智谱·AI | ChatGLM

开启AI
智能化新纪元

免责申明：

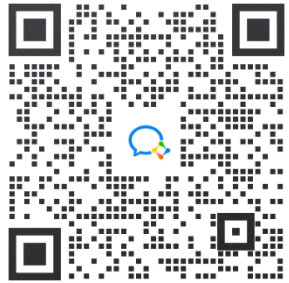
本内容非原报告内容；

报告来源互联网公开数据；如侵权请联系客服微信，第一时间清理；

报告仅限社群个人学习，如需它用请联系版权方；

如有其他疑问请联系微信。

行业报告资源群



微信扫码 长期有效

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6+份行研精选、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息

微信扫码 行研无忧

知识星球 行业与管理资源

专业知识社群：每月分享8000+份行业研究报告、商业计划、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等；已成为投资、产业研究、企业运营、价值传播等工作助手。

引言

在人工智能的漫长征程中，大模型的出现无疑是一座闪耀的里程碑。2023年，在全球科技领域，大模型无疑成为了最炙手可热的话题，这股热潮由美国创业公司 OpenAI 率先掀起，中国科技公司紧随其后，纷纷投入到这场大模型的竞争中。如果说深度学习技术的突破让计算机拥有了“看”和“听”的能力，那么大模型则让计算机具备了“理解”和“创造”的潜能，重新定义公众对 AI 的认识，为各行各业的智能化转型开辟了一条新的道路。

开发“大一统”，让 AI 真正普惠与落地

在大模型出现之前，人工智能从未如 ChatGPT 这般普及，超过 1 亿用户主动体验的背后，是业务发展需求驱动 AI 应用场景探索与实践的重大转变。在过往企业对于 AI 的应用中， 较高的开发门槛、应用场景的复杂性与多样性，以及对大量标注数据的依赖，构成了 AI 大规模部署的难题。而预训练大模型则具有良好的通用性、泛化性，可以显著降低人工智能应用门槛，用户基于大模型通过零样本、小样本学习即可获得领先的效果；同时，“预训练+精调”等开发范式，让研发过程更加标准化，显著降低了人工智能应用门槛，成为 AI 走向工程化应用落地的重要手段。

走向产业落地，助力组织加速发展新质生产力

统一数据、统一算法、统一模型，解决所有问题，这将是和过去任何一次 AI 风潮都完全不同的新时代，AI 将变成社会的基础生产要素。大模型的「海平面」正在逐渐没过人类能力的「山头」，过去被认为只有人类才能完成的事情开始逐步被大模型的洪水所淹没。而大模型作为新质生产力的代表，正在推动传统产业的转型升级和新兴产业的快速发展，为社会经济的高质量发展注入新的动力。

“时间永远分岔，通向无数的未来”，无需短期高估技术的影响，也不要长期低估技术的魅力。在企业智能革命的舞台上，我们一次又一次被大模型的能力惊艳，也必将看到越来越多的新场景、新应用随之产生，真正拥抱“AI Native 的到来”。

CONTENTS

目录

大模型

智谱·AI

在组织数字化中的应用

从 2023 年初迄今，以 AIGC为主角的跨年度大戏高潮迭起：从ChatGPT引爆人工智能通用化的话题，到大模型形成百舸争流的局面。如果说，各大厂商纷纷推出大模型产品并形成“百模大战”的局势，是大模型这场“战役”的上半场，那么这场“战役”的下半场将聚焦在大模型的垂直化应用以及价值转化发展。

大模型应用场景概述

文本生成

文本生成指的是通过指令（Prompt）让大模型自动生成文字，包括电子邮件、短信、文章、新闻报道、社交媒体帖子等各种文本内容。相较传统以规则和模板的方式，大模型提供了完全不同的体验，这也是大模型最先跑通的商业模式。

- 新闻/小说生成
- 广告文案生成
- 会议纪要生成
- 数据报告生成
- 直播脚本生成
- 周报/邮件生成

信息抽取

信息抽取是指将长段文字中的信息抽取出来并且以结构化的方式输出。相比起传统 NLP 的方式，大模型在泛化能力上有非常大的提升，并且开发成本要低 2 个数量级。

- 用户需求提取
- 用户画像提取
- 舆情分析
- 文章阅读辅助
- 助贷数据清洗
- 销售质检

信息检索

传统的信息检索系统只能以文字来匹配正文，并且只能以原文片段返回，或者对于垂直场景只能是结构性的卡片，而大模型则可以为你通读结果并根据你的查询生成针对性的回答，带来全新的搜索体验。

- 知识搜索
- 视频搜索
- 文档检索
- 商品搜索
- 简历检索
- 房产检索

智能对话

对话系统是指机器和用户进行对话的系统，通常用于客服和助手类的场景，但原有客服都基于问答对或者规则来进行对话，难以达到真人的水平，而大模型则能在上下文理解和回答生成上带来全新的体验。

- 智能客服
- 语音助手
- 游戏NPC
- 虚拟社交
- 虚拟导购
- 智能陪练

指令代码生成

自动生成代码，提高开发效率，减少人工编写代码的工作量。自动分析已有的代码并提供重构和优化的建议，减少人工编写测试代码的工作量。同时，大模型可根据用户提供的部署描述自动生成部署脚本，并监控应用程序的性能。

- NL2 SQL
- AI建站
- 智能RPA
- 代码生成
- 测试用例生成
- 代码审查

其他

- 语言翻译/优化
- 复杂指令识别
- 文章扩写/缩写
- 作文评分/润色
- 意图洞察
- PPT生成
- 解数学题
- 车载助手
- 合同审查
- 作文批改

... ...

大模型行业应用分享

大模型 + 智能汽车

智谱·AI

汽车行业一直在电动化、智能化领域持续不断地发力，截至 2023 年 9 月，新能源车国内零售渗透率已经达到了 36.9%。而从汽车研发到生产、制造，再到营销和服务的各个阶段，都能与 AI 技术有广泛的结合，大模型与汽车行业的结合可能作为支撑整个行业的基础设施，通过与现有的基础架构深度融合，从而催化对整个产业和行业的深层次变革。

智能座舱

- 复杂车控
- 车书问答
- 闲聊陪伴
- 生活服务

市场营销

- 客户标签提取
- 销售话术质检
- 营销话术辅助
- 销售话术训练
- 车友社区运营
- 战败归因分析

售后服务

- 智能客服
- 智能工单
- 舆情分析

生产制造

- 软件开发
- 产线生产
- 质量检验

典型场景分享

大模型赋能复杂车控：拓展服务边界，提升用户体验

随着智能汽车的发展，汽车座舱已逐步演变为人们的第三生活空间，集舒适、娱乐、工作于一体，实现人车生活的深度交融。大模型的加入，则可以进一步加速人与车交互方式的转变，例如从单一的任务方式逐步转变成基于多任务的应用场景，从现有的单一语音交互向多模态的交互方式进行转变，为车主带来更为便捷、实用的驾驶体验。

传统车控方案的现状和痛点

语义泛化能力弱，意图识别准确率低

依赖大量人工标注，维护成本高

不支持单句多意图，用户体验差

大模型的业务价值

更聪明：指令理解更精准，单句多意图无遗漏

更好玩：交互趣味性更强，改善用户体验

更简单：人工维护工作量更低

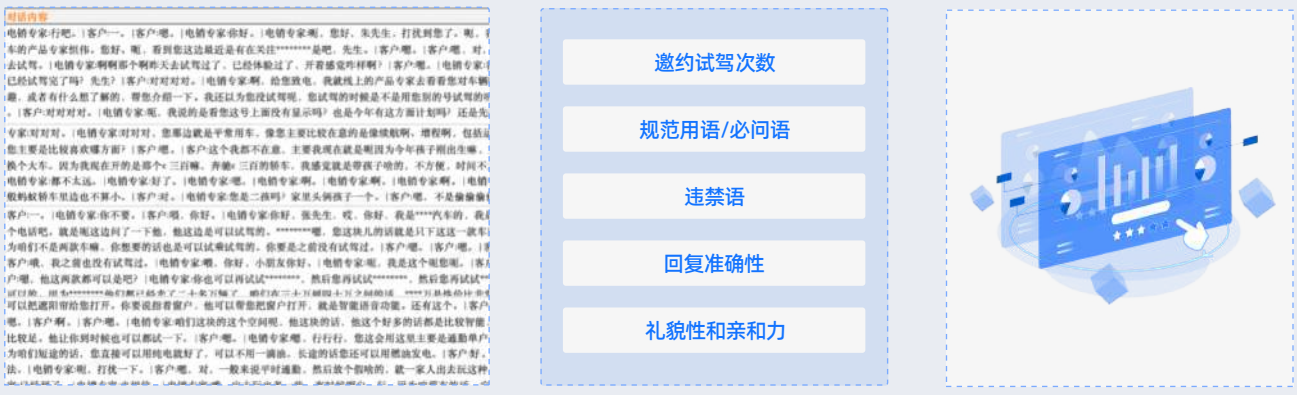
更有用：帮助处理简单任务，提供更有帮助的知识

工具属性 → 大模型 → 体验属性

分类	场景说明	示例
简单意图	简单明确的意图解析 传统小模型可以支持	“打开右后车窗”、“空调调到20度”、 “播放周杰伦的稻香”、“导航回家”
条件意图	满足A条件执行B动作 传统小模型部分支持	“10分钟后关闭空调”、 “到家后提醒我买菜”
并列多意图	一句话中包含多个意图 传统小模型不支持	“打开车窗，关闭空调。然后过5分钟关闭车窗，空调调到23度” “播放周杰伦的稻香，单曲循环，音量调到30%”

