

2025 大模型平台 落地实践研究报告

版权声明

本报告版权属于北京百度网讯科技有限公司、中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：北京百度网讯科技有限公司、中国信息通信研究院”。违反上述声明者，编者将追究其相关法律责任。

前言

以大模型为核心的人工智能技术正以空前的速度、规模和深度重塑全球格局，深刻改变着人类的生产生活方式，对世界经济发展、社会变革及文明进程产生重大而深远的影响。大模型平台作为连接大模型技术与应用场景的桥梁，近两年来发展迅速，从早期单一功能的模型训练调优能力，到如今集模型训练、推理部署、服务供给、应用开发、运营管理于一体的综合性平台，成为大模型规模化落地的有力支撑。

大模型平台是大模型落地的工程化基座，将基础模型转化为业务场景下的大模型应用，构建起覆盖模型开发、调优、部署、管理、应用的全生命周期工具链与技术支撑体系，为大模型落地过程中面临的技术转化、场景适配、效率提升等核心挑战提供系统性解决路径，破解从模型开发到实际应用的落地壁垒。然而大模型平台在建设落地过程中，面临技术选型难、场景应用复杂多样、无建设指引等问题，且如何为数据和模型提供安全可信的管理和应用环境，如何为大模型应用提供灵活便捷的构建能力，成为大模型平台是否好用的关键。

本报告旨在为各行业企业在建设落地大模型平台的过程中，提供有效的参考和指引，助力大模型更高效更有价值地规模化落地。本报告系统性梳理了大模型平台的发展背景、历程和现状，结合大模型平台的特点提出了具体的落地策略与路径，同时详细分析了大

模型平台的关键能力及其建设要点，并对多个行业的落地案例进行了深入剖析，最后对大模型平台的发展趋势进行展望。

希望本报告能够为行业提供有益参考，共同推动大模型平台赋能千行百业，助力数字经济高质量发展。由于大模型等人工智能技术产业仍处于快速发展阶段，大模型平台的能力也在持续演进，我们的认知将随着产业实践不断深化，报告存在的不足之处，恳请大家批评指正。

目 录

一、 大模型平台发展概述	1
（一） 大模型落地政策加持产业需求旺盛	1
（二） 大模型落地存在诸多问题与挑战	5
（三） 大模型平台成为化解落地难题的关键力量	6
二、 大模型平台落地策略与路径	9
（一） 大模型平台落地原则	9
（二） 大模型平台落地步骤	10
（三） 大模型平台落地框架	14
三、 大模型平台关键能力建设	16
（一） 模型开发层	16
（二） 模型服务层	18
（三） 应用开发层	20
（四） 大模型平台产品实践	21
四、 基于大模型平台的实践案例分析	26
（一） 互联网电商行业案例	26
（二） 金融行业场景实践案例	27
（三） 教育行业案例	28
（四） 企服行业案例	30
（五） 智能硬件行业案例	32
（六） 制造行业场景实践案例	33
五、 总结与趋势	36

图 目 录

图 1 AI 应用产业链分布图5

图 2 大模型平台地步骤图 11

图 3 企业智能化能力等级图 12

图 4 大模型平台建设方案图 13

图 5 大模型平台落地框架图 14

图 6 百度智能云千帆大模型平台 3.0 架构图 23

图 7 微软 Azure AI Studio 示意图24

图 8 亚马逊 AWS Bedrock 平台示意图25

图 9 爱库存超级助手架构图 27

图 10 泰康保险集团基于大模型平台的场景示例 28

图 11 考试宝工具大模型平台一站式应用架构 29

图 12 WPS AI 大模型平台应用架构31

图 13 智能硬件大模型应用架构图 33

图 14 山港大模型框架 34

一、大模型平台发展概述

在我国大模型技术快速发展的进程中，政策驱动与市场需求形成双重引擎，推动技术成果加速向各行业渗透，展现出强大的产业赋能潜力。然而大模型落地实践中，顶层规划模糊、技术壁垒高起、运营管理机制缺失等问题亟待破解。作为关键基础设施，大模型平台通过构建“建模型、用模型、管模型”的全链条支撑体系，系统性化解技术转化与工程化应用难题，为大模型从研发创新迈向规模化落地搭建桥梁，有力推动人工智能与实体经济深度融合的高质量发展。

（一）大模型落地政策加持产业需求旺盛

1. 大模型技术带来先进生产力

在全球人工智能蓬勃发展的浪潮中，大模型技术成为推动产业变革的核心力量。其技术发展经历了多个阶段，演进路线逐步清晰，起初主要以语言模型的形式出现，专注于自然语言处理任务。随着技术的不断进步，多模态模型逐渐兴起，融合了文本、图像、音频等多种数据类型，打破了单一模态的限制，实现了跨模态的理解和生成。以 OpenAI 的 GPT 系列为例，从 GPT-1 到 GPT-o1，模型的能力不断提升，上下文窗口长度得以扩展，知识密度增强，推理能力显著提高。同时，视觉大模型也在快速发展，Transformer 架构在视觉领域的应用，如视觉 Transformer（ViT）模型，推动了图像理解和生成技术的进步。

国内大模型发展迅速成果斐然，近年来，我国大模型技术在核心能力提升与场景落地方面取得显著进展，深度求索的 DeepSeek-R1 在数学推理、代码生成等复杂任务中展现国际领先水平，其在 AIME2024 和 MATH-500 测试中赶超 OpenAI o1-1217¹，后续发布的 DeepSeek-V3-0324 进一步优化了推理效率。百度文心大模型 4.5 及 X1 在多模态与深度思考领域实现突破，最新发布的文心大模型 4.5 Turbo 及 X1 Turbo，能力更强、成本更低。如文心大模型 4.5 通过多个模态联合建模实现协同优化，提升多模态理解能力，在 C-Eval、MMLU-Pro 等多个公开数据集上表现赶超 GPT-4o²，其 API 调用成本仅为 GPT4.5 的 4%³。阿里通义千问的开源推理模型 QwQ-32B 在测试数学能力的 AIME24 评测集上，以及评估代码能力的 LiveCodeBench 中，表现远胜于 OpenAI-o1-mini⁴。字节跳动 Doubao-1.5-pro 在知识、代码、推理、中文等相关的多个评测中表现优异，综合得分优于 GPT4o、Claude 3.5Sonnet 等业界一流模型⁵。

2. 政策大力支持大模型及人工智能发展

我国自上而下高度重视大模型及人工智能发展，出台了一系列政策以推动其落地。

从国家层面看，“十四五”规划将智能化领域作为数字中国建设的主要方向之一，为大模型发展提供有力支撑。从技术攻坚、产业落地及标准规范方面，发布了多项政策规范，如国家自然科学基金委员会

¹ <https://github.com/deepseek-ai/DeepSeek-R1>

² <https://mp.weixin.qq.com/s/-K-6wLyQsia3yziWcF0i3g>

³ 信通院根据模型官网价格统计

⁴ <https://help.aliyun.com/zh/cap/use-cases/two-ways-to-quickly-deploy-and-experience-qwq-32b-reasoning-model>

⁵ <https://console.volcengine.com/ark/region:ark+cn-beijing/model>

2025 年 1 月发布《可解释、可通用的下一代人工智能方法重大研究计划 2025 年度项目指南》，国资委 2025 年 2 月部署深化中央企业“AI+”专项行动，推动更多大模型标志性成果和突破性进展，工信部等四部门 2024 年 6 月份联合发布《国家人工智能产业综合标准化体系建设指南（2024 版）》，推动人工智能标准化发展。

从地方政策看，各地因地制宜、各有侧重。北京发布《北京市推动“人工智能+”行动计划（2024-2025 年）》，充分发挥本市大模型创新动能强劲、算力保障多元、数据供给海量和应用场景丰富的综合优势，推动北京成为具有全球影响力的人工智能创新策源地和应用高地。上海发布《关于人工智能“模塑申城”的实施方案》⁶，瞄准技术前沿，坚持应用牵引，通过推进人工智能“模塑申城”，计划 2025 年底，建成一批上下游协同的赋能中心和垂直模型训练场。深圳则发布了《深圳市加快打造人工智能先锋城市行动计划（2025—2026 年）》⁷把握人工智能前沿发展态势，发挥深圳产业链完备、机电一体化、应用场景丰富等优势，加快打造具有全球重要影响力的产业科技创新中心。

从行业政策看，助力多行业领域创新发展，涉及制造、教育、医疗等。在制造行业，工信部发布《2025 人工智能+制造行动计划》，推动人工智能与制造业深度融合，形成“通用大模型+行业大模型”结合的智能生产体系。在教育行业，教育部启动《人工智能赋能教育行动》，推动人工智能与教学融合，提高全民数字教育素养与技能；在医药行业，国家药监局器审中心发布《人工智能辅助检测医疗器械（软

⁶ <https://www.shanghai.gov.cn/nw12344/20241227/670ecb743f954adfb4fa60aaf3c688b.html>

⁷ https://www.sz.gov.cn/cn/xxgk/zfxgj/tzgg/content/post_12052799.html

件）临床评价注册审查指导原则》，推动人工智能辅助检测类医疗器械落地应用。这些政策有助于推动大模型及人工智能在各领域的广泛应用和深度融合，为其发展创造良好环境。

3. 大模型市场广阔，落地应用需求旺盛

大模型市场发展前景广阔，无论是全球企业在生成式人工智能领域的支出趋势，还是国内大模型解决方案市场规模的增长态势，都展现出强劲的发展潜力。据 2024 年 IDC 预计，到 2028 年全球人工智能支出将实现翻倍增长，2024-2028 年内以 29.0% 的年均复合增长率（CAGR）攀升至 6320 亿美元⁹。企业对生成式人工智能技术的重视程度不断提高。据统计，国内大模型解决方案市场规模持续上升，预计到 2028 年 AI 大模型解决方案市场规模将达 211 亿元⁹，未来大模型解决方案将为企业数智化转型注入新活力。

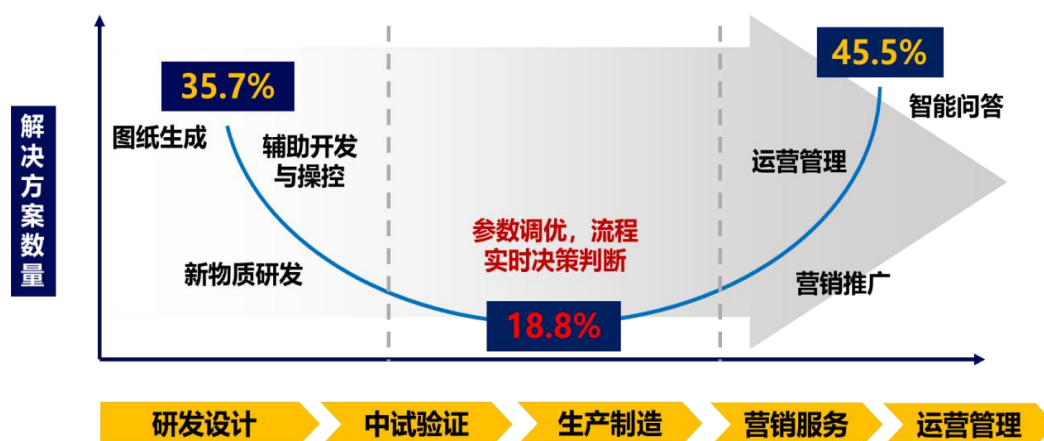
企业对大模型的应用需求快速升温，大模型在诸多行业逐渐落地应用，形成了以头部企业和头部场景为引领的头雁效应。根据 2024 年公开的大模型中标信息统计显示，中标项目数量和金额是 2023 年的十倍之多，且项目数量平均每月以 40% 的速率增长¹⁰。从采购方企业看，近六成为大型央国企，以运营商、金融和能源三大行业为主，中标项目金额超过总金额的一半¹¹。从场景分布看，高附加值场景更为集中，大模型应用场景呈现微笑曲线分布，“两端快、中间慢”特征明显。在研发设计、营销运营等高附加值场景率先发力，该类场景的

