. URL: https://openai.com/ blog/chatgpt-plugins.

[2-34] Mialon G, Dessì R, Lomeli M, et al. Augmented language models: a survey[J]. arXiv preprint arXiv:2302.07842, 2023.

[2-35] Qin Y, Hu S, Lin Y, et al. Tool learning with foundation models[J]. arXiv preprint arXiv:2304.08354, 2023.

[2-36] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents[J]. arXiv preprint arXiv:2308.11432, 2023.

[2-37] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey[J]. arXiv preprint arXiv:2309.07864, 2023.

[2-38] Tsinghua University & ModelBest. XAgent: An autonomous LLM agent for complex task solving.https://blog.x-agent.net/.

[2-39] Chen W, Su Y, Zuo J, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents[J]. arXiv preprint arXiv:2308.10848, 2023.

[2-40] Ye Y, Cong X, Tian S, et al. ProAgent: From Robotic Process Automation to Agentic Process Automation[J]. arXiv preprint arXiv:2311.10751, 2023.

[2-41] Qian C, Cong X, Yang C, et al. Communicative agents for software development[J]. arXiv preprint arXiv:2307.07924, 2023.

[2-42] Talebirad Y, Nadiri A. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents[J]. arXiv preprint arXiv:2306.03314, 2023.

[2-43] Li G, Hammoud H A A K, Itani H, et al. Camel: Communicative agents for" mind" exploration of large scale language model society[J]. arXiv preprint arXiv:2303.17760, 2023.

[2-44] Hong S, Zheng X, Chen J, et al. Metagpt: Meta programming for multi-agent collaborative framework[J]. arXiv preprint arXiv:2308.00352, 2023.

[2-45] Park J S, O'Brien J, Cai C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023: 1-22.

[2-46] Zhou X, Li G, Sun Z, Liu Z, Chen W, Wu J, Liu J, Feng R, Zeng G. D-Bot: Database Diagnosis System using Large Language Models[J]. arXiv preprint arXiv:2312.01454. 2023.

[2-47] Lilian Wang. LLM Powered Autonomous Agents. https://lilian weng.github.io /posts/ 2023-06-23-agent/

[2-48] 中国人工智能协会. 中国人工智能系列白皮书——大模型技术（2023版）

[2-49] Händler T. Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous LLM-powered multi-agent architectures[J]. arXiv preprint arXiv:2310.03659, 2023.

[2-50] Zhou W, Jiang Y E, Li L, et al. Agents: An open-source framework for autonomous language agents[J]. arXiv preprint arXiv:2309.07870, 2023.

[2-51] Wu Q, Bansal G, Zhang J, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation framework[J]. arXiv preprint arXiv:2308.08155, 2023.

[2-52] Liu X, Yu H, Zhang H, et al. Agentbench: Evaluating llms as agents[J]. arXiv preprint arXiv:2308.03688, 2023.

[2-53] CLUE中文语言理解评测基准团队.SuperCLUE-Agent: Agent智能体中文原生任务能力测评基准[EB/OL]. https://www.cluebenchmarks.com/ superclue _agent.html

[2-54] Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models[J]. Transactions on Machine Learning Research, 2022.

[2-55] Zhao J, Zhang Z, Ma Y, et al. Unveiling A Core Linguistic Region in Large Language Models[J]. arXiv preprint arXiv:2310.14928, 2023.

[2-56] 大数据协同安全技术国家工程研究中心.大语言模型提示注入攻击安全 风险分析报告[R].2023

# 白皮书发行单位介绍

## 清华大学自然语言处理实验室简介

清华大学自然语言处理实验室，是国内开展 NLP 研究最早、深具影响力的科研单位之一，也是国内开展大模型研究最早的团队。实验室目前在职教师 4 人、博士后 5 人、硕博研究生 30 人。团队在孙茂松（清华大学人工智能研究院常务副院长、欧洲科学院外籍院士 /ACL FELLOW）、刘洋（国家杰青）、刘知远（万人计划青年拔尖人才）等老师带领下，长期深耕自然语言处理的核心技术研究，涉及大模型、知识图谱、机器翻译、AI Agent 等，在论文发表和技术开源上取得了系列有影响力的成果。

团队近三年在自然语言处理和人工智能领域的高水平国际会议和期刊上发表相关论文 200 余篇，其中包括 ACL、EMNLP、NAACL、NeurIPS、ICLR、AAAI 等会议，以及 TACL、IEEE/ACM TASLP、IEEE TKDE 等期刊，累计获得国际主流会议的最佳论文或提名十余次，论文谷歌学术引用累计超过 8 万次。获国家发明专利授权 50 余项。