销售话术质检：高转化率、精细管理、持续改进的营销闭环

随着客户量的增加、沟通轮次和渠道的增多、汽车产品复杂度的提高，车企销售面临的挑战越来越多，难度越来越高；而对于管理者来说，也难以实现对销售过程的全方位管控和精细化管理、找到能提升转化率的症结。车企通过将大模型应用到销售话术场景，通过大模型自动分析对话中的关键信息点，实时抓取并分析销售人员与客户之间的对话记录，从专业性、合规性、服务态度等多个维度进行评估。对于汽车性能介绍是否准确详尽、是否严格按照官方信息及行业规范进行解说、是否存在过度承诺或误导消费者的情况，大模型都能做到精准识别；同时，还能对销售人员的情感表达、沟通技巧等方面进行质检，确保其服务质量达到高标准。不仅有助于提升汽车销售团队的整体业务素质，保障消费者的合法权益，也能为企业优化销售策略、提高客户满意度提供有力的数据支持，进而推动整个汽车行业的服务质量和品牌形象的持续升级。



大模型 + 大消费

智谱·AI

大消费产业的复杂营销和销售环节为大模型提供了广阔的应用空间，如何把大模型应用在品牌营销、用户运营过程中，也是众多消费科技企业正在不断思考的话题。企业希望通过大模型，率先将商业经验转化为行业的专有数据并进行有效管理，从而构建起企业的竞争优势；同时，通过大模型更智能的交互能力，在消费者旅程日益碎片化的今天，能帮助企业找到与品牌和消费者习惯相符合的交互点，不断构建品牌专业性保持高速增长。

服务对象	用户	技术人员	运营人员	市场人员	...
场景应用	产品研发	市场营销	产品销售	售后服务	企业经营
	用户分析	营销文案撰写	销售话术质检	智能客服	企业知识库
	市场趋势分析	营销素材生成	销售话术辅助	坐席助手	智能办公
	商品汰换	文稿优化	智能导购机器人	舆情分析	数据分析

典型场景分享

智能导购机器人：重构用户交互、购买决策模式

如今，各大购物App的商品推荐已成为引导消费者购买的重要渠道之一，而大模型的加入将重构与消费者交互的方式，使得更多样化、更自然、更精准的导购推荐模式成为可能，即基于消费者的历史购物记录、浏览行为和搜索关键词等数据，并通过与消费者进行智能互动，分析消费者的购物习惯、喜好和需求，解答其关于商品的疑问，提供专业的购买建议。

多条件筛选产品(快捷购物)

购物建议-推荐产品-多轮交流 (精挑细选)

舆情分析：实时聆听消费者之声，为企业战略决策提供有力支持

零售行业面临激烈的市场竞争，实时准确地了解客户之声对于树立良好的品牌形象和声誉至关重要。企业通过大模型实时抓取和解析海量网络文本数据（如社交媒体、电商评价、新闻报道等），能够全面而深入地汇总分析消费者情感倾向、需求变化、产品反馈等信息，迅速把握市场趋势，精准预测消费热点，并用以适时调整产品策略与营销方案，有效预防和应对负面舆情，保障品牌形象，助力企业在竞争激烈的市场环境中实现智慧化运营与精细化管理。

信息准

无需设置监测关键词，输入系统用户关注的信息即可进行舆情监测，实现内容智能化

研判准

100%理解检测到的负面内容，让自动化推送达到专家级别的研判水平

更实时

自动形成热点聚类，总结每日舆情要点

大模型 + 智能制造

智谱·AI

过去几年，国内工业制造领域经历了智能制造与 AI 1.0 阶段的洗礼，不少企业已对AI应用有了相当程度的认知，并完成了部分场景的智能化升级。而 AI 大模型的出现，将会融入工业企业的研发设计、生产工艺、质量管理、运营控制、营销服务、组织协同和经营管理等方方面面，极大加速各领域的智能化升级进程。

生产制造
- 工业知识问答
- 设备维修SOP生成
- 工业代码生成

客户服务
- 智能客服
- 智能工单
- 客服质检

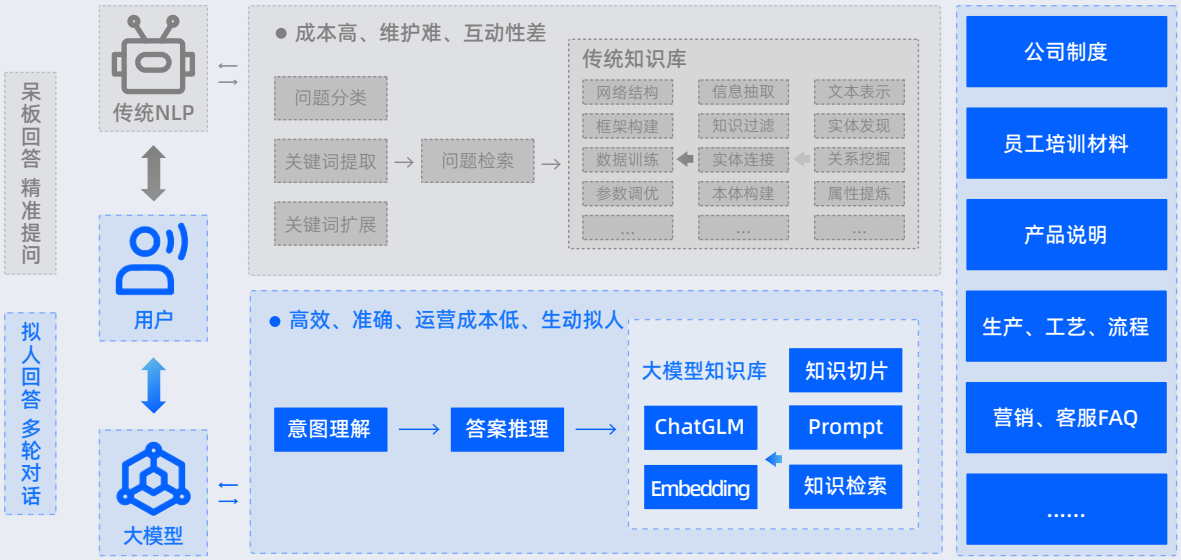
智能产品
- 智能控制
- 智能问答
- 闲聊陪伴

经营管理
- 文档处理
- 内部制度问答
- 对话式数据查询

典型场景分享

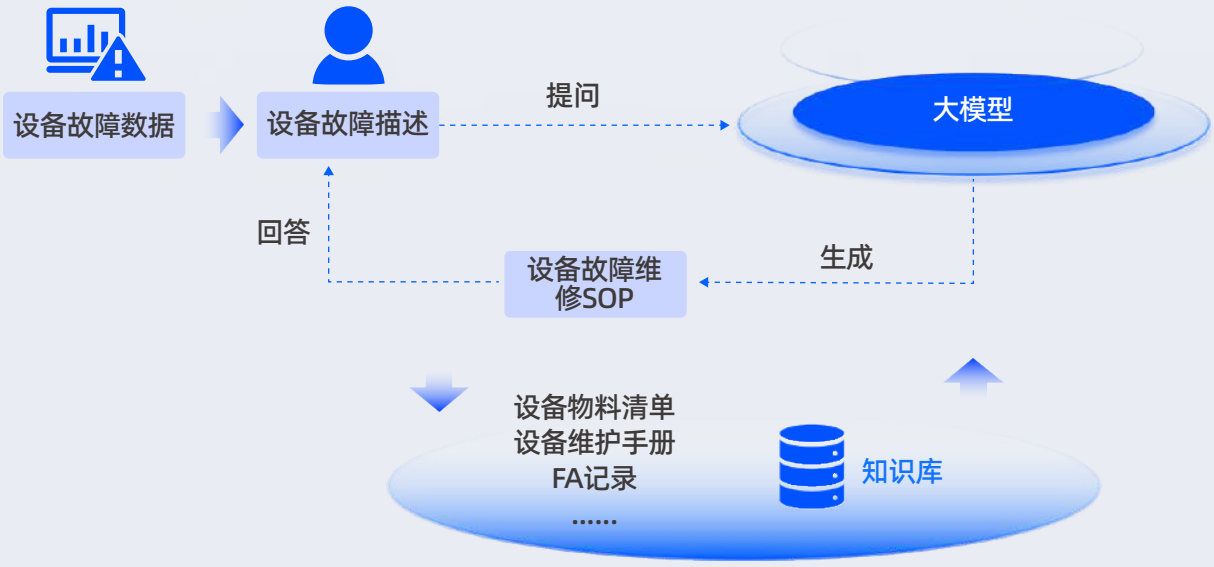
知识检索：更有效地将经验数据转化为可用的知识资产

专业性是工业制造的关键要求之一，而中国的工业知识和工业数据在很多企业都面临着因人才流转而遗失，难以转化为企业知识资产的难点和痛点，而大模型的出现可以更有效地解决这一问题，并通过企业知识库的训练微调，最大化提高准确率、避免幻觉问题。



设备维修SOP生成：提升故障处理的效率和准确性

制造业设备故障检测对于处理时效性、人员经验性有着极高的要求。而大模型可以通过对行业知识、企业知识的理解，并结合设备的历史数据、运行参数以及周围环境信息进行综合分析等，建立起一套完整的设备健康管理体系。以便工作人员能够迅速分析故障数据，精确识别问题根源，高效制定修复策略，提升故障处理的效率和准确性。



结合行业知识、企业知识等，工程师能够迅速分析故障数据，精确识别问题根源，高效制定修复策略，提升故障处理的效率和准确性。

大模型应用优势

- 基于向量模型和知识库，对相关的知识和信息进行向量化存储，以便于后续的SOP推荐和生成
- 通过历史维修记录分析，构建设备事件知识库，对比设备状态及事件属性，进行故障归因诊断，并推荐维修方案
- 利用大语言模型生成能力，将设备知识库及故障信息通过Prompt方式生成新的设备维修SOP
- 通过对FA记录的更新，重组Prompt并运用大模型进行SOP的自迭代，不断完善其在实际工作中的适用性和有效性