⁹ 《中国模型即服务（MaaS）及 AI 大模型解决方案市场追踪，2024H1》

¹⁰ 中国信通院调研统计

¹¹ 中国信通院调研统计

应用价值更加明确，可有效减少人力投入，实现降本增效，而中部的生产制造等低附加值场景因落地路径不清晰、投入产出不明确等因素，但这也恰恰是大模型落地应用的蓝海。



来源：中国信息通信研究院

图 1 AI 应用产业链分布图

（二）大模型落地存在诸多问题与挑战

随着大模型技术的快速更迭，以及各行业数智化转型需求的紧迫性，大模型落地的规模化已初见端倪。但是，企业在落地应用时仍面临诸多难题。

在顶层规划和需求分析方面存在盲目性。一是部分企业对大模型应用边界认知模糊，难以精准定位应用场景。企业难以判断大模型能否切实解决业务流程中的特定问题。二是企业面对市场上众多大模型产品，难以评估其在实际应用中的效果差异，也不清楚自身业务对模型性能的具体量化要求，导致在选择大模型产品时存在盲目性。

在落地过程中存在技术壁垒高筑情况。大模型的训练、调优与部署涉及诸多复杂技术难题，让许多企业望而却步。在训练方面，大模

型对数据和算力要求极高。高质量数据是训练出优秀模型的关键，但数据收集、标注、清洗和预处理工作不仅耗时耗力，还需要专业的知识和技能。以医疗领域为例，获取大规模高质量的医疗数据难度大，数据标注还需医学专业知识，成本高昂。在调优方面，因模型结构复杂、参数众多，如何针对特定场景进行有效调优是个难题。不同行业需求差异大，要使模型在不同场景下表现出色，需要深入了解模型原理和行业特点，选择合适的调优算法和路径，以达到最优调优效果。在服务部署方面，需考虑大模型服务的接口兼容性、安全性以及推理性能等问题。将大模型部署为可灵活调用的接口服务，一方面应考虑各业务系统的兼容性需求，另一方面应考虑服务的稳定性和安全性。

在大模型运营管理方面存在系统性挑战。一是大模型生产过程的管理，涉及训练、调优、评估、部署等全流程的标准化与自动化。当前技术瓶颈主要体现在两方面，其一，数据闭环构建困难，新数据采集、清洗及标注依赖人工干预，导致迭代效率低下；其二，持续优化机制缺失，模型动态适配市场需求的敏捷性不足，多数企业缺乏跨团队协同的技术中台支撑。二是大模型资产的管理，核心在于全生命周期价值激活。行业普遍面临三大问题，知识资产沉淀不足，导致模型能力复用率低；版本控制体系缺失，多版本并行引发维护成本激增；用户赋能机制不完善，缺乏针对性的培训体系与效果评估框架。

（三）大模型平台成为化解落地难题的关键力量

大模型平台在大模型的建、用、管方面发挥着重要作用。通过精准选型、高效训练、模型调优、部署交付、便捷调用、集成开发、持

续迭代、AI 资产管理和运营管理等一系列措施，大模型平台可帮助企业应对建设、使用和管理等方面的挑战，推动大模型在各行业的广泛应用。

1. “建”模型：筑牢大模型落地根基

大模型平台通过全流程技术支撑，系统性解决模型构建的核心挑战。在精准选型阶段，平台基于企业业务场景需求，构建多维度评估体系，结合行业基准测试筛选最优模型架构。例如，为高并发场景匹配轻量化推理模型，为高精度需求场景选择复杂结构模型，确保技术能力与业务目标精准对齐。在高效训练阶段，整合分散的异构数据资源，建立标准化数据清洗与标注流程，缩短数据准备周期。通过动态算力调度机制，并采用分布式训练框架，提升资源利用率，缓解算力瓶颈。在模型调优阶段，平台内置自动化优化工具，支持模型参数自适应调整与模型量化，帮助用户选择适合的调优路径，结合行业数据进行局部优化，快速适配特定场景需求。在部署交付阶段，平台提供异构硬件适配方案与模型压缩技术，确保大模型在云边端有效运行，同时通过标准化接口与系统集成方案，实现与企业现有业务系统的对接，保障部署后的服务稳定性和推理效率。

2. “用”模型：拓展大模型应用价值

大模型平台通过能力封装与场景化服务设计，降低技术使用门槛并释放业务价值。在模型调用阶段，平台提供统一的模型调用接口和开发工具包，企业无需具备深厚的技术背景，即可轻松调用大模型服务，同时平台对模型推理服务进行工程化优化，提高模型调用响应速

度和稳定性。在应用开发阶段，平台为不同技术储备的用户提供零代码、低代码和代码态多种开发环境，支持企业基于预置组件与行业模板，通过拖拽式操作完成复杂应用的搭建。

3. “管”模型：保障大模型可持续发展

大模型平台支持构建大模型的全生命周期管理体系，实现模型的可持续运营与合规治理。在持续迭代阶段，平台建立动态数据反馈机制，实时收集业务运行数据与用户交互信息，并通过自动化微调工具，定期优化模型，确保其持续适配场景需求。在 AI 资产管理阶段，平台对模型、服务、组件、应用等资产进行集中化管理，并通过权限控制与审计追踪机制，保障核心资产的安全性。在运营管理阶段，平台提供用户培训体系与效果评估框架，帮助企业建立从技术落地到价值转化的完整链路，消除用户使用障碍，提升模型的实际应用效能。

二、大模型平台落地策略与路径

大模型平台作为推动大模型规模化落地的有力支撑，基于技术门槛低、模型可共享、应用适配易等特点，对推动企业智能化转型、提升业务效率与竞争力具有重要意义。本章首先总结大模型平台建设落地的原则，在此基础上深入剖析落地步骤，并梳理归纳大模型平台通用落地框架。

（一）大模型平台落地原则

大模型平台落地是指企业通过建设模型开发层、模型服务层、应用开发层，将 AI 模型及其相关能力打包成可重复使用的服务，使企业能够快速高效地构建、部署、监控、调用模型。中国信息通信研究院通过对电信、科技、互联网等行业关于大模型平台落地实践的调查显示，大模型平台作为大模型落地应用的重要形式，通常遵循以下几个主要原则。

战略引领原则：以企业战略定位作为大模型平台落地的核心指引。企业需先明确自身长远发展规划，以及人工智能技术在其中的战略角色与定位，基于此制定大模型平台的落地策略与实施方案，确保平台建设与企业整体战略目标高度契合，以战略为导向驱动平台建设，保障平台发展的长期价值与方向。

因地制宜原则：从企业实际情况出发，全面评估 AI 基础、技术实力、人才储备、资金状况等要素，制定合理策略。一方面，充分利用现有 AI 基础设施与资源，避免重复建设，实现资源高效利用；另一方面，依据实际需求与预算，合理采购资源，优化大模型平台的功能。

能需求及服务部署能力，使平台建设贴合企业实际能力与发展需求。

标准化原则：参考行业标准规范开展大模型平台能力建设，确保平台的功能、性能及质量。在大模型生产调优、大模型部署服务化、大模型应用开发等过程，建议以行业标准为依据，减少摸索成本，提高平台建设质量与效率，保障平台在行业内的通用性与竞争力。

开放协同原则：积极引入外部模型与服务，拓展平台能力边界，丰富应用场景。加强与合作伙伴的协同合作，共同推动大模型技术的应用与发展，整合外部资源，提升平台的创新能力与市场竞争力，构建开放共赢的大模型生态。

安全性原则：围绕平台建设全生命周期的安全保障机制，降低各类风险。有效识别、评估并降低模型开发、服务和应用阶段面临的各类潜在风险，构建全方位、多层次的安全举措，确保平台安全稳定运行，切实保障模型、用户数据与业务安全。

持续改进原则：构建持续反馈机制和数据驱动流程，通过监控和维护闭环，持续优化平台性能。依据用户数据反馈，不断改进平台服务能力，确保性能稳定性与用户体验，实现平台的迭代升级与持续发展。

（二）大模型平台落地步骤

大模型平台的落地应用是一个系统化过程，构建科学合理的落地步骤，将为构建落地大模型平台提供良好指引，推动大模型产业健康有序发展。大模型平台落地过程可划分为评估诊断、方案制定、部署实施、持续优化四个关键步骤。

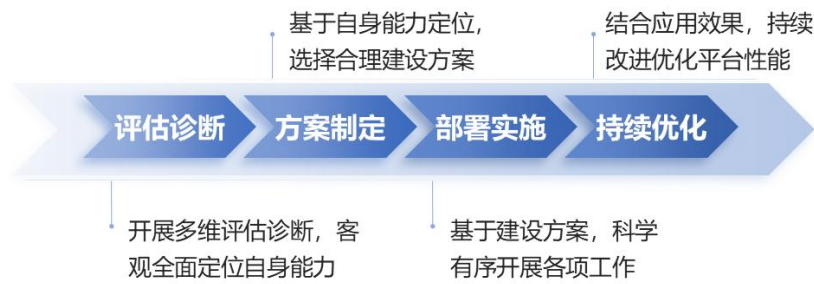


图 2 大模型平台地步骤图

开展多维评估诊断，客观全面定位自身能力。在诊断维度上，通过应用场景诊断，详细梳理企业现有业务场景及其规模，明确当前迫切需要解决的问题，确定大模型平台在企业内部落地必要性，梳理当前需求和未来规划；通过技术能力诊断，客观审视企业在人工智能领域的技术储备，明确自身是否具备研发、运营、维护大模型平台以及技术创新的能力；通过基础设施诊断，全面评估企业现有的算力资源、存储资源、网络资源等，完成大模型平台建设的基础准备；通过安全可信诊断，明确企业自身对安全可信的要求，如数据安全性要求、模型安全性要求、应用安全性要求、平台安全性要求等，从而便于制定安全保障机制。在能力定位上，依据战略规划、基础设施、技术储备、经费预算、安全要求等多方面情况，结合企业在大模型平台建设、技术应用等方面的具体表现，综合研判企业的智能化等级，可化为五个不同等级。