团队开源发布多个有学术与业界影响力的大模型：发布全球首个知识指导大模型 ERNIE、国内首个中文大模型 CPM-1、百亿参数大模型 CPM-Bee（登顶少样本评测排行榜 ZeroCLUE）、中英双语多模态大模型 VisCPM、生物医学领域大模型 KV-PLM（入选 Nature Communications 亮点文章）、指令对齐大模型 UltraLM（登顶斯坦

福 Alpha-Eval 开源模型榜首）。研发了面向大模型训练、微调、压缩、评测、对齐、智能体等技术工具套件，其中获 ACL2022 最佳展示论文奖、入选 Nature Machine Intelligence 封面论文等，指令微调数据集 UltraChat、UltraFeedback 成为国内外开源社区大模型应用的重要基础设施之一，以及 ToolLLM、ChatDev、AgentVerse、XAgent、D-Bot 等智能体开源工具成为领域的代表性工具。

## 易慧智能简介

北京易慧涌现智能科技有限公司（简称“易慧智能”）是业界领先的 AI 产品和业务解决方案提供商。公司致力于将尖端的 AI 科技与最佳的业务实践相结合，缔造全球首个汽车行业大模型驱动的群体智能协同工作平台，提供精研的组织孪生解决方案，和卓越的 AI 员工运营服务。全面助力企业在智能化时代实现跨越式升级，共创更加璀璨的商业前景。

凭借汽车行业互联网营销解决方案 20 余年的深厚积累，已累计服务 300 余家汽车品牌和覆盖超两万家汽车经销商的营销经验，拥有汽车行业千亿级别用户行为数据和行业覆盖度第一的综合知识库，共同支持客户智能化的最后一公里。

## 面壁智能简介

北京面壁智能科技有限责任公司（简称“面壁智能”），成立于 2022 年 8 月，是一家源自清华大学自然语言处理实验室（THUNLP）

的人工智能科技公司，深耕通用 AI 领域，专注大模型技术的创新与应用转化。公司拥有人工智能领域享有盛誉的清华系研发创始团队，依托在自然语言处理方面的多项世界级前沿技术，正在构建大规模预训练模型库及配套工具，旨在推进大模型技术与应用标准化。

秉承“智周万物”的企业愿景和理念，面壁智能致力于让 AI 技术安全普惠地服务人类美好生活，为 AGI 世界的到来打下坚实基础。公司自主研发了 CPM 系列大语言模型，其中包括国内首个中文大模型 CPM-1、国内首个开源免费商用基座模型 CPM-Bee、千亿多模态大模型 CPM-C 及多个商用行业大模型。

基于 CPM 系列基座模型的强大能力，面壁智能已成功助力多个行业实现智能升级与业务提效，并已正式公开发布千亿多模态大模型对话助手“面壁露卡 Luca”。面壁 Luca 不仅在中英文语言对话方面表现卓著，还具备极强的代码、知识、逻辑及图片理解等能力。基于持续的科技前沿探索与多年行业积累，面壁智能也是最早在 AI Agent 技术取得突破的 AI 大模型科技公司之一，目前已推出由大模型驱动的 AI Agent“三驾马车”创新成果，包括：大模型驱动的智能体通用平台 AgentVerse、超强 AI 智能体应用框架 XAgent、多智能体协作开发框架 ChatDev。目前，ChatDev 已通过 SaaS 平台形式面向软件开发行业开发者和创业者开放服务。

我们相信「大模型 +Agent」将会引起新一轮的应用爆发，为行业和用户带来更多新的能力与服务，推动 AI 大模型的场景落地。

面壁智能致力于基于核心自主研发的大模型底座，构建一套旨在全面提升人类智能的架构。依托自研的强大单体智能和群体智能协作的前沿技术探索，打造标准、易用的智能体化（Agentization）协作平台 AgentVerse，该平台不仅服务于开发者，也致力于为终端用户带来革命性的生产与作业方式，从根本上改变千行百业和人们的日常生活。平台面向所有用户全面开放、拥抱生态，期待更多产业、用户齐聚于此，共同迈入 AGI 未来。

此外，面壁智能还联合清华大学 NLP 实验室、OpenBMB 开源社区打造完成“一体两翼”的大模型“产学研用”开放式平台生态布局，已为数百家企业提供商用服务，覆盖金融、汽车、商业、工业、医疗、教育、法律、媒体等多行业和领域。我们正在积极践行和推动 AI 大模型落地千行百业，协同全行业伙伴共建、共赴 AGI 未来。

清华大学自然语言处理实验室

易慧智能

面壁智能