大模型 + 医疗健康

智谱·AI

大模型与医疗行业有着天然的契合性。医疗领域存在大量模态种类丰富的数据，且呈现出多学科、跨领域的特点，而大模型的长项之一就是对多类数据进行整合总结、分析判断和自动摘要。而根据发布在《急诊医学年鉴》、BMJ等期刊的研究，医疗大模型在部分测试中比肩甚至超越了医生，在保证医疗服务准确率与公平性、提升医疗系统工作效率等方面展现出应用优势与价值。

典型场景分享

医生诊断助手：提高问诊效率，大幅释放就诊时间

在医生诊断助手的场景中，医生需要花费大量的时间和精力去分析各种检查结果，如血液检测、影像学检查等，而传统的临床文本分析通常依赖于规则引擎或浅层机器学习方法，对复杂的医学文本难以处理。大模型的加入使得医生可以直接将检测报告交给大模型进行解读，自身则更加聚焦于与患者的深度沟通，由大模型帮助医生更快速地做出诊断和制定治疗方案，提高诊断的效率和准确性，也提高了医疗服务的质量和效率。

传统辅助诊断痛点

效果不佳：目前CDSS系统，数据少、标准化低，标注成本高，辅助诊断效果不佳，多为科普或参考。

诊断者差异：医疗资源分布不均，专业程度参差，存在漏诊误诊现象；多疾病交叉，存在经验不足、考虑不周全现象。

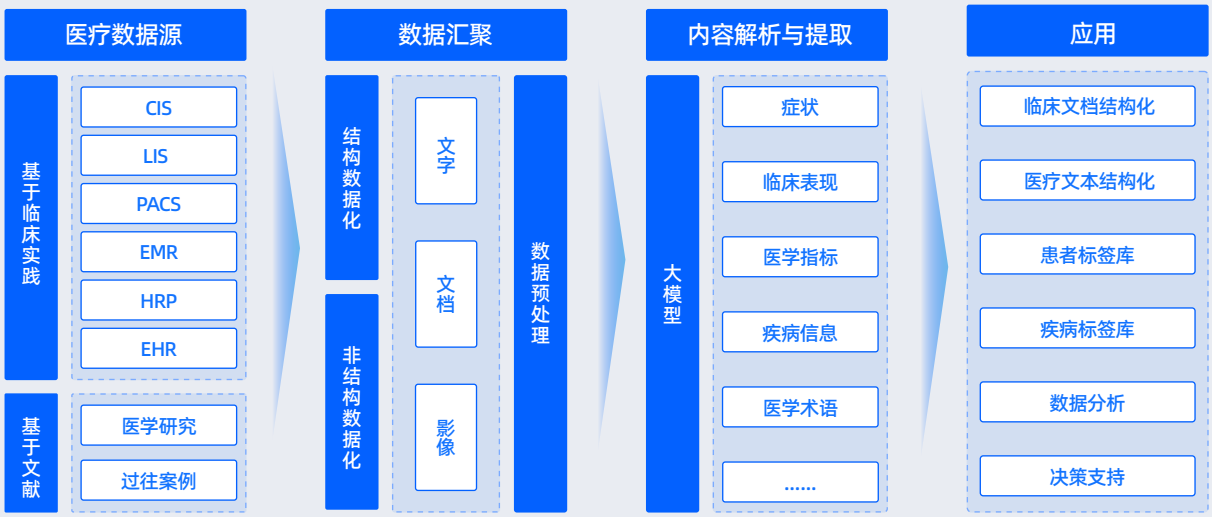
大模型的业务价值

高准确性判断：具备复杂病例识别能力，提供高准确性建议，降低误诊、漏诊风险。

全面性分析：根据患者的病症和病史，预测每种可能疾病的概率，并将它们从高到低排序。给医生全面建议，避免遗漏。

医疗报告研究助手：释放更多精力，聚焦于数据的分析及应用

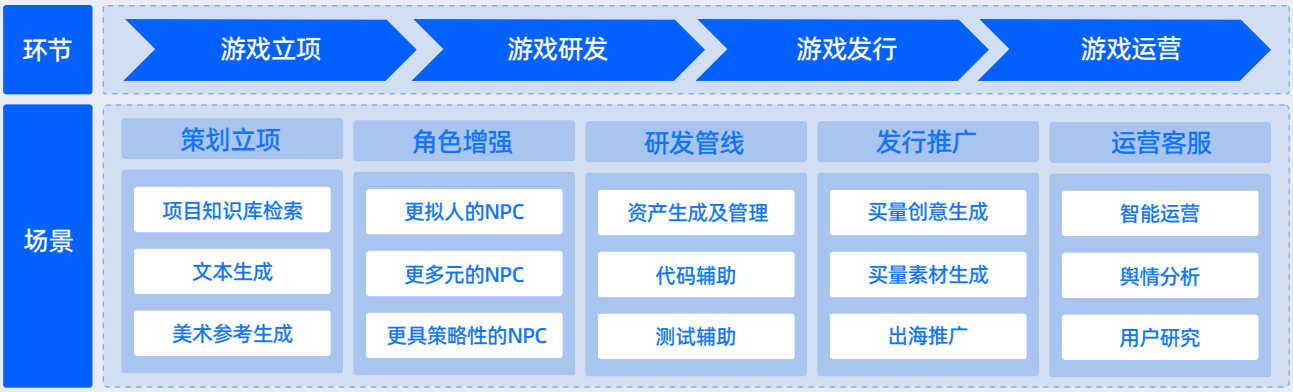
在医疗行业中，大模型正逐步成为医生和科研人员不可或缺的研究助手，尤其在处理和解析复杂的医疗报告方面展现出巨大潜力。通过大模型深度学习海量医学文献、病历报告和诊断数据，精确理解并提炼其中的关键信息，让医护人员及科研人员更聚焦于数据的分析及应用，可以极大提升医疗行业的工作效率和决策准确性，为个性化医疗和精准诊疗提供了强有力的技术支撑。



大模型 + 游戏娱乐

智谱·AI

1994 年中国迎来了第一本游戏期刊——《电子游戏软件》，被视为中国游戏产业的开端。2024 年中国游戏产业进入而立之年，从最初的热爱驱动到如今的科技赋能，中国游戏产业也在逐步走向成熟。对于游戏娱乐行业而言，“质量、成本、速度”是普遍公认的三大基本要求，而大模型的创作生成能力则可以帮助游戏研发、发行商等多角色机构兼顾三者变成现实，高效、低成本、快速地进行新文娱产品的开发、发行、运营。



游戏

01 内容生成：辅助生成剧情、任务、场景文本

02 玩家互动：构建更真实和自然的 NPC，增强游戏体验

03 AI 增强：狼人杀、剧本杀等对话类游戏拥有更多策略空间

04 玩家运营：精细化私域运营，全面提升玩家忠诚度与营收水平

影视/短视频/直播

01 内容创作：使用大模型辅助生成剧本、角色、故事，作为创作工具

02 用户互动：利用直播中的机器人主播，实时响应用户打赏、留言、评论

03 媒资查找：大模型超强的理解能力和网络查询能力，可以精准匹配媒资

小说/媒体

01 写作助手：大模型可以提供创作建议，帮助生成场景、角色描述等

02 内容生成：自动化生成定制化的新闻剪辑、摘要、标题等

03 读者互动：进行书籍推荐、内容检索、解答问题等

社交

01 虚拟好友：用于用户交互，提供个性化情感陪伴和生活工作问题解答

02 社区运营：通过大模型分析社区文本，提升互动安全性与活跃度

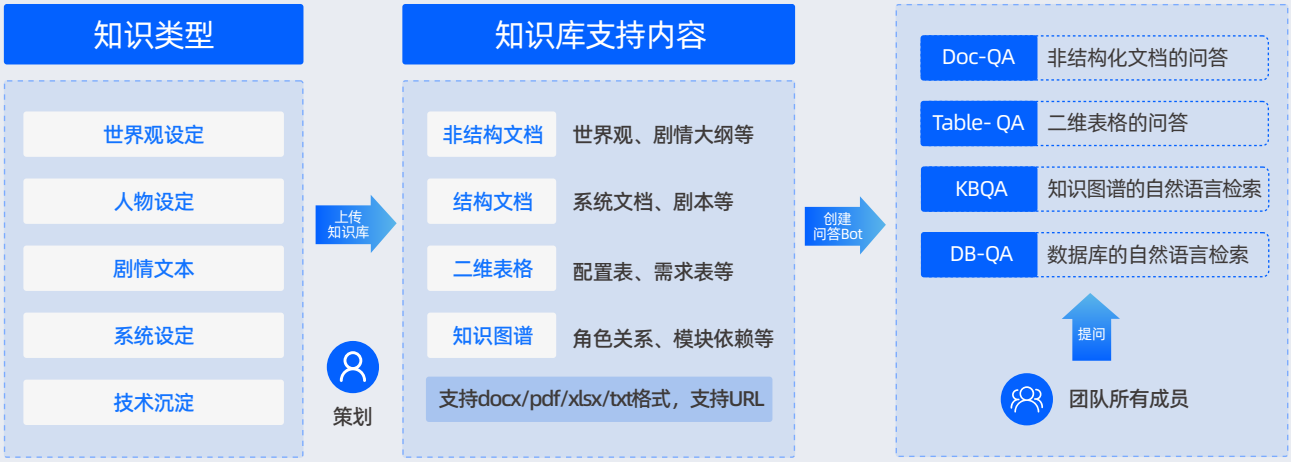
典型场景分享

超拟人 NPC：千人千面，打造自由探索的全新体验

在生成式AI的加持下，游戏可能不再会有千篇一律的枯燥剧情，而是产生更多千人千面、自由探索的开放玩法，大模型正革新着非玩家角色（NPC）的设计与互动体验。大模型可以根据游戏角色的身份背景、情感状态及剧情发展动态生成相应的对话脚本，确保 NPC 的语言表达既符合情境逻辑，又充满生动性和独特性，并通过对玩家行为和选择的即时响应，调整 NPC 的对话策略，实现真正意义上的动态交互。这不仅提高了玩家的代入感，也拓展了游戏的可玩性和重玩价值。