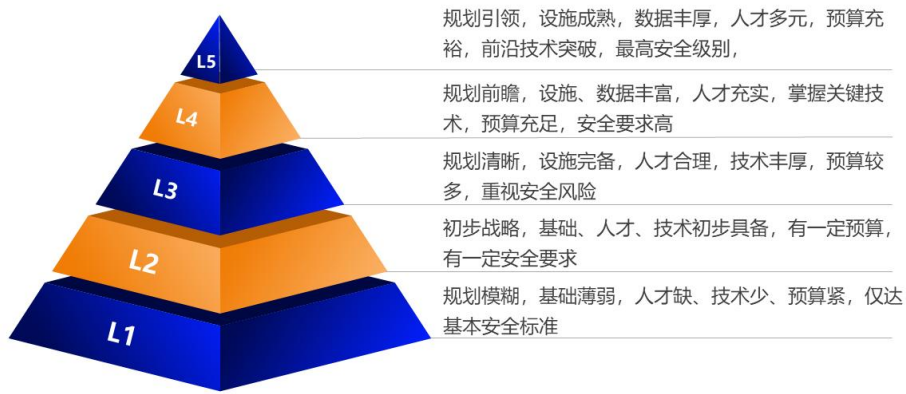


图 3 企业智能化能力等级图

基于自身能力定位，选择合理建设方案。L1 的企业适宜全部借助外部成熟力量，如公有云模式下的全栈大模型平台，快速生成大模型应用或接入现有业务系统；L2 的企业适合借助部分外部成熟能力，如公有云模式下的大模型服务平台，快速接入大模型接口（API）服务，企业内部私有化建设大模型应用开发平台；L3 的企业适合采购大模型平台，实现私有化的大模型微调和服务化部署推理，对内提供各类大模型服务及应用开发能力，以满足多样化应用开发需求和安全性要求；L4 的企业结合企业自身技术优势，对采购的大模型平台进行适当定制化开发，可进行大模型的重训与调优，为企业提供适合垂类行业或场景的大模型服务；L5 企业进行大模型平台的全栈自主研发，深度定制平台功能，提高与基础算力的适配，使其具备高度的可扩展性和灵活性，满足复杂业务深度需求。

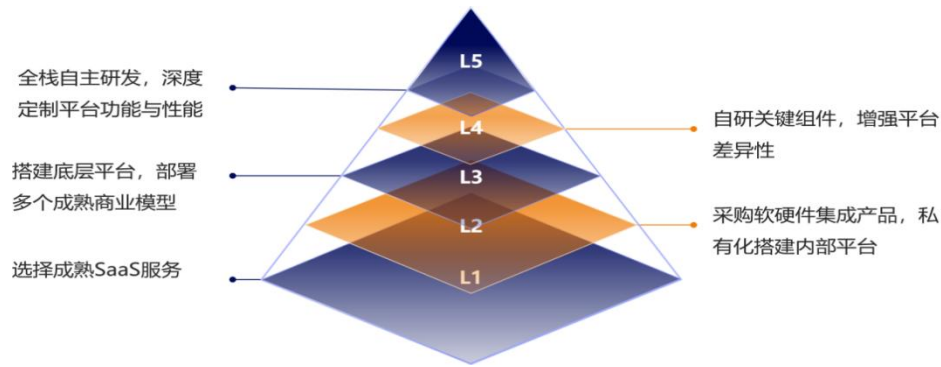


图 4 大模型平台建设方案图

基于建设方案，科学有序开展各项工作。一是精细规划，筑牢项目根基。明确关键时间节点，精确配置人、财、物等资源，并制定完善的风险管理方案；二是精确采购，夯实资源保障。依据企业能力等级与平台建设需求，开展针对性的软硬件采购工作，为后续工作提供坚实物质基础；三是严谨部署，保障平台质量。运用科学的评估体系和测试方法，从准确性、稳定性、安全性和可靠性等多维度对平台进行全面检测，确保平台可稳定、高效运行，满足企业业务需求；四是培训赋能，试点先行探索。组织专业培训，提升人员在大模型平台的认知和操作技能，使其能够熟练运用平台开展工作，并提供全方位的技术支持；五是优化拓展，实现全面推广。逐步将平台推广至更多团队，实现平台价值的最大化，助力企业全面提升业务效率 and 创新能力。

结合应用效果，持续改进优化平台性能。一是构建完善指标体系，强化实时监控机制。搭建一套全面、科学且有效的评估指标体系，涵盖模型平台性能与应用效果多个关键维度，并建立自动化监控机制，开展实时持续监测。二是收集分析多维数据，精确定位优化方向。定期收集分析监控数据、用户反馈数据、业务运营数据等多源数据，并

对数据进行深入分析，精确定位平台存在问题，明确后续改进方向。
三是建立维护更新机制，提升平台整体质量。

（三）大模型平台落地框架

大模型平台落地框架由三部分组成，包括模型开发层、模型服务层和应用开发层，每层可作为单独服务供用户调取，也可搭载下层设施提供更完整的服务。



图 5 大模型平台落地框架图

模型开发层为大模型定制开发提供数据、工具支撑，降低模型开发技术门槛。数据工程通过多源异构数据集成、标注优化，结合数据回流机制，为模型训练调优筑牢根基；模型调优借助多元调优工具、强化学习策略，融合多维评估体系构建闭环优化机制，实现模型性能提升契合不同业务需求；模型交付运用量化、蒸馏等算法实现模型压缩与多端适配，协同多环境部署方案及流程优化策略确保模型稳定高效落地；提示词工程借助预置提示词模板与在线实时提示词调优技术，降低用户使用门槛，保障模型生成内容质量。基于上述功能，模型开发层推动大模型高效开发与应用，为企业智能化转型注入强大动力。

模型服务层以大模型为主体，为用户提供可直接调取的大模型服

务，确保服务便捷稳定。模型服务部署实现对大模型的加速、封装、资源配置、服务部署，加速模型落地应用；模型服务测试集成模型能力、服务性能、服务风险等测试，提供模型测试与选择的能力，确保用户选择到满足需求的模型；模型服务管理严格把控模型服务的发布、检索、下线，保证模型服务的高质效、可持续；模型服务运营采用全链路运行监控机制，结合数据收集政策，切实保障模型服务稳定高效运行。

应用开发层完成基于大模型的 AI 应用开发，直接触达应用场景。应用构建涵盖代码态、低代码态、零代码态三种实现路径，匹配差异化场景需求；能力扩展配备丰富组件库、专业知识库、强大记忆能力，提升应用多样性、专业性与交互连贯性；应用调试提供灵活调试配置、实时效果预览及详细数据统计，助力优化应用性能和体验；发布及运营支持一键快捷发布、用户多渠道反馈及数据深度分析，驱动应用持续迭代升级。

运营管理聚焦全生命周期治理，构建“生产-运营”双轮驱动体系。生产过程管理通过自动化流水线实现应用持续迭代，集成代码审查、性能优化与安全合规工具，确保版本升级的稳定性和可靠性，同时结合用户行为反馈动态调整功能模块与模型组合策略，强化场景适配能力；运营管理搭建多维评估体系，融合业务价值指标与技术性能指标，通过策略对比优化资源配置，并建立用户反馈响应机制，将问题分类映射至知识库更新、交互逻辑优化等改进方向，形成“问题识别-根因追溯-闭环修复”的敏捷运维链路。

三、大模型平台关键能力建设

大模型平台是推动大模型技术落地应用的核心支撑。本章从模型开发层、模型服务层、应用开发层解析平台核心技术架构。模型开发层突破数据处理与模型调优技术瓶颈，模型服务层提升模型服务性能与运营能力，应用开发层降低场景化应用开发门槛。同时结合国内外平台实践，解析技术工具与行业需求的深度融合，为大模型从技术研发走向规模化应用提供参考。

（一）模型开发层

1. 聚焦解决的关键问题

模型开发层旨在突破数据处理和模型开发技术方面的瓶颈。一是解决数据处理效率低下且质量不稳定、多模态数据融合困难等问题；二是缓解模型架构设计复杂带来的训练及调优等技术门槛问题。

2. 关键技术及能力要素

数据工程：在数据处理方面，模型开发层支持各类数据源接入，能高效采集结构化数据库数据以及文本、图像、音频等非结构化数据，同时通过数据标注、数据清洗去重等能力，统一管理标注数据集，并提升数据集质量。在数据回流方面，支持业务数据回流至数据集，为模型持续优化训练提供新数据，助力模型不断进化，更精准地适应业务变化，持续提升性能。

模型调优：在模型精调方面，模型开发层为开发者提供丰富的调优工具，实现自动化或半自动化调优，并针对不同行业特点，基于私

域数据集重训模型，从而生成专属模型，同时借助人类反馈数据开展强化学习，支持多种强化学习框架及策略，使模型得到不断优化。在模型评估方面，支持从准确性、稳定性、泛化能力等多个关键维度对模型进行严格全面的评估。通过分析评估结果，精准定位模型的优势与不足，为进一步精调提供有力依据，形成优化闭环，持续提升模型性能，满足多样化业务需求。

模型交付：在模型压缩方面，模型开发层提供了量化、蒸馏、剪枝等多种压缩方法，且支持预估不同算法下的压缩效果，帮助用户在保持模型性能的前提下，尽可能地减小模型体积，使得模型能够在更多类别的计算设备上适配运行。在模型部署方面，支持多种环境下的快速部署，将大模型服务化，并对其准确性、稳定性等各类指标进行测试评估，确保上线后的模型能够稳定运行。

提示词工程：在提示词管理方面，平台通过对提示词的有效管理，预置了多种丰富的模板，涵盖不同领域和应用场景，极大地方便了用户快速调用，满足多样化需求。部分先进平台还具备自动生成提示词的功能，能依据用户的简单描述或任务类型，智能生成初始提示词。在提示词优化方面，挖掘潜在的优化方向，提升引导大模型生成内容的质量。此外，在线调试能力让用户能实时调整提示词，即刻获取更符合需求的结果，降低使用门槛，助力不同技术水平的用户都能更好地利用大模型。

（二）模型服务层

1. 聚焦解决的关键问题

模型服务层旨在解决模型服务选型、保障服务稳定性和可靠性、服务管理运营等方面的难题。一是解决企业选择大模型服务时难以对比效果，并通过工程化方法助力各类大模型推理加速；二是解决用户管理和权限控制、服务监控与故障排查、计费模式等方面的问题，以提升大模型服务的高质效应用。

2. 关键技术及能力要素

模型服务加速：在基础架构方面，通过异构资源协同与动态资源调度，构建多参数规模模型的算力支持体系，提升资源利用率；在模型优化方面，结合量化、剪枝等轻量化技术及软硬件协同加速策略，降低推理延迟与算力消耗；在适配性方面，支持主流框架与多硬件平台的模型封装，标准化 API 设计兼顾跨平台兼容性与安全性；在资源管理方面，提供按需弹性扩展与可视化配置界面，支持基础、自适应及精细化计费模式；在部署能力方面，实现批量更新、负载均衡与服务降级策略，保障复杂环境下的高可用性与快速迭代。

模型服务测试：在评测维度方面，支持基于公开或私有数据集的多指标对比测试，覆盖响应时延、吞吐量、成功率等核心性能参数；在风险验证方面，开展数据安全过滤、服务稳定性及合规性专项测试，降低潜在风险；在流程闭环方面，通过自定义指标设计与批量评测机制，形成测试结果分析与优化反馈链路，为服务调优提供量化依据。

模型服务管理：在发布管理方面，建立审核机制与多策略部署流程，支持版本回滚与发布信息追溯，确保服务规范性；在服务查找方面，提供分类标签、关键词及智能联想功能，提升服务资源定位效率；