项目知识库检索：高效准确拉齐信息，精细化追求与现实挑战矛盾的化解新思路

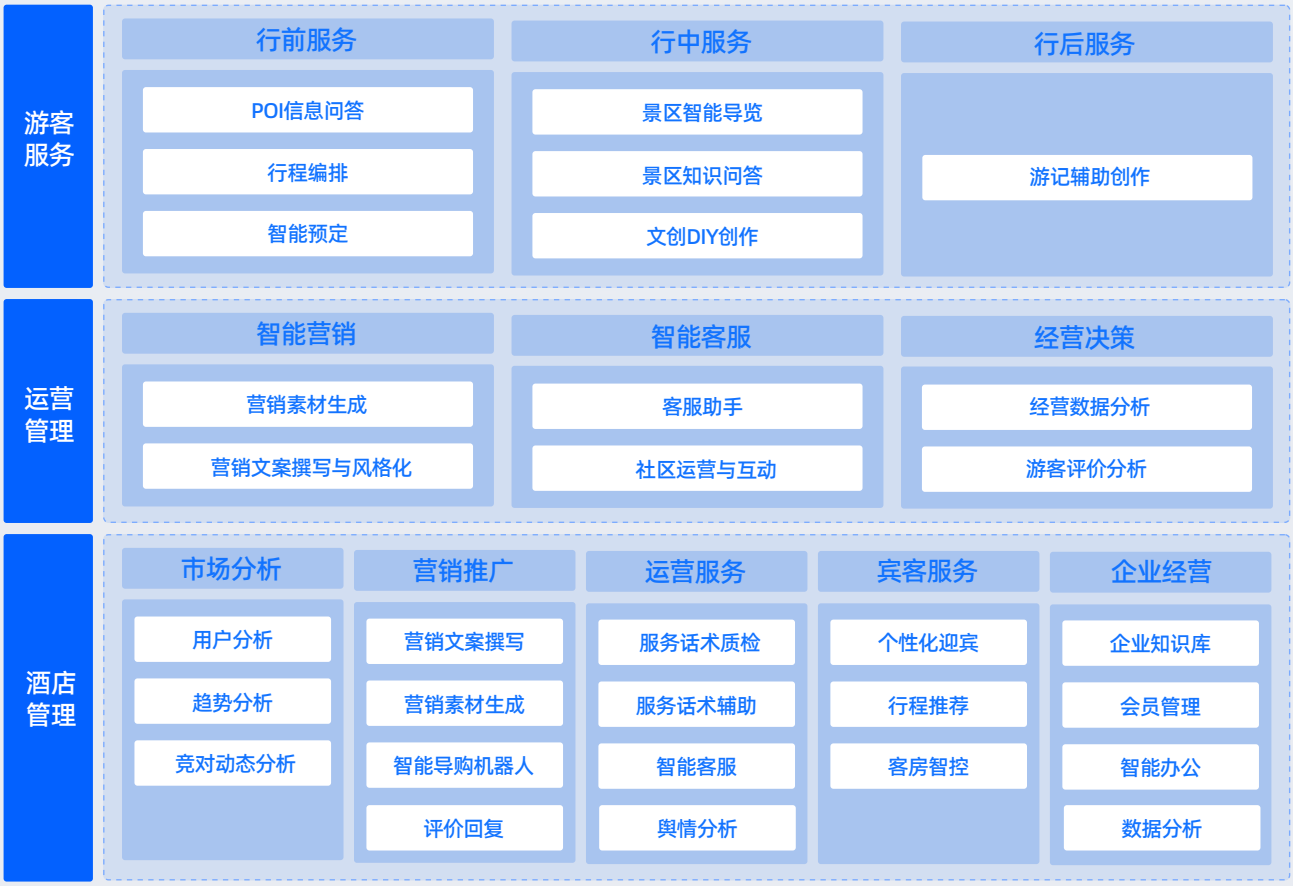
游戏开发是一项复杂且耗时的工程，从创意构思到实际落地的过程中，涉及项目策划、美术设计、程序开发、平衡性测试等多个环节，每个环节都需要精细打磨，任何细微改动都可能导致整体进度延宕；而因为游戏引擎技术更新迭代快速，开发者需要不断学习新技术并进行整合，特别是开放世界、多人在线等大型游戏项目，其内容丰富度、平衡性和稳定性测试都需要投入大量时间和精力去完善。大模型则能够帮助游戏公司构建一个全面而深入的知识库，在团队知识对齐上起到关键性作用，不仅包含了游戏设计的核心要素，如角色设定、故事情节、游戏机制等，还能够根据玩家行为和市场趋势进行动态更新，让策划团队在能够快速获取最新的信息和洞察的同时，确保不同团队成员和项目阶段之间的设计理念和目标保持一致性，使得游戏策划团队能够更加专注于创意和创新，为玩家带来更加丰富和引人入胜的游戏体验。



大模型 +文化旅游

智谱·AI

自从大模型技术应用以来，旅游业一直密切关注其发展，从客户使用的应用程序和小程序，到供应端的管理系统，大模型技术已经无声息地融入了旅游业的各个环节，例如供应链管理，大模型辅助的智能化运营管理能够有效降低成本并提高效率；同时对于消费者而言，大模型技术的融入使得行程规划更加高效，旅行体验也得到了显著提升。



典型场景分享

酒店营销素材生成：高效打造品牌差异化，帮助构建行业竞争壁垒

2023年文旅消费渐渐复苏，酒店行业的市场热度也居高不下，无论是从行业趋势、品牌发展必要性还是客户群体的迫切需要来看，酒店体验管理的必要性都在逐步凸显。酒店的环境、服务体验是酒店行业的基础，但是新一代主流客群更倾向于个性化、多样化需求，他们定义的酒店不再作为简单的差旅、歇宿和饮食等为一体的场所。酒店借助智能手段持续打磨“个性化体验”已成为趋势。而在智能营销领域，大模型能够根据酒店的品牌定位、目标客户群体和市场趋势，理解并生成符合特定营销策略的海报设计，还可以实时分析消费者反馈和市场动态，动态调整海报设计，确保营销素材的时效性和相关性，不仅提升了营销内容的质量和效果，也为酒店行业带来了创新的营销策略和方法，进一步推动了行业的数字化转型。

根据指令生成素材

帮我设计一个酒店早餐产品炸酱面的营销宣传图

这是我为您设计的酒店早餐炸酱面营销宣传图。图中展示了一碗诱人的炸酱面，热气腾腾的面条和丰富的酱料，以及优雅的餐具和舒适的环境，希望这样的设计能够吸引顾客对您的酒店早餐产品产生兴趣。您可以查看图片，看是否符合您的需求。

并能根据指令进一步微调

突出面本身，配料体现一些有营养的配菜

这是根据您的要求调整后的酒店早餐炸酱面营销宣传图。图中更加突出了炸酱面本身，同时加入了营养丰富的配菜，如新鲜的蔬菜，以展示早餐的健康和丰富性。希望这个设计更符合您的需求您可以查看图片，看是否满意。

POI 问答：打造旅游行业专属问答

在旅游行业中，大模型的应用正在深刻改变游客获取景点信息和服务的方式。在针对兴趣点（Point of Interest，简称 POI）的问答场景上，大模型不仅能够精准回答有关景点历史背景、文化特色、最佳游览时间、交通路线等各类问题，还能综合考虑季节、天气等因素，为游客量身定制游览建议。此外，大模型还可以模拟真人对话模式，针对游客的需求变化灵活回应，从而提升游客的整体旅行体验。帮助旅游业客户构建了一个智能化、人性化的信息服务桥梁，有效填补了传统静态信息展示方式的不足，促进了旅游业的服务升级与创新。

游客：POI信息问答

问题		样例问句
基础知识	地点、分类、特点等	Q:西安有哪些推荐的自然类旅游景点?
理解分析	区别、优缺点等	Q:坐公交和打车哪个更方便?
规划指导	日程安排、游玩攻略等	Q:请规划一下大唐不夜城的游玩攻略?

GLM 行业特色优势

智谱·AI

2023 年 3 月以来，在短短一年间，智谱AI 已拥有了千余家付费客户，并与超过 200 家企业进行了深度共创，越来越多的行业、业务、场景都已经实现了大模型的落地。而作为企业，更要主动拥抱大模型，探索这片愈发宽广的海域，找到模型既能解决又能节省较大成本的节点，找到模型擅长又能真正为业务带来价值、解决痛点的场景，锚定业务和大模型的最大公约数。最终，企业的“敢为人先”会让所有的尝试成为正向循环，一方面能够源源不断产生新的大模型时代的数据资产，构成新的壁垒和竞争力；另一方面，这些数据资产又能够灌回模型当中，变成企业独有的业务价值，耦合到业务场景当中。

完整的模型尺寸系列

中英双语、多模态　1.5B ~ 130B 全参数尺寸　完善的开源生态

语言大模型

ChatGLM3-1.5B 端侧模型	**ChatGLM3-6B** 开源/免费商用	**ChatGLM3-12B** 快速敏捷	**ChatGLM3-32B** 高性价比
ChatGLM3-3B 端侧模型	**ChatGLM3-130B** 最强大	**Character GLM** 高情商拟人对话	**GLM-4** 最新上线ALL tools

代码大模型

CodeGeeX3-32B 最新升级	**CodeGeeX3-6B**

多模态大模型

GLM-4V 多模态理解模型	**CogVLM** 多模态模型	**CogView** 文生图模型	**CogVideo** 文生视频模型

丰富的行业生态伙伴

基座模型调优

50+ 生态伙伴：应用定开、行业场景挖掘、行业系统对接

支持国产信创

全栈自研不依赖开源模型　支持国产GPU算力　避免“卡脖子”业务中断

丰富的落地经验

1000+ 付费客户、200+ 深度共创，覆盖汽车、大消费、制造、医疗健康、金融、游戏娱乐、文旅等 20 + 行业

大模型的部署与落地

智谱·AI

大模型部署常见问题

模型选择与场景选择

- 不同大小的模型到底怎么选择?
- 哪些场景适合用大模型解决?
- 大模型如何接入实际业务流程?