在标准化治理方面，通过文档更新与全生命周期管理工具，保障服务信息的完整性与可维护性。

模型服务运营：在监控体系方面，实时统计调用量、流量及服务状态，结合数据大盘与趋势分析实现效能可视化；在运维支持方面，提供故障报警、弹性扩缩容及自动化修复能力，确保服务高可用性；在数据闭环方面，采用加密存储与敏感信息过滤技术，支持用户反馈与业务数据回流，驱动服务持续优化；在安全管控方面，通过权限分级与审计功能，实现服务全流程的精细化权限管理与操作追溯。

（三）应用开发层

1. 聚焦解决的关键问题

应用开发层旨在解决模型准确性提升难、系统融合复杂和开发技术门槛高等难题。一是解决直接使用大模型遇到准确性不足的问题，通过工程化方式提升大模型落地的可用性；二是解决大模型应用与原有系统融合的问题，推动大模型在企业已有的业务中发挥价值；三是解决开发大模型应用面临的高技术门槛问题，推动大模型应用的规模化、普适化发展。

2. 关键技术及能力要素

应用构建：应用构建包括三种主流方式，一是代码态，提供可编程的完整开发套件，支持多种开发环境、语言和框架，便于专业开发者深度定制应用。开发者能利用各类组件与第三方 API，依据复杂业务需求，实现精细化功能设计和业务逻辑构建；二是低代码态，拖拉

拽式 workflow 支持低代码灵活编排，能构建复杂且可控的任务流程，以可视化操作搭建业务场景，降低开发门槛，提升开发效率，减少开发成本；三是零代码态，支持通过自然语言创建智能体（Agent），操作简单便捷。同时，三步即可完成应用创建与分发，普通用户无需编写代码，借助开放的工具组件和应用框架，轻松构建大模型应用。

能力扩展：在组件方面，为开发者提供丰富组件资源，涵盖多种功能类型，如搜索、代码解释器等组件，可满足不同业务场景下的应用开发需求，助力实现设备的多样化功能；在知识库方面，可接入多种格式数据，进行抽取、切片、索引构建等操作，部分还支持多样化构建策略和周期性数据接入，能够不断丰富知识储备，为模型提供专业、精准的知识支撑，提升模型回答的准确性与专业性；在记忆方面，在记忆方面，为应用开发提供强大的记忆支持能力。它支持记忆变量、记忆表和记忆片段，可灵活记录多种信息，让应用能更好地留存关键数据，提升交互的连贯性与智能化水平。

应用调试：在调试配置方面，平台提供丰富的配置选项，支持开发者对应用的各项参数进行灵活调整。无论是接口调用的参数设置，还是与外部系统对接时的配置信息，都能方便地进行修改和优化，以满足不同业务场景的需求；在效果预览方面，具备实时预览功能，开发者可以随时查看应用在不同输入条件下的运行效果。通过模拟真实用户操作，直观展示应用的界面展示、交互逻辑以及数据处理结果，便于及时发现并解决可能存在的问题，提升应用的用户体验；在用量参考方面，提供详细的用量统计数据，包括模型调用次数、数据流量

消耗等关键指标。

发布及运营：在发布方面，提供便捷且多样化的发布方式，支持一键式快速发布，同时针对不同的应用类型和部署环境，提供个性化的发布配置选项。开发者可以轻松选择将应用发布到公有云、私有云或者混合云环境；**在使用及反馈方面**，支持收集用户在使用过程中的反馈信息，具备多渠道反馈机制，如在线反馈表单、应用内反馈入口等。通过对用户反馈的及时收集和整理，为后续优化提供依据；**在迭代调优方面**，基于用户反馈和应用运行数据的深度分析，平台助力开发者快速定位问题并进行针对性优化。支持对应用功能进行快速迭代，更新模型算法、改进界面设计等。

（四）大模型平台产品实践

大模型平台作为连接技术与应用的重要桥梁，落地需求逐渐增多，落地场景日趋丰富。当前行业中大模型平台在整体发展、主流平台特点、价值体现以及未来发展趋势等方面呈现出鲜明的特征。

从整体数量来看，截止 2025 年 3 月，根据公开数据显示，已涌现出 70 余款大模型平台工具产品¹²，众多企业积极布局，竞争较为激烈。从各产品特点来看，百度智能云的千帆模型开发平台和千帆应用开发平台，分别聚焦模型调用、模型开发与应用开发，功能明确；亚马逊的 AWS Bedrock 平台在模型开发、模型服务和应用开发工具方面提供了全面支持；微软的 Azure AI Studio 提供完整的模型开发到应用部署链路。从种类分布来看，大模型平台涵盖了大模型开发、大模

¹² 中国信通院根据公开信息统计

型服务、大模型应用开发、大模型管理四大子平台方向，据中国信通院 2025 年 3 月份问卷调研统计，大模型开发子平台作为基础数量相对较多，占比约 30%，为领域模型的调优提供支撑；大模型应用开发子平台则面向市场的应用需求，助力开发者快速搭建应用，占比也近 30%，大模型服务子平台向用户提供便捷调取的模型服务 API，模型管理子平台协助公司统一管理和运营 AI 资产，二者各占比 20%。

1. 百度智能云千帆大模型平台

在大模型技术快速发展的进程中，百度智能云千帆大模型平台不断迭代升级以适应行业需求，最大限度地降低客户和伙伴的试错成本，让整个创新更容易。该平台提供包括模型开发、模型服务和应用开发在内的一体化工具能力。在模型开发层提供最全面的模型开发工具链，包括深度思考模型开发工具链、多模态模型开发工具链、模型蒸馏工具链等，支持一站式模型开发，快速实现效果调优；在模型服务层，提供了 100 多个模型，既包括文心大模型 4.5 Turbo 等一系列百度自研模型，也包括像 DeepSeek、Llama、通义、Vidu 等第三方模型，来灵活满足用户需求；在应用开发层提供企业级大模型应用开发平台和端到端的应用开发工具，降低开发门槛，满足企业客户各类应用开发需求。



图 6 百度智能云千帆大模型平台 3.0 架构图

平台支持 DeepSeek 系列模型调用、精调及蒸馏,预置 RAG/Agent/ 工作流等大模型应用开发工具链,在能源、制造、教育、企业服务等 行业加速企业级大模型应用落地。同时千帆应用开发平台在国内率先 支持 MCP 协议,并正式推出千帆企业级 MCP 服务,第一批已经有 1000 多个 MCP Servers 接入了这个平台,而且还在持续地扩展中,促 进工具生态的高效共享。

2. 微软 Azure AI Studio

微软 Azure AI Studio 在生成式 AI 的发展中不断升级,提供高效、 安全的开发体验。最初专注于模型快速部署和基础开发,为企业提供 安全可靠的平台。随着技术的发展,平台增加了丰富的模型资源和强 大的开发工具,降低了开发难度。随着生成式 AI 在各行业的广泛应 用,Azure AI Studio 的使用量和应用场景大幅增加。平台升级为 Azure AI Foundry,支持模型探索到部署的全流程开发。

Azure AI Studio 提供完整的模型开发到应用部署链路。模型开发层提供全流程工具，从数据准备到模型微调，通过 Prompt Flow 简化开发流程，并支持模型性能对比。模型服务层整合了 OpenAI、微软等系列模型，提供超过 1800^[1]个预训练模型，并支持统一 API 和 SDK（Software Development Kit，软件开发工具包）集成。应用开发层提供端到端工具，支持企业级应用构建和部署，利用 Azure AI Foundry 的项目管理功能，高效组织团队协作，实现从概念验证到生产部署的全过程。

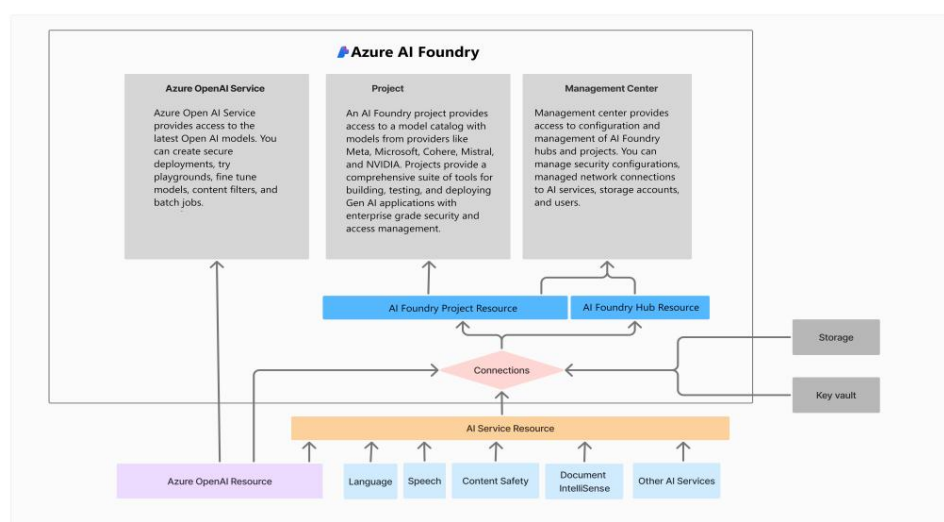


图 7 微软 Azure AI Studio 示意图¹³

3. 亚马逊 AWS Bedrock

AWS Bedrock 作为亚马逊生成式 AI 平台，自推出以来经历了多次功能更新和版本迭代。其核心优势在于提供统一 API 接口，让开发者能够轻松访问来自 AI21 Labs、Anthropic 等领先 AI 公司的多种基础

¹³ Azure AI Foundry architecture - Azure AI Foundry | Microsoft Learn

模型。随着市场需求的增长，Bedrock 不断扩展其模型库，并优化服务器架构，支持无服务器部署，从而加速 AI 应用的开发和部署。

AWS Bedrock 平台在模型开发、模型服务和应用开发工具方面提供了全面支持。在模型开发层，Bedrock 通过 Amazon SageMaker 等工具，为开发者提供了强大的模型开发工具链，支持一站式模型开发，并实现数据、模型和算力资源的统一管理和调度。在模型服务层，Bedrock 通过无服务器架构简化了模型部署，开发者无需管理复杂的基础架构，即可快速实现模型的调用和优化。在应用开发方面，Bedrock 提供了丰富的 API 和开发工具，支持端到端的应用开发，帮助开发者快速构建生成式 AI 应用。



图 8 亚马逊 AWS Bedrock 平台示意图¹⁴

¹⁴ AWS | Community | Deep Dive into AWS Bedrock: Simplifying Access to Foundation Models for Enterprises

四、基于大模型平台的实践案例分析

在数字化转型的浪潮中，大模型平台在各行各业中逐渐发挥出其关键作用，并涌现出诸多优秀实践，进一步证明了大模型平台在产业效率重构方面的普遍价值，并形成了“技术基础+场景适配+价值闭环”的转型模式。

（一）互联网电商行业案例

梦饴科技旗下爱库存专注于商品分发智能解决方案平台，为各类电商提供电商软件工具、供应链服务、私域运营标准作业流程（SOP）等能力。随着业务规模的持续增长，该企业希望利用大模型对传统业务进行重构，帮助爱库存提高运营效率，降低成本，提升用户体验，以响应激烈的市场竞争，但是在落地过程中存在如下问题。

1) 大模型缺乏电商行业相关的领域知识，不能有效与业务系统进行交互，在复杂业务场景中容易出现幻觉；

2) 单一智能体无法支撑多元化的业务需求，亟须建立一个能让业务人员和开发人员快速便捷进行智能体搭建的平台。

1. 解决方案

针对当前问题，该企业基于千帆大模型平台打造了“爱库存超级助手”，基于平台配置工作流，设计确定性更高的智能体流程，从而有效提升了智能体在复杂业务场景中的识别准确性。此外，借助大模型平台中的 RAG 知识外挂及工具调用等功能，强化智能体的业

务理解和与现有系统交互的能力。

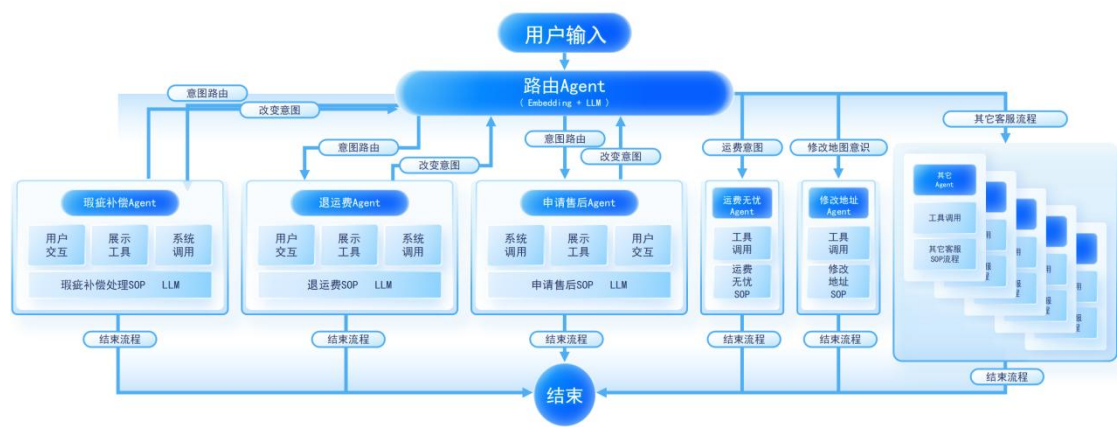


图 9 爱库存超级助手架构图

2. 价值和收益

基于大模型平台搭建的智能体服务成功在智能客服、瑕疵品识别等场景成功落地。其中，智能客服场景通过多智能体协作，实现了 70% 以上的问题解决率。瑕疵商品识别场景通过使用智能体将识别准确率提高至 90%，有效提升了客户满意度。

（二）金融行业场景实践案例

泰康保险集团积极探索数字化水平提升之路，基于大模型在多个业务板块搭建了应用，进一步增强用户体验，推动了保险行业的数字化转型。在推动过程中面临的行业痛点问题长期存在。

1) 保险的业务系统复杂,技术架构需具备高度的灵活性和适配性，以应对各种复杂的业务流程和系统架构。

2) 保险业知识资源庞大，但知识的统一建设、高效获取、价值激活还需要进一步探索。

1. 解决方案

在此背景下，泰康保险集团基于大模型平台的大模型服务、端到端应用开发工具链，搭建了统一的知识-工具-模型底座，面向总子公司提供创新的标准化、高可用、高性能、高精度的 AI 应用。

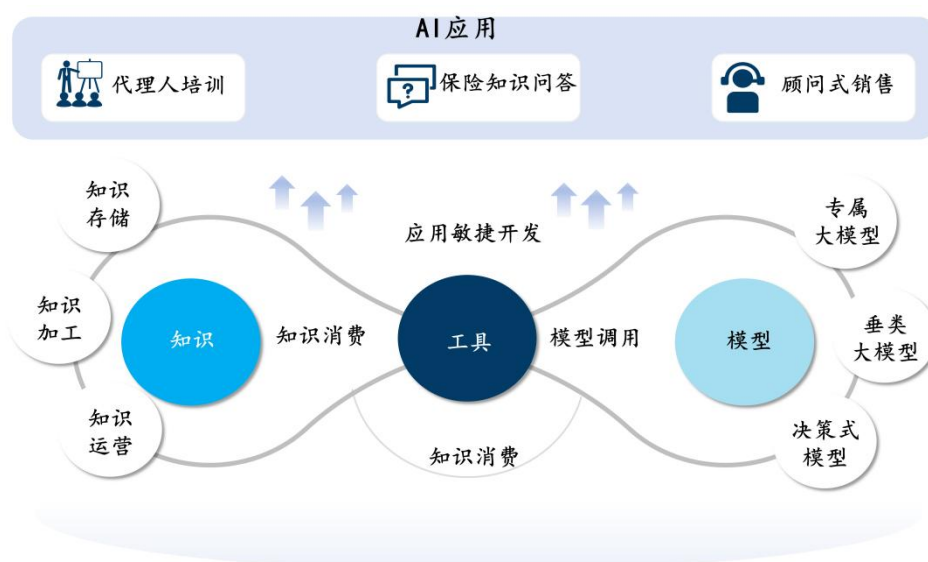


图 10 泰康保险集团基于大模型平台的场景示例

2. 价值和收益

在知识构建及价值激活方面，构建泰康保险集团的知识网络，接入了 9 个场景中海量知识。在模型及应用构建方面，基于大模型平台能够对全集团数据进行自动挖掘、解析和加工，形成精细化知识点，大幅提升知识资产价值。在应用效果方面，已在 3 个场景落地使用，通过语言交互即可实现高效搜索，大幅提高知识获取和办公效率。

（三）教育行业案例

考试宝是上海巨闲网络科技有限公司推出的职业考试在线学习

服务平台，其业务涵盖建筑工程、安全生产、特种作业、考公考编、医药卫生等 1 万多个职业考试培训类型。该平台用户量的大规模增长带来了多重挑战。

1) 运营成本高，该平台题库涉及 10000+行业，试题数量庞大，且存在诸多长尾行业场景，其题库上传、更新、解析等运营成本高昂。

2) 用户体验不佳，该平台日均新增 2000 万道用户上传的试题亟需专家答疑解惑，由于题目量巨大，专家资源紧缺，无法及时为用户反馈结果导致用户体验差。

1. 解决方案

针对上述行业痛点，该公司立足海量用户数据与丰富应用场景，基于大模型平台上丰富的大模型和调优工具，创建了适合多行业场景的试题解析应用和 Prompt 模板，实现解析生成、试题讲解、错题精选、考点总结等功能，帮助用户提升学习效率，同时提升用户体验。



图 11 考试宝工具大模型平台一站式应用架构

2. 价值和收益

考试宝通过大模型平台的落地，使得其运营效率和成本控制上实现了重大突破，同时考试宝用户规模和付费率也因此大幅增长，为在线教育行业树立了标杆。

1) 效率成本，试题解析效率提升 98%，阅卷评分效率提升数百倍，内容生成成本从每题 1.5 元降至 0.03 元，效率提升超过 1000 倍。

2) 用户体验，用户日均使用时长减少 40%，搜索准确度提升 30%，大幅增强用户黏性。

3) 商业效益，AI 技术的引入带来的付费用户和日活跃用户增长率达 100%，显著提升了平台的市场竞争力和品牌影响力。

（四）企服行业案例

随着 AI 技术快速渗透到智能办公场景，用户对于智能办公的应用需求持续增加。金山办公的 WPS Office、WPS 365 等办公助手产品面临巨大智能化转型压力，其传统产品的应用存在诸多痛点。

1) 内容创作效率低下，在制作 PPT 时，从构思大纲到设计排版的过程耗时较长，用户通常缺乏设计能力，在模板选择、配色等方面存在困惑，导致创作效率低下。

2) 文档质量难以保证，文档中的语法、拼写错误不易被发现，且整体风格难以达到专业水准，影响了文档的质量。

3) 多语言处理困难：涉及外语文档或跨国交流时，翻译缓慢且不准确。

1. 解决方案

针对上述一系列问题，WPS AI 通过大模型平台接入所需的大模型，使得用户可通过自然语言即可解决内容生成、内容理解、指令操作等方面的难题。

1) 技术接入与平台支持，通过大模型平台接入大模型服务，为产品提供了基础技术支持，同时借助平台为应用提供多重安全防护措施能力，确保生成内容及用户数据的隐私与安全得到保障。

2) 内外循环迭代助力场景落地，采用大模型平台接入、Prompt 提示词工程、精标数据驱动的三步走策略，构成了外部能力接入和内部能力提升的双循环，结合实际办公场景，实现大模型能力的共建、共创和落地交付。



图 12 WPS AI 大模型平台应用架构

2. 价值和收益

基于大模型平台优化后，WPS AI 大幅减少了文档和 PPT 的制作时间。用户通过与 WPS AI 的自然语言交流，轻松处理复杂办公

任务，使办公更智能，进一步提高了企业的运营效率 and 创新能力，更好地适应数字化时代。

（五）智能硬件行业案例

深圳酷旗将 AI 技术广泛应用于智能家居、智能教育、消费电子等领域，旗下生态产品已覆盖 95% 以上的生活与生产场景，然而在 AI 与物联网（IoT）融合的过程中，存在以下问题：

1）小模型聊天体验差。日常生活中互动场景种类多，而传统机器学习模型专注细分领域泛化性差，导致用户与 AI 硬件互动效果不佳。

2）大模型效果和服务性能难以平衡。在对话场景下中英混 TTS 效果不佳，答案准确度与延时无法同时达到完美状态。

1. 解决方案

酷旗引入大模型平台能力，基于平台的一站式大模型精调与调用服务，将大模型能力对接至 IoT 设备，为各类智能家居、智能硬件提供全面的解决方案。同时，酷旗基于大模型平台可进行超级智能体搭建，在 IoT 设备上实现端侧大模型应用落地。



图 13 智能硬件大模型应用架构图

2. 价值和收益

该解决方案通过大模型平台与 IoT 的深度融合，实现了三重产业升级：首先，基于大模型平台的于端云协同架构提升了 IOT 设备智能化响应速度，同时保证了中英混合 TTS 准确率，突破交互体验瓶颈；其次，通过大模型将智能体开发周期大幅缩短，助力服务机器人、AI 耳机等产品实现精准语义理解与多模态交互；第三，基于大模型平台可快速构建 Agent 应用，形成“模型即服务-设备即终端-场景应用”的 AIoT 闭环生态，推动智能硬件行业进入实时响应、自主进化的新阶段。

（六）制造行业场景实践案例

山东省港口集团有限公司是山东省属国有重要骨干企业，已初步建设信息化及数字化的基础设施，但随着业务规模增长，复杂性和不确定性增加，亟需推进以大模型技术赋能港口智能化升级，落地过程中也面临诸多挑战：

- 1) 大模型项目技术复杂度高,需要强大的技术实力和人才储备。
- 2) 大模型项目对于数据质量问题、归集慢、预处理时间长也带来困难。
- 3) 港口数据敏感,必须严格遵守隐私和安全法规。

1. 解决方案

山东港口通过建设大模型平台推进大模型技术赋能山东港口智慧化转型升级,为各场景应用提供基于大模型的原生应用。

1) 通过大模型平台引入公有云大模型,建设山港集团大模型统一门户,支持港口私有知识问答,赋能数智化办公,支持多轮对话管理、信息来源定位功能;

2) 基于大模型平台快速搭建智能化应用,建设智慧运营助手,结合港口多年数据治理的数据积累,采用语音、文字等交互方式,对各级组织生产经营指标进行多维度快速问询和总结分析;