PoC 测试

- 无 Prompt 优化经验，测试效果总是不好。
- 是否需要微调？微调预计能达到什么样的效果?
- 数据标注有什么重点事项?
- 数据训练轮次、参数如何调整?

测试与验收

- 效果问题怎么定位?
- 性能维护要重点关注哪些指标，怎么扩容?
- 新场景合不合适?

双维度看模型选择

厂商模型选择

01 科研团队强大，模型能力优秀

02 项目经验丰富，落地效果更好

03 支持私有化，性能有保障，支持场景模型微调

04 持续服务能力强大，效果持续迭代能力强

模型规格选择

算力: 算力成本与资源评估

性能: 首响时间、对话效率要求标准

场景复杂度: 是否需要模型推理能力、Agent能力等，内容质量要求，知识复杂度等因素

环境部署要求: 是否私有化、是否本地、是否信创等因素

预算: 在符合以上所有条件中，推荐符合预算内的、规格尽量大的模型，保证最优效果

常见误区

Q 是否模型越大效果越好?

A 否，要根据场景选择，配合工程、指令等优化工作，效果也很好。

Q 是否所有场景，都能在小规格模型上做出一定效果?

A 否，小规格模型一定不是全能模型，在复杂场景中做不出效果，但发挥擅长的能力，结合大规格模型，做出更好的效果。

Prompt 调优准备

01 需求清晰：能够清晰定义产品需求，梳理功能流程图，确定大模型调用的输入和输出。

02 知识准备：结合产品功能，准备大模型推理需要的业务知识。

03 评测集构建思路：评测集数据分布与实际业务场景数据分布一致，数据类型涵盖全面，评测标准逻辑统一。

常见误区

Q 数据集是否可以人工构建？	A 最好是人工构建，结合业务目标，挑选各种难度、场景的数据作为Prompt调优的参考。
Q 数据集是不是越大越好？	A 否，通常生产环境中随机抽取的100~200条数据已经足够，过大只会增大标注评测成本。
Q Prompt是否所有大模型都通用？	A 否，Prompt有方言性，用同一个指令测所有大模型的方式是不科学的，每个大模型敏感的指令内容是不同的。

微调的判断与方案选择

01 基础效果评测：采用较大批量的评测数据集评测，按照交付标准分析badcase，拆分类型、归因，确定微调方向。

02 数据集构建：针对badcase，组建微调数据集，按照一定比例混合线上其他类型的数据，防止微调过程梯度爆炸以及知识遗忘。

03 数据标注：重新构建评测集，分别评测第一次和第二次评测集结果，评估微调效果，以及新一轮badcase；迭代多轮，直至达到上线标准。

常见误区

Q 是否所有场景都需要微调？	A 否，基于通识能力的场景，基本不需要微调，垂直行业、专业度要求高、知识复杂的场景可能需要微调。
Q 微调后，模型是否就具备了行业知识？	A 否，微调主要对齐的是标准、行业术语等，可以采用知识注入等方式。

智谱AI GLM

企业级解决方案

五个方向判断是否大模型 ready

01 决心： AI业务价值体现，需要的不仅仅是一次性的投入，还需要持续的迭代优化

02 投入： 确定预算范围，选择合适的模型、部署方式

03 企业现阶段数字化程度： 足够的高质量数据量，与一定量的专属知识库沉淀

04 明确的业务目标： 确定效果验证方法/指标

05 明确的试点场景： 大规模、高重复场景提效；提升服务品质；提升服务边界

业务场景落地全生命周期服务

行业分析

行业成功案例分享

业务流程梳理

价值场景选择

PoC 测试

PoC方案设计与实现

- 确定兼顾难度和业务需求的PoC场景
- Prompt指令设计与优化

...

系统流程架构设计分析

专属落地方案设计

- 大模型与业务系统接入交互逻辑设计
- 分阶段落地规划

...

部署交付

模型与效果交付

- 数据集准备、数据标注
- Prompt指令与微调

...

评测与验收

问题归因与定位

测试上线与验收

迭代与优化

模型升级方案

新场景规划

更适合中国市场的多样化部署模式

部署交付

方式一：API

开箱即用
方案灵活
成本低廉

适用企业 中小企业、个人开发者

- 敏捷迭代，希望快速实践大模型应用
- 公开数据信息丰富，通用行业适用

场景特征

- 快速上线，低成本试验
- 通用类场景，单轮对话只需完成一个任务
- 可通过指令调优获得优质效果

部署交付

方式二:云端私有化

专属模型
灵活微调
高性价比

适用企业 中、大型企业

- 希望探索行业垂类场景大模型场景
- 经验壁垒、专业知识壁垒高要求行业场景特征

场景特征

- 对于并发、首响时间有特殊要求
- 场景涉及较多行业知识或企业自有沉淀
- 指令功能复杂或需要微调，单轮对话需要实现多个目标

部署交付

方式三：本地私有化

私有大模型
二次优化
可控的高可用

适用企业 大型、超大型企业

- 计划构建专属垂类大模型，打造竞争优势
- 自有算力、硬件充足，或有特殊要求

场景特征

- 数据不能出内网，有特殊安全保密要求
- 支持最大程度灵活扩展与最高级稳定性要求
- 需要私有数据训练模型

助力大模型企业落地Plus服务

大模型咨询服务	PoC项目验证	部署陪跑	模型共创共建	内部培训	联合实验室

名词注释：

Prompt/指令调优：对大模型输入的提示语（Prompt）进行精心设计与优化的过程，旨在通过调整提示词句的表述、结构、风格等元素，引导模型生成更加准确、相关、高质量的文本输出，以满足特定应用场景的需求。

微调：在预训练大模型的基础上，针对特定任务或领域进行的小规模附加训练过程。通过在有限的标注数据集上调整模型参数，微调能够使模型在保留其泛化能力的同时，提升对特定任务的理解与表现，实现对新场景的快速适应与精准响应。

并发：在大模型应用中，指模型能够同时处理多个请求或任务的能力。并发执行意味着模型能够在同一时间段内高效地响应来自不同用户的请求，或者并行处理多个相互独立的数据单元，从而提高整体系统的吞吐量和服务效率。

首响：在大模型交互或服务场景中，特指用户首次发出请求后，模型生成并返回初始响应的时间间隔。首响时间是衡量模型响应速度和用户体验的重要指标，反映了模型从接收到请求到生成有效输出的即时性与效率。

德勤中国

Deloitte
德勤

资深报告撰写专家


文档解析


数据切分


信息提取


报告草稿生成


快捷翻译

10%左右

整体员工工作效率提升

项目背景

德勤中国为中国本地和在华的跨国及高增长企业客户提供全面的审计及鉴证、管理咨询、财务咨询、风险咨询和税务服务。德勤中国持续致力于为中国会计准则、税务制度及专业人才培养做出重要贡献。

传统运营模式问题

传统的工作模式，需要由有经验的顾问通读所有的文档，并摘取关键指标或文件信息，然后按照自己的经验来撰写报告草稿。中间会面临几个问题：

01 摘取关键信息耗时较长，且容易遗漏。

02 撰写草稿需要对报告规范很精通，对顾问的经验要求较高，且撰写中容易出错。

03 复核报告中的关键信息过程较长，寻找信息来源很费时。

04 报告需要单独请人翻译，耗时较长且费用较高，翻译质量也因人而异。

05 终端使用人群如何交互。

应用方案

在为客户提供非鉴证类报告撰写服务的场景中，顾问需要从若干访谈纪要或客户提供的大量文档资料中提取关键指标或文字信息，按照规范的格式生成报告草稿，人工复核后再提交给客户。

报告生成智能助手，充分考虑到客户对数据保密要求以及遵从跨境数据传输的监管政策，采用云私有部署智谱GLM系列大模型的方式，由用户自由上传采集的多种格式的文档资料，智能实现文档解析、数据切分、信息提取、报告草稿生成、快捷翻译等功能。

客户反馈

报告生成智能助手，借助智谱GLM系列大模型优秀的文本生成能力，解决了传统的信息提取、报告草稿生成、信息复核、文档翻译中面临的难题，极大的提升了顾问的效率，给客户带来了更满意的体验。通过对比测试该场景，在中文环境下表现出超过同类模型的能力。

实现方式

使用数据

咨询报告、行业研报。

分众传媒

营销必备文案专家“众智AI”

传媒/数字营销行业

广告语生成　　文案创作　　产品智能分析

项目背景

分众传媒成立于2003年，是主营电梯媒体的数字化媒体集团，拥有中国最大的户外媒体网络。2005年分众成为首家在美国纳斯达克上市的中国广告传媒股，并于2007年入选纳斯达克100指数；2015年回归A股，市值破千亿，成为中国传媒第一股。分众用户覆盖4亿中国城市主流消费人群，2021年营业额达160亿人民币。

传统运营模式问题

- **人力成本高、效率低**　　所有广告编写都需要由专业广告编辑编写。
- **数据统计易错漏**　　销售收集、整理广告主体的产品信息工作量大，信息不全面。

应用方案

智谱GLM系列大模型助力客户开发“众智AI”营销行业大模型，进行广告语生成、文案创作、广告主产品智能分析等操作。用户输入产品相关的信息， 模型即可输出符合分众广告风格的广告语。

客户反馈

智谱AI帮助分众构建营销行业大模型，实现传媒行业智能化路径。

实现方式

使用数据

5000条微调数据，如：品牌、产品、广告文案。

技术难点

数据收集与构建：
仅仅使用 Prompt 无法完成此项任务， 需要收集相关信息，并自动化构建训练数据。
模型基础能力：
广告文案对于基础模型的要求极高，需要基础模型已经学习过广告相关的内容，并且广告文案本身对于语言的运用更加严谨、复杂， 对于模型的基础能力有更高的要求。
对比结果：
对比于GPT-4，微调后的智谱GLM系列大模型生成的广告更加简练、准确、在风格和语言运用上更加贴近。

未来展望

01
当前阶段

客户的降本增效工具

02
第二阶段

智能创意制作平台

能够根据不同广告主的多样化需求，产出符合不同人群，不同媒体形态，不同营销诉求的广告创意。

03
第三阶段

营销领域智慧AI伙伴

利用AI和算法能力，基于分众历史投放的数据和后续效果数据的沉淀和学习，为广告主提供包含创意策略、预算策略、地域策略、人群策略、媒介策略的智能营销方案；持续成长的营销领域智慧AI，成为客户增长的可信任的伙伴。

华泰金融

华泰证券 HUATAI SECURITIES

金牌财富管理专家“涨乐财富通”

证券行业

智能客服

10~20%

效果提升
相较于通用大模型

项目背景

华泰证券，是中国领先的科技驱动型证券集团,是一家在上海、香港、伦敦三地上市的中国金融机构，以金融科技引领业务创新,为投资者提供专业、多元的金融服务,包括财富管理、机构服务、投资管理和国际业务。

传统运营模式问题

此前基于传统知识库机器人和模板引擎等能力，已经初步构建多场景客服服务体系。但仍存在产品形态孤立、意图识别泛化性不足、缺乏多轮会话理解能力等问题，导致客户使用体验不佳，未充分形成一站式财富管理助手形式的服务能力。

应用方案

基于智谱GLM系列大模型，我们构建了新一代的财富管理助手，提升用户体验。解决了传统技术无法对意图进行精准识别、无法与客户之间进行多轮交互的问题，加强了与用户的交互体验和一站式服务的能力。

基于智谱GLM系列大模型，叠加了40-50G金融专业书籍、资讯、百科、法规、上市公司公告等金融专业数据进行增量训练，以及投研、客服场景上万条高质量指令集进行指令微调，形成华泰金融大模型1.0，表现出明显的金融领域优势。

实现方式

使用数据

公告、年报、研报等专业金融数据。

技术难点

对于多轮多意图的语义理解能力，以及文档精准有效的总结，在国内外模型能力中领先。智谱GLM系列大模型作为纯国产、自研大模型，同时支持本地私有化部署，有效保证了数据安全。

未来展望

进一步整合全业务系统，对更加复杂的任务多次调用工具。

华信永道

Yondervision
华信永道

24小时在线的公积金智能客服

政务、公积金、金融科技行业

项目背景

华信永道（北京）科技股份有限公司（股票代码：837592）成立于2007年，是国内行业领先的政务及银行业数字化解决方案的供应商和服务运营商。为贯彻落实国家发展新一代人工智能的决策部署，公司全面适配国家“信创”要求，选用“智谱AI”等AI大模型底座，积极研发大模型在政务场景的应用实践。于2023年成立了专门的研发团队并建立了华信永道（北京）人工智能科技有限公司，旨在引领大模型在政务领域的发展和创新。2023年11月2日，公司与北京智谱华章科技有限公司生成式人工智能（AIGC）大模型在数字政务服务领域签订了《人工智能大模型共建战略合作协议》，正式建立战略合作伙伴关系。

传统运营模式问题

基于大模型的客服相比传统机器人客服，展现出更强的语义理解能力和更流畅的对话处理能力。它们能够提供个性化服务，快速响应用户需求，降低运营成本，并提高回答的准确率。相反，传统客服机器人受限于预设规则，缺乏灵活性，难以应对自然语言的复杂性，且需要频繁的人工更新和维护，导致用户体验和效率受限。

应用方案

用户通过公积金中心线上公众号、官网、政务公共服务互联网入口等与在线客服交互。公积金智能客服即可解答市民对公积金政策的问题。

客户反馈

智谱AI与华信永道在政务垂直领域展开深度合作。智谱AI是专注知识智能的创新型科技公司，是国内可以比肩OpenAI的行业龙头。华信永道在政务公积金及软件工程领域有着深厚积累，且有着丰富的产品线和高市场占有率。智谱GLM系列大模型在先发优势和强有力的技术团队加持下，已取得了显著效果和进展。面向未来，随着智谱GLM系列大模型垂直领域深耕与落地，创新服务场景，为政务条线客户提质增效、为广大市民提供卓越服务体验，形成新质生产力，增强民生福祉、推进社会事业发展。

实现方式

使用数据

苏州公积金政策知识、全国公积金政策背景知识。

技术难点

知识专业度很高、深（涉及金融、房产、法律、公积金等多领域知识），需要算法成为业务专家。

数据与业务逻辑融合较复杂（数据的组织、清洗、聚合等需要结合领域知识），迭代周期长。

为了增强对话数据质量，需要在训练数据中添加 大模型思维链 COT（chain of thought）以及使用智谱GLM系列大模型辅助对数据进行润色，使其更符合智谱GLM系列大模型数据分布，降低模型学习难度。

未来展望

未来会在其他城市上线公积金智能客服，如济南、深圳等。

智己汽车

智己汽车

感情与精神寄托的第三生活空间

汽车行业

700+ 车控场景

95%+ 座舱意图支持

识别准确率

30%+ 响应速度提升

项目背景

智己汽车-上汽集团旗舰品牌，是阿里巴巴智慧赋能，专注打造的高端纯电智能车。智己汽车聚焦「智能化」，旨在成为智能时代出行变革的实现者。

传统运营模式问题

· 一问一答的交互限制：

目前的车机系统对于语音指令理解率低，用户需要明确指令才能执行操作，车机被动执行命令。

· 回复内容偏生硬：

目前的回复内容是基于传统的模版方案生成。内容回复缺乏时效性和趣味性。

应用方案

· 模型意图二次确认：

对于模型的意图，通过引导对话的方式确认。

· 趣味内容生成：

包含大量笑话、故事等趣味内容,能够根据场景和用户的描述智能生成相关的趣味内容，增强互动趣味性。

· 语音交互游戏：

支持语音交互的游戏,如成语接龙、猜谜语、智力问答等。

· 多元人设： 可参照不同人设风格进行聊天。

客户反馈

“IM生成式大模型”深入挖掘海量座舱交互体验数据并构建最新模型智能体能力框架进行整体升级换代，与端侧大模型协同配合，将综合复杂场景进行云上分流，共同实现更丰富、更贴心、更惊艳的多模态AIGC智能化场景，于人车路互联之际，以AI驱动万物创新，打造智能出行峰值体验新世界。

实现方式

使用数据

客户提供座舱内的语料数据。

技术难点

传统车机系统对于语音指令理解率受限，用户需要明确指令才能执行操作，当用户有更加自由的说法时，车机无法理解用户的真实意图，智谱GLM系列大模型通过AI技术使用更少的语料支持座舱更多的意图，赋予座舱更准确、更流畅的语音识别功能，以及更丰富的知识储备和语义理解能力。

未来展望

利用智谱GLM系列大模型算法能力，基于车机座舱交流的对话记录和后续效果数据的沉淀和学习，为智己汽车用户提供更多元的功能，未来或可能由机械控制演进到电子控制，实现“无按键交互”，且各独立的电子信息系统逐步整合，组成“电子座舱域”。打破指令式，一问一答的交互限制，从“简单问答”进化到无障碍、直观的自然交流，重塑人机交互体验，让座舱具备个性化、情感化、自由化的交互能力。

无缝交互，有问必答

不分大小，事事回应
多轮交互，上下文记忆
无割裂感的全场景交互

多重意图，瞬间感知

多重意图识别，感知真实需求
专属定制，独特出行体验

人设多元，沉浸超然

灵活切换多种人设
跨时空、跨文化的心灵交流

情感理解，温情对话

冷暖共情，情感陪伴
提升语音的智能化和情感化

马蜂窝

马蜂窝

你的靠谱旅游小助手

旅游行业

AI旅行伙伴

个性化旅游定制

80+分

准确性、完整性、实用性全链路测评分

项目背景

马蜂窝是中国年轻旅行者首选的旅行社交平台、旅行APP，以其独特的“内容+交易”模式，提供全球旅游目的地的一站式信息服务和产品预订服务。通过AI技术和大数据分析，马蜂窝为用户打造个性化旅行体验，同时以“内容获客”模式助力商家提升获客效率，推动旅游行业的创新发展。

传统运营模式问题

- **产品体验：**在海量的旅游内容中，用户难以快速、精准地筛选出所需有价值的信息，导致体验感下降。
- **内容质量：**无法根据用户不同的旅行偏好和需求，自动化地提供个性化的旅游内容和产品服务。
- **数据价值：**缺乏开发、变现大量非结构化高质量内容的有效手段。

应用方案

在马蜂窝APP内打造AI旅行伙伴（小蚂）应用，用户可通过与AI小蚂对话进行全球旅游问题咨询。

- **意图理解：**

通过大模型完成用户个性化、多样化的意图理解，用于提供更精准的问题回答、个性化的内容服务。

- **语义补全：**结合历史对话信息、指代信息进行用户输入补全，使对话内容更具连续性。
- **信息抽取：**

从历史对话中抽取POI关键信息、天气信息、问题核心点等，建立用户行程信息库，用于后续的对话回忆、问题引导、精准回答。

- **内容生成：**

结合海量数据、旅游知识，以“新、准、全”为标准，为用户提供靠谱的旅游攻略、咨询服务。

- **内容总结：**多轮对话历史总结，增强对话内容的连贯性。

客户反馈

智谱AI凭借专业的技术实力，为马蜂窝提供了可靠的大模型底座，携手探索全球旅游“新玩法”，致力将个性化服务规模化，共同推动行业创新。

实现方式

使用数据

官方旅游攻略、用户旅游笔记、用户旅游游记、历史QA总结。

技术难点

- **模型能力：**

需要基于多个知识切片进行总结回答，效果不佳，易产生幻觉。行程规划等旅游问题复杂，需模型具备强大的语言理解、逻辑推理能力。

- **效果评估：**

用户咨询问题个性、多样，答案较为主观，需建立完善的效果评测体系。

未来展望

图文问答

引导用户旅游问题问答，多模态能力结合，能够实现随拍随攻略，丰富AI旅行伙伴（小蚂）的问答形式。

内容助手

用户旅游笔记、游记内容生成助手，提升内容发布效率和用户体验。

运营助手

社群运营，自动化、个性化为社群用户提供答疑、推荐等服务，并增强属地化社交属性。

蒙牛

健康守护营养师“AI蒙蒙”