3) 建设大模型试点应用所需算力资源池,确保为大模型微调、推理提供充足的算力支持。



图 14 山港大模型框架

2. 价值和收益

该方案实现港口智能化三重突破：其一，构建港口知识中枢，实现专有数据查询响应速度的提升，同时数据归集预处理耗时缩短，破解数据治理难题；其二，实现港口主要生产经营指标的一问知全局，为管理者提供智能决策辅助功能；其三，打造全国首个港口调度大模型，通过生产要素动态建模使车辆调度匹配率提升，形成“数据治理-智能决策-生产优化”的港航大模型中枢系统，为全球港口行业提供可复制的 AI 工业化落地范式。

五、总结与趋势

在大模型时代蓬勃发展的浪潮中，各行业正面临前所未有的数智化转型挑战，同时也迎来了与之并行的历史性机遇。在此背景下，大模型平台凭借其强大的工程化能力，为行业转型与变革提供了坚实支撑与强劲动力。

大模型平台围绕“建、用、管”三个维度有效化解大模型落地难题。通过“建”模型提供行业模型或场景模型的生产方式，通过“用”模型提供大模型与应用有效结合的落地方法，通过“管”模型提供有效管理模型、组件、应用等 AI 资产的渠道。基于“建、用、管”的思路，大模型平台为解决大模型落地挑战提供了良好的解决方案，推动化解大模型选择难、大模型技术壁垒高、大模型运营和管理复杂等系列问题。

大模型平台以三层落地框架构筑大模型应用路径的有力支撑。大模型平台的落地路径通常包括评估诊断、方案制定、部署实施和持续优化四大阶段，企业在落地过程中根据自身能力选择不同的落地方法。而大模型平台的落地框架包括模型开发层、模型服务层和应用开发层，为大模型落地应用的各阶段提供全方位的平台能力。

大模型平台应用日趋广泛落地效果逐渐显现。一方面大模型平台产品日益增多，平台能力侧重各有千秋，为不同的应用场景提供了丰富的基础；另一方面，在外需和内驱双向推动下，互联网、金融、教育、制造等行业涌现出诸多优秀的大模型平台落地实践，并为业务带来了实际价值。

但当前大模型平台发展仍处于初期阶段，未来将从技术、应用和生态角度持续发力，且安全可信能力将持续跃升，为大模型落地应用及全行业智能化转型带来全新动力。

技术方面：大模型能力持续拓展，大模平台能力将随之提升。大模型平台与大模型能力相辅相成，一方面大模型自身能力将从深度和广度持续演进，从参数规模、长链路推理方面纵向提升模型推理能力，从多模态、具身智能的维度扩大模型与物理世界的连接；另一方面大模型平台将提供更强大的多模态服务、加速服务和应用开发服务能力，以支撑大模型的能力提升，包括更简单快捷、更低门槛的方式支持多模态的训练调优和应用，同时支持更优的模型推理加速能力，并为多 Agent 等应用的建设提供强大的平台支撑。

应用方面：助力数据飞轮快速运转，企业级应用将迎来大爆发。一是通过平台能力，将大模型与各业务系统或工具紧密结合，推动数据飞轮的运转，使大模型推理与业务系统或客户的反馈形成互馈，从而充分发挥和挖掘业务价值，助力场景化应用的扎根落地。二是基于 Agent 框架的智能应用飞速发展，单 Agent 能力持续提升，同时面对更具开放性、更复杂的任务时，多 Agent 协作将成为主流，企业级应用新生态即将形成。根据 Gartner 预测，到 2026 年，超过 80% 的企业将使用生成式 AI 的 API 或生成式 AI 模型，或者在生产环境中部署支持生成式的应用。

生态方面：大模型平台将更加开源开放，推进生态新发展。首先基于大模型的应用开发新范式将持续演进，通过自然语言和跨模

态快速生成特定领域的大模型原生应用，开发门槛大幅降低，这将吸纳更广泛的用户群体参与其中，推动应用开发生态持续扩大。其次围绕三横一纵的平台架构构建标准体系推动生态健康有序发展，围绕大模型开发、大模型推理部署、大模型应用构建、大模型运营管理各维度，制定统一的技术标准和接口规范，促进不同系统和应用之间的互联互通，降低集成和迁移的成本，为生态的可持续发展提供保障。

安全方面：围绕安全可信继续筑牢应用安全围栏。未来将从平台安全、数据隐私保护、模型可控性、服务可靠性及应用安全等维度，持续健全大模型平台安全保障体系，根据 Gartner 预测，到 2026 年，对 AI 应用部署信任、风险和安全管理控制措施的企业，将能过滤掉 80% 的错误和非法信息，从而提高决策的准确性¹²。通过健全的保障体系，保护用户在使用模型服务过程中的数据安全与隐私权益，筑牢模型服务风险管理防线，帮助企业生产更加负责任的 AI，推动各行业“AI+”行动进展。

扫一扫
获取更多资料



百度智能云公众号



百度智能云服务号

北京百度网讯科技有限公司
中国信息通信研究院人工智能研究所

附录

(一) 教育行业场景实践案例

华北电力大学通过教育数字化转型，不断提升学校治理能力和师生满意度，为更好解决服务师生的“最后一公里”问题，实现信息化应用“简约高效，精准连接”，华北电力大学针对以下场景应用大模型技术进行优化和升级。



图 15 华北电力大学

1.需求痛点

华北电力大学是国内较早开展智慧校园建设的高校,功能齐全,有流程 200 余项,应用 100 余个,过去主要是在官网和移动端,以集成门户形式呈现,在师生实际应用过程中主要存在以下问题:

- 1)师生获取校园服务信息需要通过多个系统和菜单,操作复杂,效率低下;
- 2)系统后台关键词与用户输入可能不匹配,导致用户搜索特定

信息可能无法找到结果；

3) 师生在查询奖学金、预约盖章、开发票、请假报批等信息时，需要在多个系统间切换，耗时耗力。

2.落地方案

针对以上存在的问题，2024 年华北电力大学与百度智能云率先合作，应用千帆应用开发平台进行联合开发，得益于千帆应用开发平台支持编排外部接口，优质的自主规划编排能力，能够快速高效地将外部 API 服务编排成为专属插件，可快速完成应用开发。平台同时也能支持线上调试，当效果满足预期即可快速发布应用。“i 华电”智能助手仅用三周即实现上线。

1) 通过百度智能云千帆应用开发平台，华电 AI 助理将学校已有的各信息化系统的相关信息集纳成一套可供 AI 学习、理解的文本“知识库”，以此解决师生获取服务信息操作复杂、效率低下的问题；

2) 利用百度智能云千帆应用开发平台接入大模型能力，AI 助理能够立即理解用户的意图并精准给出结果，解决系统后台关键词与用户输入可能不匹配导致用户搜索特定信息可能无法找到结果的问题；

3) 通过 AI 助理的智能服务，师生只需一句话描述需求，就可以快速获取所需信息和服务，极大地减少在多个系统间切换的时间，

解决了师生在查询奖学金、预约盖章、开发票、请假报批等信息时需要在多个系统间切换、耗时耗力的问题。



图 2「i 华电」应用架构

3.实践效果

“i 华电”智能助手上线后，获得师生的广泛好评，在新生引导的场景中可以直接向“i 华电”提问，比如学校有哪些学院、奖助学金的政策、体测要求、四六级考试、图书馆借阅等，不同于在网上开放式搜索的答案五花八门，“i 华电”给的答案就是基于大学自建知识集，准确可靠。在校务管理场景中，过去充饭卡要去移动微校园找到点进一级入口、二级入口，流程较为繁杂，现在直接在手机上打几个字，就可以在“i 华电”中完成支付。通过“i 华电”智能助手，一是师生可以快速获取各类信息和服务，减少了在多个系统间切换的时间，提高了工作效率；二是 AI 助理能够理解师生意

图，提供个性化的服务推荐，提升了用户的使用体验；三是“i 华电”本身也在进化，逐步强化能力，丰富应用，推进 AI 在人才培养、科学研究等各个领域的深化应用，加快教育数字化转型发展，不断塑造学校高质量发展的新动能和新优势。



图 3 华北电力大学「i 华电」AI 应用

(二) 电商行业场景实践案例

返利网是一家电商导购平台，为消费者提供消费优惠服务。返利网业务探索过程中发现，大模型对于用户需求意图理解有着巨大优势，尤其是经过精调后模型，返利网通过大模型能力探索推出个性化推荐和全网比价系统，助力用户高效做出明智购物决策，实现大模型技术在电商导购行业实现新的发展。

1.需求痛点

在当前电商导购服务中，返利网识别出几个关键的业务痛点：

1) 用户购物决策周期长：由于商品信息复杂多样，用户在选购时常常难以迅速做出最佳选择，亟须一个高效、准确的决策辅助工具。

2) 个性化推荐需求未满足：面对消费者需求的日益多样化，现有的推荐系统在提供高度个性化商品推荐方面力不从心。

3) 全网比价过程繁琐：用户在追求性价比时需手动在各大电商平台之间进行繁琐的比价，急需一种更高效的途径来提升比价效率。

2.落地方案

为解决上述问题，返利网采纳了大模型技术，提出了以下创新解决方案：

1) 智能辅助购物决策：结合大模型技术与商品专业知识库，构建了一个智能问答系统，能够为用户提供专业且易懂的购物建议，使他们更加高效地做出明智的购买决策。

2) 提升个性化推荐服务：借助大模型的深度学习优势，深入分析用户行为，建立细致的用户画像，进而提供更为精准、个性化的商品推荐。

3) 跨平台全网比价：通过大模型强大的数据处理能力，研发了一个自动化全网比价系统，快速识别并比较各个平台的商品价格，从而为用户提供性价比最佳的购物选择。



图 4 返利网大模型应用架构

3.实践效果

返利网通过大模型技术的深度应用，实现了用户体验与业务效能的双重突破。针对用户决策周期长的痛点，基于大模型构建的专业智能问答系统显著减少了用户的决策犹豫时间。依托大模型异构

数据处理能力的全网比价系统，可大幅节省用户比价时间，购物效率提升从而带动订单量增长。在个性化推荐领域，有效打破传统推荐系统的信息局限，全网比价点击率、销售转化率均有提升。该些解决方案的应用，标志着大模型技术已构建起电商导购场景用户体验和价值提升的一项重要竞争力。

(三) 招聘行业场景实践案例

1、智联招聘

智联招聘作为人力资本生态价值链平台，正在推进大模型技术在职位发布、简历筛选、人才推荐、面试等招聘全流程服务中的应用。

(1) 需求痛点

招聘平台主要工作是促进供需匹配，即把合适人匹配到合适岗位。企业招聘往往面临着最主要的两大难题：一是用人部门提出的需求描述主观且模糊，HR 无法据此得出可量化的简历挑选标准；二是招聘流程链路较长，各环节中 HR 要处理大量琐碎工作。如果无法解决这两大问题，不仅会消耗 HR 的时间和精力，企业招聘周期变长，更严重的是导致企业错失与岗位匹配的合适人才。

(2) 落地方案

利用大模型技术重构人岗匹配环节，通过自然语言对话，迅速为求职者和岗位画出精准匹配画像，提高匹配的准确性；通过自然语言对话获取求职者关键信息，精准筛选人才并推荐职位，提升用