乳制品行业

健康评测

营养健康问答

营养健康内容生成

提升>10%

用户在私域和公域内消费意愿

项目背景

蒙牛集团（简称“蒙牛”）1999年成立于内蒙古自治区，总部位于呼和浩特，是全球乳业八强，2004年在香港上市（股票代码2319.HK）。蒙牛专注于为中国和全球消费者提供营养、健康、美味的乳制品，共有6大品类，400多款单品，形成了包括液态奶、冰淇淋、奶粉、奶酪等品类在内的丰富产品矩阵；拥有特仑苏、纯甄、冠益乳等明星品牌。在高端纯牛奶、低温酸奶、高端鲜奶、奶酪等领域，市场份额处于领先地位。2022年，蒙牛实现全年收入近1000亿元。
蒙牛不是从现有痛点出发，而是前瞻的、从更好服务消费者的角度，引入了AIGC技术，构建了营养健康领域模型MENGNIU.GPT。

传统运营模式问题

人们对营养健康的日益重视（高品质牛奶、功能性牛奶高速增长），但专业的营养健康服务供给严重不足（中国每10万人平均只有0.3名营养师，远低于全球27名营养师的水平，且价格昂贵难以普及）。渴望获得相关专业知识的需求不断增长，越来越多消费者及家庭希望有一个能够随时随地提供专业解答的营养师助手。

应用方案

基于MENGNIU.GPT构建了AI营养师蒙蒙（应用），消费者可以7*24h通过自然语言与AI营养师蒙蒙自由沟通，获得专家级的个性化营养健康服务。
此外，还有AI planner功能。基于个人健康评估结果，个性化地制订营养健康计划，实时提醒互动，并记录执行过程。当执行过程和计划发生偏差时，智能调整计划。

蒙牛与智谱AI联手打造MENGNIU.GPT、AI营养师蒙蒙，致力于让每个家庭都可以拥有7*24h、个性化、专业级的营养健康服务，场景涵盖：

· 智能健康营养专家：

专业知识解答、个性化营养建议。

· 用户营养健康评测：

包括缺钙风险评估、免疫力状况评估、营养均衡测试等。

· 营养计划制定及监督：

营养配餐、智能共情陪伴、智能提醒、过程辅导和激励、目标和计划动态调整等一系列服务，如瘦身计划、肠健康养护计划等。

客户反馈

在营养健康领域，我们看到了大模型带来颠覆性影响。相信未来通过和智谱AI更加深入的合作，可以让更多家庭享受到更加个性化、陪伴式的专业营养健康服务。

实现方式

使用数据

蒙牛二十多年积累的营养健康相关的私域知识。
合作的营养健康权威机构的知识数据。
与多位知名营养健康领域的专家学者，以及多名中医专家、教授、名医等进行学术研讨，确保在学术和实践领域都具实用性。

技术难点

蒙牛作为业内领先的龙头企业，在交付要求上保持着一贯的高水准，而且营养健康领域的专业知识需要更加严谨的训练。为了深化AI技术与领域知识的融合，智谱AI需投入更多的技术资源和人力成本进行前期沟通与后期交付。

智谱AI凭借国产化背景、自主研发能力和与国际先进水平相媲美的模型技术等优势，有效地协助蒙牛集团拓展更多合作伙伴，如将AI营养师助手解决方案引入医院营养科，提供更加智能化和个性化的营养服务。

未来展望

素材生成

内部知识库

数据分析

上汽集团

汽车维修AI师傅

汽车行业

售后接待

维修
（动态排故流程）

80%准确率

单轮对话抽取达到业务人员

90%

多轮对话修正后可以达到

项目背景

上汽乘用车：上海汽车集团股份有限公司乘用车公司，是上海汽车集团股份有限公司的分公司，承担着上汽自主品牌汽车的研发、制造与销售。 从诞生之日起，乘用车公司就依托上汽集团20多年合资合作所积累的技术、制造、采购、营销和管理优势，以国际化的视野，创造性地集成全球优势资源，以高品质的产品与服务，满足消费者高品位需求，以优秀的国际合作团队，打造中国人自己的国际汽车领导品牌。

传统运营模式问题

售后/维修工能力水平一致性希望从当前状态得到进一步提升。需要增强总部支持的“预诊断”功能，赋能维修工，统一积累经验知识。

应用方案

智谱GLM系列大模型赋能维修助手、一线的售后咨询人员、车辆维修人员可以通过耳机+麦克+触屏的智能终端方式，实时与维修人员交互，提供需确认问题/处置方案/预诊断结论等。

- **初诊：** 初诊计划＋概率排序、过程引导、实时描述＋计划修正、初检结论（可能故障＋概率排序）；
- **诊断：** 诊断方案＋概率排序、过程引导、实时描述＋计划修正、排查结论（故障定位）；
- **修理：** 修理计划、过程引导、实时描述＋计划修正；
- **验收：** 验收测试计划、过程引导、实时描述；
- **报告：** 过程记录自动生成维修记录。

实现方式

使用数据

包含车辆维修手册，车辆排故手册，历史维修案例。

技术难点

单轮对话实现多特征抽取，以及语义判断，并通过业务多轮反问实现问题定位。

未来展望

以现有对话语料及手册为数据基础，通过All Tools构建以模型为主不受流程限制的售后服务的Agent。

金山办公

金山办公 KINGSOFT OFFICE

人工智能办公应用“WPS AI”

办公软件


文档生成


内容改写


续写


公文写作

95%+

用户满意度

项目背景

办公软件是企业和个人用户使用最为高频的工具之一。随着人工智能技术的发展，用户对于办公软件智能化提效的需求不断提升。大模型带来的技术突破，让办公软件的智能化大幅提速，文本生成等应用有望变革用户的工作方式。对于办公软件厂商而言，这也意味着变革性的产品创新机遇和新的战略增长点。金山办公是国内领先的办公软件和服务提供商，在办公软件领域有三十多年研发经验及技术积累，主要产品包括 WPS Office、WPS 365、WPS AI等办公能力产品矩阵以及各类行业的数字办公解决方案。截至 2023 年 9 月 30 日，金山办公主要产品月度活跃设备数达5.89 亿。

传统运营模式问题

作为行业领先的办公软件，WPS一直以来都注重结合AI技术为用户提供智能、高效的产品体验。但是，基于传统NLP技术开发智能化功能，面临以下问题：

受限于技术能力边界，基于传统NLP技术开发的智能化助手功能在实际用户体验上不够智能化，生成的内容可用率低，用户满意度不高。

针对各个小场景，金山办公需要分别训练相应的小模型，研发成本高，产品开发周期长，导致产品化难度高。

同时，针对各个小场景单独进行研发维护，需要大量的人力投入。在人力有限的情况下，上线的智能化产品场景少，体验不佳。

应用方案

大模型为办公软件智能化带来了全新可能。基于大模型技术，金山办公推出了全新的智能办公助手产品WPS AI，提供文字、PPT等文档生成，内容改写、续写，公文写作等功能，帮助用户提高生产力。

WPS AI可以实现的典型场景包括：

在文字编辑场景中，可以通过大模型辅助用户完成创作，包括续写、总结、缩写、扩写、改写（口语化、正式化等）、大纲、翻译、头脑风暴等功能。尤其是可以发挥公文模板库的作用，包括会议纪要、公文通知、简历、合同、喜报、参数提取、会议、请假条等等。用户只需要输入简单的文档标题和关键信息，就可以快速生成文档内容。

在PPT演示文档场景，可以通过大模型生成PPT的大纲、内容、演讲稿等。用户只需要输入PPT的主题，WPS AI会通过多次调用大模型逐步产出以上内容，并完成PPT的排版工作。

目前，WPSAI已接入金山办公的WPS文字、演示、表格、PDF等产品，解决用户在内容生成、内容理解、指令操作等方面的日常办公难题，以及搭载在面向企业组织提供服务的新产品WPS 365，进一步发展办公新质生产力，为企业数字化转型搭建智能基座。

从用户反馈来看，对比传统NLP小模型支撑的智能化场景的“人工智障”体验，基于智谱GLM系列大模型能力反复打磨的WPS AI用户满意度超过95%，用户满意度大幅提升。

客户反馈

在与我们产品功能的深入融合中，智谱GLM系列大模型表现出了强大的综合能力，在多个场景具备更优的效果，很好地支撑了我们文字、PPT等产品线的智能化升级。

实现方式

技术难点

产品直接面向C端用户，为了保证用户体验，对场景功能采纳率的要求较高。

智谱GLM系列大模型在大部分核心场景中表现出更好的效果。

未来展望

未来，金山办公将探索知识库问答等更多大模型能力与产品的结合。

智联招聘

智联招聘

招聘提效助手

在线招聘行业

人岗匹配

80%以上

筛选简历
时间节省

简历优化

100%

求职者
写简历效率提升

面试助手

50%

招聘方
面试效率提升

项目背景

在线招聘行业，旨在利用互联网平台和技术，为求职者和企业提供招聘服务。在线招聘行业的发展受到多种因素的影响，包括宏观经济环境、互联网技术的发展，以及求职者和企业需求的变化等。随着技术的不断进步和用户需求的变化，这个行业仍在持续发展和创新中。