人和求职双方效率。现在双方已合作沉淀出系列提示词模板，在数万条数据中验证，场景平均准确率高达 93%，这无疑为招聘行业带来革新。

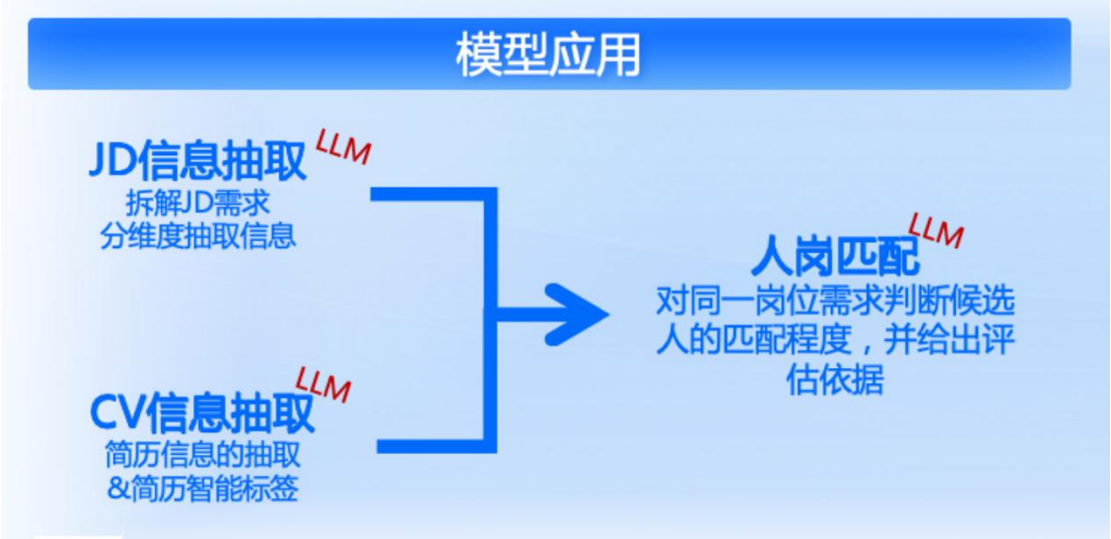


图 5 基于大模型平台的招聘场景示例

(3) 实践效果

为求职者推荐合适职位，为用人单位筛选所需人才，提高招聘效率；此外，协助 HR 解答工作时长、薪资福利、工作地点等问题还大大减少了 HR 的工作量，使得他们能够专注于更有价值的工作，提升招聘效率。百度和智联招聘合作，沉淀提示词模板，在数万条实际数据中验证，人岗匹配平均准确率高达 93%，大模型使用成本下降约 90%。让求职者轻松找到心仪工作，让企业高效获得所需人才，为招聘行业带来全新变革。

2、万码优才

万码优才是一个专注于数字技术人才招聘和职业发展的平台。随着技术的快速迭代和业务模式的不断创新，企业对数字人才的需

求迅速增长。在此背景下，如何精准高效地选拔人才成为关键问题。

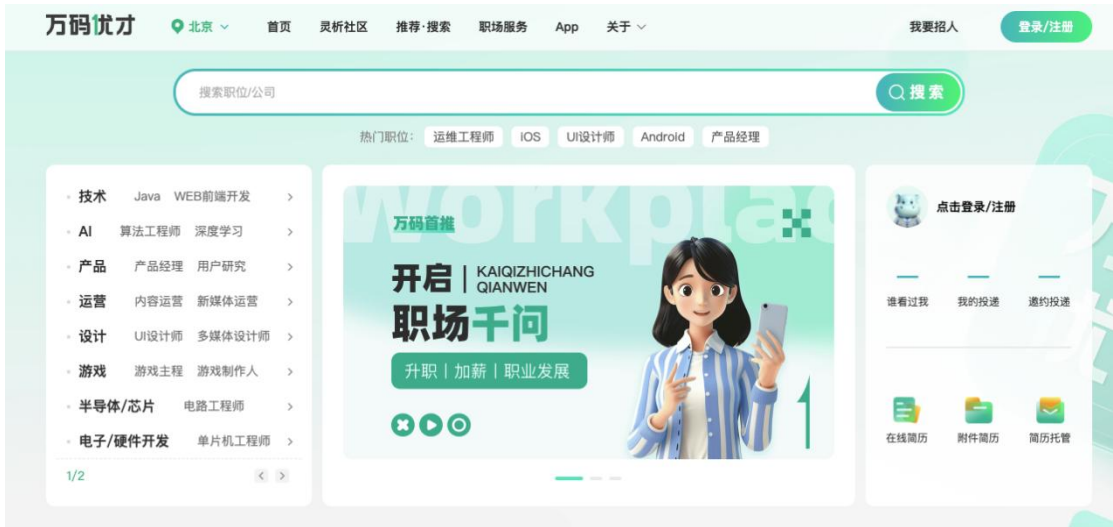


图 6 万码优才官方网站

(1) 需求痛点

当前，虽然市场上的数字人才众多，但企业在招聘过程中常面临无法精准匹配人才与岗位的困境。这一问题主要源于传统招聘流程的低效率和信息筛选的不精确。企业需求迅速变化，而传统方法无法及时反映和适应这些需求的变化，从而导致人才选拔不精准、效率低下。此外，人工审核的高成本和低效率进一步加剧了这一挑战。

1) 人岗不匹配问题：企业在招聘过程中常遇到的主要困难是如何精准地将合适的人才与相应岗位匹配。这导致了人才资源的浪费和招聘效率的降低。

2) 招聘流程低效：传统招聘方法在处理快速变化的市场需求时响应不够及时，导致招聘决策过时或不符合当前业务需要。

3) 审核成本和时间高：人工审核职位信息既耗时又耗力，且容

易出错，影响整个招聘流程的效率和精确性。

(2) 落地方案

为了解决这些问题,万码优才使用百度智能云千帆大模型平台,通过 AI 技术深度赋能招聘流程。首先,通过 ERNIE 4.0 模型强化文本理解能力, AI 可以自动审核职位信息,提高审核的精确性和专业性。其次,万码优才构建了全面的技能标签体系和知识图谱,通过智能算法分析求职者的技能和经验, AI 解构岗位的人才画像,快速识别与需求匹配的人才。此外,新功能“万码智搜”结合技能标签体系,能快速匹配推荐合适人选,极大简化招聘工作流程。

1) AI 增强的人岗匹配:通过使用百度智能云千帆大模型平台,万码优才应用 ERNIE 4.0 模型的强大文本理解能力,借助智能算法分析和解构岗位需求与人才画像,实现更精准的人岗匹配。

2) 提升招聘流程效率:AI 技术的引入极大提高了信息处理的速度和准确性,使招聘流程更加高效。AI 自动审核职位信息,显著提升审核效率,达到了人工审核速度的 5 倍。

3) 降低审核成本和时间:通过自动化的 AI 审核代替传统的人工审核职位信息,减少了人力成本,同时降低了错误率 and 提高了专业性,确保职位信息的准确无误且及时更新。



图 7 万码优才落地方案

(3) 实践效果

通过引入AI技术和智能化服务,万码优才显著提升了招聘的精准度和效率,岗位审核效率提升5倍。人岗匹配的精准度和质量得到提升,日推简历点击率提高35%,同时线上内容错误率降低3倍。这些改进不仅使企业能更快速地筛选和聘用适合的人才,而且减轻了人力资源部门的工作负担,提高了整个招聘流程的专业性和效率。此外,借助千帆应用开发平台快速上线关键模块,进一步提升了服务的灵活性和创新性。



图 8 万码优才效果展示

(四) 手机行业场景实践案例

当前智能手机已成为人们重要的 AI 接入终端,用户群体对先进的 AI 服务期待高,手机提供 AI 相关功能有助于冲击高端市场。头部手机厂商均将 AI 技术作为核心能力建设,通过引入 AI 大模型服务提升手机智慧服务能力,以满足用户对智能交互和功能需求,助力打造个人的智慧世界。

1.需求痛点

手机具备强交互体验属性,无论是面向用户的手机应用场景,还是面向企业的办公提效,用户在使用传统智能手机时,通常会有以下痛点:

- 1) 手机助手无法支撑开放自由的问答,对话回复体验不佳,多轮问答不佳;
- 2) 手机使用交互命令自然语言理解差,无法准确分发任务;
- 3) 手机轻办公场景,由于操作屏幕大小有限,难以完成一些较复杂的任务;
- 4) 传统浏览器的搜索体验不佳;
- 5) 助手的形象不够拟人化;
- 6) 能力拓展流程复杂。

2.落地方案

针对传统智能手机用户体验不佳的场景,使用大模型做意图识

别和任务分发，通过不同量级的大模型调用，支持手机助手，实现智能创作、数理计算、会议纪要、多文档问答、通用闲聊、角色扮演、情感陪伴等多个能力。构建统一的大模型服务能力解决方案，快速支撑业务场景落地。

- 语音助手实现基于文心大模型的通用问答、知识问答，打造一个无所不知的智能手机助手；
- 自然语言的提示词改写，提升意图理解准确率，准确任务分发到具体的应用中，让用户实现一个入口，自然对话，完成交互任务，极大地缩短交互点击的步骤，提升手机用户的交互体验；
- 轻办公场景，一键翻译、续写、改写、润色，文档问答、思维导图快速生成，极大提升移动端和 PC 端的办公体验；
- 将搜索体验和语音助手结合，自动联网抓取互联网数据，总结搜索内容，体验到更加汇聚和自然交互的搜索体验；
- 大模型帮助用户打造个性化、千人千面的数字人助理形象，让手机助理更加拟人化；
- 百度智能云千帆应用开发平台能够快速帮助客户新建“智能体”，实现不同的任务类型，让助理能力无限拓展，常用常新；
- 充分和百度生态合作开发第三方的应用插件，在手机助手中快速触发。

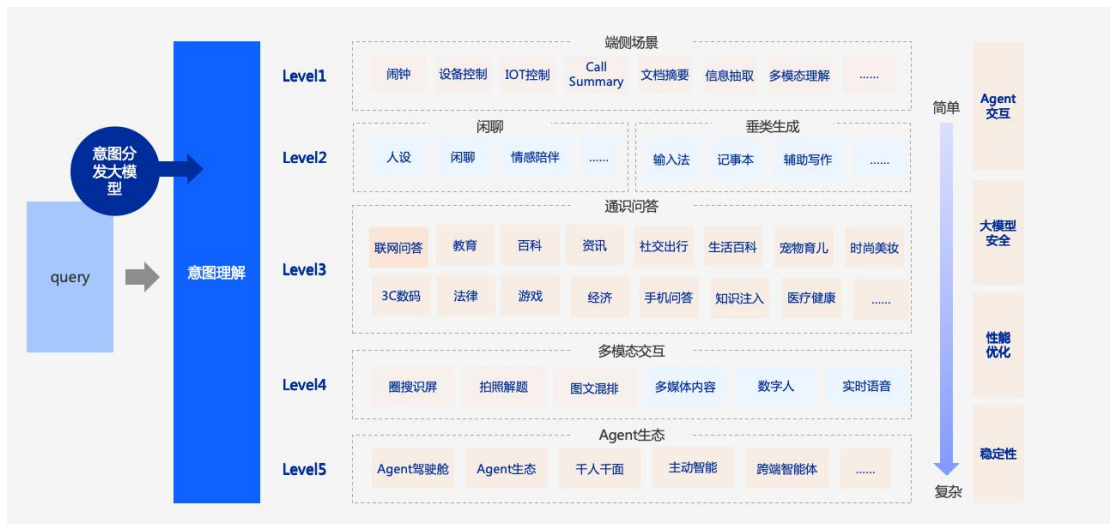


图 9 大模型赋能智能手机架构图

3.实践效果

通过手机系统应用提供大模型服务能力，在通用闲聊、多文档问答、智能创作等场景中基于长文本专业的能力做支撑，目前已在手机用户实际体验中，比如文案写作助理、智能会议纪要助理等，帮助用户快速生成写作草稿，提供写作思路、进行写作辅助等。在AI实时互动场景中，封装场景化AI Agent SDK，对接门槛低，最快1天便可快速实现AI实时互动能力，极大降低了接入成本。用户问题是否启用百度搜索，可由模型自动判断，也可以接口控制强制搜索，精准溯源，多条返回，模型推理总结，弥补了大模型实效性检索效果。应用手机智能问答助手后，用户有效问答轮次提升90%，相较传统问题应用效果提升明显。