智联招聘是中国领先的人力资本生态价值链平台，为大型公司和快速发展的中小企业提供一站式专业人力资源服务，包括网络招聘、校园招聘、海外招聘、猎头服务、人才测评和培训、BPO等。智联招聘拥有超过3.49亿职场人用户，累计合作企业数超过1341万，业务遍及全国400多个城市。

传统运营模式问题

· 平台信息量大：

平台拥有大量的职位信息和求职者简历，企业筛选简历效率较低；人岗匹配精准度进一步提升后，可以提高招聘求职效率。

· 求职者需求增多：

求职者简历缺乏竞争力，衍生出增值服务，包括简历优化、职业规划、行业分析等。

应用方案

面向求职招聘垂类场景提供简历优化、筛选简历、招聘助理等应用能力，具备知识问答、文本生成、内容总结、模拟对话等基础能力。

· 企业端（招聘方）：

筛选简历：通过对话的方式理解招聘方的需求，为招聘方筛选简历，进行定向推荐。

招聘助理：在面试中帮助招聘方分析求职者能力，给出面试后的综合评价。

· 求职者端：

简历优化：求职者输入个人优势等信息，自动生成完整简历，并且可以根据求职者需求对简历进行优化。

客户反馈

在智谱GLM系列大模型的支持下，深度优化了人岗匹配、通过对话进行信息收集的效率和准确性，助力智联在招聘领域持续保持领先。

实现方式

使用数据

岗位JD、简历、通用行业知识。

技术难点

AI模拟面试场景产品需要垂直领域知识作为辅助，才能提出有专业性有深度的提问，这对大模型基础能力要求高，需要强大的信息检索、信息匹配能力，用来学习理解海量的岗位信息和垂直领域知识。

未来展望

在AI面试、智能约面、撮合企业和求职者成交等招聘全流程应用大模型技术，提升用户体验和效率。

坚持完全自研
打造新一代认知智能大模型

智谱AI自研了具有完全知识产权的预训练框架GLM，并自建训练平台，拥有从零开始搭建平台和运维平台的能力，并在国内外测评中均取得了领先效果优势。
2024年1月,新一代基座大模型GLM-4正式推出。GLM-4的整体性能相比上一代大幅提升,逼近GPT-4。

GLM-4，新一代基座大模型

GLM-4支持更长上下文,具备更强多模态能力,推理速度更快、支持更高并发,从而大大降低推理成本。
GLM-4大幅提升智能体能力,GLM-4 AllTools实现自主根据用户意图，自动理解、规划复杂指令,自由调用网页浏览器、CodeInterpreter代码解释器和多模态文生图大模型以完成复杂任务。
GLMs个性化智能体定制功能亦同时上线,用户用简单的提示词指令就能创建属于自己的GLM智能体。

模型能力全面对齐世界先进水平

- 不分大小，事事回应
- 多轮交互，上下文记忆
- 无割裂感的全场景交互

智能体能理大幅增强

- 灵活切换多种人设
- 跨时空、跨文化的心灵交流

模型能力全面对齐世界先进水平

性能全面提升

GLM-4在基础能力、指令跟随能力、中文对齐能力等方面全面看齐GPT-4,达到世界先进水平。

基础能力评测(英文)

	MMLU (5-shot)	GSM8K (5-shot)	MATH (4-shot)	BBH (3-shot)	MMLU (5-shot)	HellaSwag (10-shot)
GPT-4	86.4	92.0	52.9	83.1	95.3	67.0
GLM-4	81.5	87.6	47.9	82.3	85.4	72.0
GLM-4/GPT-4	94%	95%	91%	99%	90%	100%

指令跟随能力(中英,数据集:谷歌 IFEval)

	Prompt 级别、中文	Instruction 级别、中文	Prompt 级别、英文	Instruction 级别、英文
GPT-4	72.4	80.0	79.5	85.4
GLM-4	63.4	71.9	67.7	76.4
GLM-4/GPT-4	88%	90%	85%	89%

对齐能力（中文，数据集：AlignBench）

	专业能力	中文理解	基本任务	数学计算	文本写作	综合问答	角色扮演	逻辑推理	中文推理	中文语言	总分
GPT-4	7.94	6.93	7.81	7.65	7.93	7.42	7.51	7.37	7.47	7.59	7.53
GPT-4 Turbo	8.65	7.33	7.99	7.80	8.67	8.61	8.47	7.66	7.73	8.29	8.01
GLM-4	8.91	8.07	7.87	7.75	8.44	8.42	8.58	7.01	7.38	8.38	7.88
GLM-4/GPT-4	112%	116%	101%	101%	106%	113%	114%	95%	99%	110%	105%

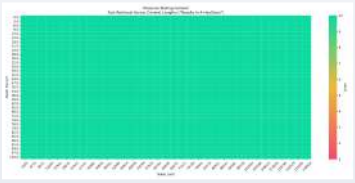
更长上下文

GLM-4在基础能力、指令跟随能力、中文对齐能力等方面全面看齐GPT-4,达到世界先进水平。

长上下文能力评测(中文,数据集:LongBench,NeedleTest)

GPT-4	71.2
GPT-4 Turbo	82.7
Claude 2.1	80.4
GLM-4	81.1.
GLM-4/GPT-4	114%

Needle Test(128K)，全绿表示100%找回精度



更多模态

GLM-4在基础能力、指令跟随能力、中文对齐能力等方面全面看齐GPT-4,达到世界先进水平。

文生图性能评测

	Alignment	Fidelity	Aesthetic	Safety	Composition &Layout	Lighting &shadow	Color User	Emotional Response
SDXL(开源最佳)	0.467	0.740	0.6945	0.978	0.739	0.717	0.720	0.597
DALLE.3	0.770	0.852	0.735	0.980	0.772	0.772	0.746	0.649
CogView3	0.706	0.802	0.702	0.973	0.733	0.728	0.708	0.593
CogView3/ DALLE.3	91.7%	94.1%	95 .5%	99.3%	94.9%	94.3%	94.9%	91.4%

关于智谱AI

智谱AI致力于打造新一代认知智能大模型，专注于做大模型的中国创新。公司于2020年底研发GLM预训练架构，2021年训练完成百亿参数模型GLM-10B，同年利用MoE架构成功训练出收敛的万亿稀疏模型，2022年合作研发了中英双语千亿级超大规模预训练模型GLM-130B并开源。2023年，智谱AI推出千亿基座对话模型ChatGLM并两次升级，开源版本的 ChatGLM-6B让大模型开发者的本地微调和部署成为可能，在开源社区受到广泛欢迎。

2024年1月，智谱AI推出新一代基座大模型GLM-4，整体性能相比上一代大幅提升，比肩世界先进水平。它支持更长上下文，具备更强多模态能力，推理速度更快，支持更高并发，大大降低推理成本。同时，GLM-4智能体能力得到大幅提升，可根据用户意图，自动理解、规划指令以完成复杂任务。GLMs个性化智能体定制功能亦同时上线，用户用简单提示词指令即能创建属于自己的GLM智能体，由此任何人都能实现大模型的便捷开发。

基于全自研基座大模型的强大能力，智谱AI构建了极具竞争力的AIGC模型产品矩阵，包括AI提效助手智谱清言、高效率代码模型CodeGeeX、多模态理解模型CogVLM和文生图模型CogView等。

践行Model as a Service市场理念，智谱AI致力于打造高效率、通用化的“模型即服务”开发新范式，通过大模型链接物理世界的亿级用户，为千行百业带来持续创新与变革，加速迈向通用人工智能的时代。

2019年

智谱AI成立，源自清华技术成果。

2020年

专注大模型算法研究。

2021.9月

设计GLM算法，发布拥有自主知识产权的开源百亿大模型GLM-10B。

2022.8月

发布高精度千亿模型GLM-130B并开源，效果对标GPT-3（175B），收到70余个国家1000余个研究机构的使用需求。

GLM-130B

2022.9月

发布代码生成模型CodeGeeX，每天帮助程序员编写2000万行代码。

CodeGeeX

2022.10月

发布开源的100+语言预训练模型mGLM-1B。

2023.3月

发布千亿基座的对话模型ChatGLM及其单卡开源版本ChatGLM-6B,全球下载量超1300万。

ChatGLM

2023.5月

开源多模态对话模型VisualGLM-6B（CogVLM）。

2023.6月

发布全面升级的ChatGLM2模型矩阵，多样尺寸，丰富场景，模型能力登顶C-Eval榜单。

2023.8月

作为国内首批通过备案的大模型产品，AI生成式助手“智谱清言”正式上线。

智谱清言

2023.10月

发布全面升级的ChatGLM3模型及相关系列产品，参数范围从6B、12B、32B、66B到130B 不等。

2024.1月

新一代基座大模型GLM-4正式推出，整体性能相比上一代大幅提升，比肩世界先进水平。

免责申明：

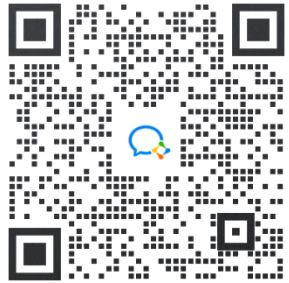
本内容非原报告内容；

报告来源互联网公开数据；如侵权请联系客服微信，第一时间清理；

报告仅限社群个人学习，如需它用请联系版权方；

如有其他疑问请联系微信。

行业报告资源群



微信扫码 长期有效

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6+份行研精选、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息

微信扫码 行研无忧

知识星球 行业与管理资源

专业知识社群：每月分享8000+份行业研究报告、商业计划、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等；已成为投资、产业研究、企业运营、价值传播等工作助手。

智谱·AI | ChatGLM

联系我们

官网：https://open.bigmodel.cn
合作伙伴申请：https://open.bigmodel.cn/partner
联系电话：400-6883-991
微信公众号：智谱、GLM大模型