（五）游戏行业场景实践案例

“AI+游戏”已成为全球游戏厂商新的竞争高地。在这一背景下，《逆水寒》手游携手网易伏羲共同推出全球首个游戏内AI竞技

场、也是全球评委人数最多（高达1亿）的AI竞技场——逆水寒AI大模型竞技场，开启了一种新的游戏体验方式。

1.需求痛点

在传统游戏中，各类互动玩法往往受限于预设的剧情文本，缺乏灵活性与自主性，玩家在游戏中会有刻板与机械化的感受。

2.落地方案

《逆水寒》手游利用大模型技术，为玩家提供了一种全新的游戏互动体验。大模型具备强大的时空和场景优化能力，能够无缝融入游戏的庞大开放世界观。

在《逆水寒》游历玩法中，放入9个故事主题下千行百业的NPC，这些NPC由大模型驱动，具有不同性格的NPC会根据玩家提出的话题进行探讨，两两PK，由玩家投票评选。玩家可直接从灵感库中选取问题进行提问，也可以自由输入问题。



图 10 《逆水寒》游戏剧情场景

3.实践效果

AI 竞技场玩法一改以往“玩家行动、AI 评分”的机制，让玩家反过来成为游戏的评委。通过这一机制，不仅可以为玩家带来全新游戏乐趣，还能够收集到玩家反馈。这一玩法获得广大玩家的喜爱，而来自真实玩家、消费者的反馈，有助于揭示 AI 在具体问题上的不足。一方面丰富了游戏的互动玩法，另一方面促进了大模型技术在游戏中的发展和应用。

（六）资讯行业场景实践案例

汽车之家互联网平台围绕看车、买车、用车、换车等多场景，为消费者提供便捷的汽车消费服务体验。通过应用大模型技术，用户个性化内容消费、客服智能化水平等多个场景下实现价值提升。

1.需求痛点

在运用大模型开拓业务时，RAG 方案成为关键路径，汽车之家的大模型实践同样依循此方案推进落地，然而过程并非一帆风顺。

1) 数据可靠性困境，汽车之家深耕多年，积累海量涵盖用户评论、论坛、社交及新闻等领域的数据，本是优势。但细究发现，部分数据存在隐患，用户评论准确性欠佳、促销政策类时效过期导致应答失准，易误导用户决策。

2) 数据全面性短板，之家的数据资源大多聚焦于车生态领域，随着用户需求日益多元、追求极致体验，亟需打破既有边界，广泛接入全网数据，以覆盖更多元的用户使用场景，满足多样化需求。

2.落地方案

1)设计采用预加工知识图谱策略。鉴于传统构建图谱方式成本高昂，在大模型的时代背景下，借助综合推理能力卓越的模型落地成为破局关键。经综合评测，选定文心大模型（ERNIE 3.5）对 3GC 内容进行图谱三元组抽取。在此过程中，同步联动汽车之家内部精细的结构化数据辅助校验，二者协同发力，高效剔除错误信息。同时，巧妙提取时效性协变量，为后续数据运用夯实基础。

2)聚焦数据全面性瓶颈，经效果评测后接入百度智能云千帆应用开发平台提供的搜索插件,实现了无缝对接全网搜索数据的能力，为业务注入新的信息源，拓宽发展边界。

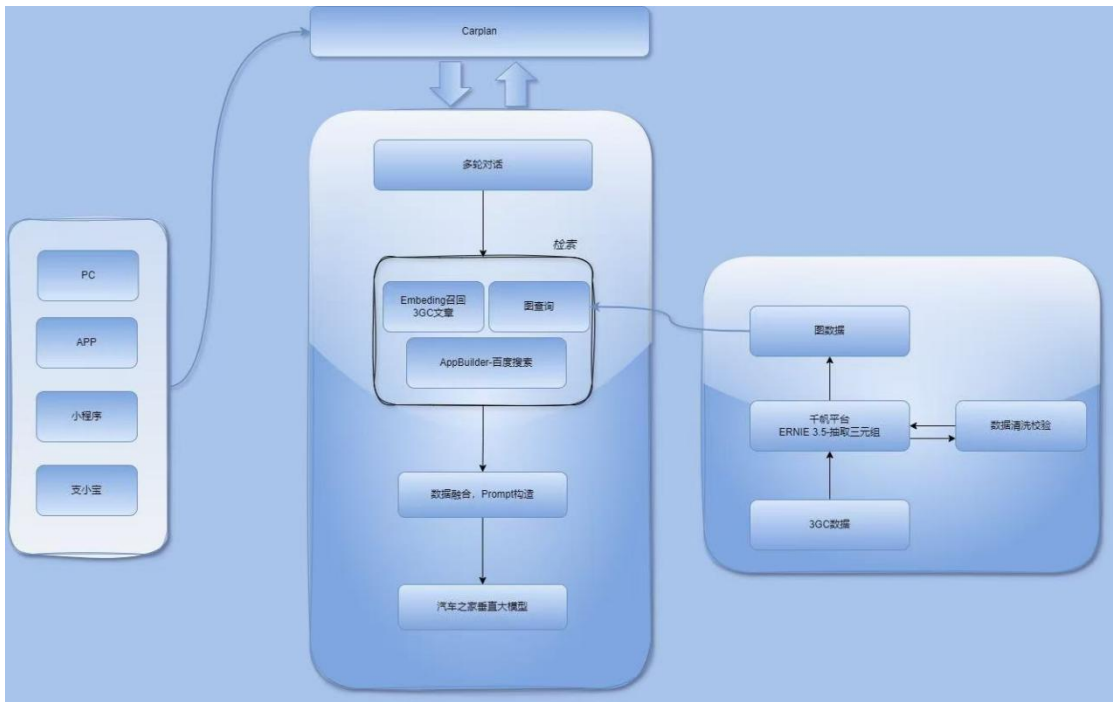


图 11 汽车之家基于大模型平台的场景示例

3.实践效果

经多维度评测，Carplan 在应对事实性问题时，其应答准确率达

到了 97%，为用户提供坚实可靠的信息保障；在处理主观性问题方面，受益于数据全面性的提升，用户的整体感受得到显著优化，满意度大幅跃升。当前，Carplan 的能力已通过小程序、支小宝等多元渠道进行了发布，持续为用户与行业创造价值。

（七）政务行业场景实践案例

1、中卫慧通

中卫惠通公司主要为基层数字治理、数字政务、数字乡村等领域提供服务。为更好赋能乡村地区发展，提升区域居民的生活质量和便利性，中卫慧通充分将大模型技术应用到数字政务场景中，与百度智能云千帆应用开发平台合作，开发了“居民服务助手”“村长 copilot”“智能填单”等能力，形成面向居民、领导、网格员业务闭环的产品矩阵。

（1）需求痛点

1) 信息获取不便捷：居民在获取医保缴费、婚姻登记、户籍办理等政务信息时，传统方式需要前往相关部门或通过电话咨询，耗时长且可能因信息不对称导致误解。

2) 服务响应慢：基层工作人员面对大量居民咨询，人工处理难以满足高效率的服务标准，尤其在春节等返乡高峰期，回复压力大。

3) 基层工作负担重：基层工作人员在处理群众诉求、编写文案、查询关键信息等工作中，常常面临工作量大、效率低的问题。

(2) 落地方案

基于千帆应用开发平台，构建了面向居民、领导、网格员业务闭环的产品矩阵，主要面向居民端开发了「居民服务助手」、面向领导端「村长 copilot」、面向网格员提供「智能填单」能力。

1) 居民服务助手：通过千帆应用开发平台构建 RAG 应用，微信群内/微信私聊便捷提问及问题 AI 智慧化回复、微信小程序打造智能客服窗口满足居民场景化及业务化咨询诉求、基于微信群与小程序的群众诉求进行问题自动转工单处置；

2) 村长 copilot：通过千帆应用开发平台 标签识别组件能力，集成至村长 copilot，完成一表通办、工单自分发、工作指南、居民信息一问知等功能；

3) 舆情监控平台：基于大模型能力进行舆情监控和管理；

4) 智能语音大屏：实现调度人员信息以对话形式快速查询，一问知其人。通过对话形式反馈待办事项，一问知何事。所言即所达，AI 将需要执行的任务一键推送给执行人并进行通知。



图 12 中卫慧通落地方案

(3) 实践效果

基于百度智能云千帆应用开发平台知识问答应用（RAG 框架）和百度文心大模型能力，构建本地政务知识库，只要居民在群内通过语音或者文本方式召唤“居民助理”并提出相关问题，如医保缴费、婚姻登记、户籍办理等，就能立刻获得详实、精准的解释说明，本地政务知识库回复不了的问题，则会通过百度文心大模型能力，结合百度搜索全网信息，在设定框架内进行回答。无需居民下载任何 App，居民助理自动@诉求提问人回复，真正实现 7X24 小时居民群内便民服务。大模型服务通过政务 Copilot 等产品，为基层工作人员在框架性文案编写、群众诉求回复等场景赋能，使其工作效率有效提升 25%，从而减轻了基层工作负担。

通过百度智能云千帆应用开发平台，乡村基层组织无需拥有深

厚的技术背景或大量的资金投入，就能通过简单易用的工具和平台接触并应用大模型技术。为乡村地区的发展带来了新的契机，提升了当地居民的生活质量和便利性。截至目前已覆盖河南省沈丘县、重庆丰都县、重庆巫溪县、天津市河西区、长沙市芙蓉区、安徽省歙县等 59 个区县，覆盖 2829 万人。



图 13 “居民助手”功能界面

2、国家环境监测总站水务

中国环境监测总站是生态环境部直属事业单位，是全国生态环境监测技术中心、网络中心、数据中心、质控中心、研究中心和培训中心，引领生态环境监测技术发展，为国家生态环境管理与决策提供监测信息、报告及技术支持，对全国生态环境监测工作进行技术指导。中国环境监测总站以生态环境监测数智化转型为引擎，探索大语言模型在生态环境监测领域的应用，培育监测业务辅助“智

能专家”，促进生态环境监测业务数智化跃升。

(1) 需求痛点

《关于全面推进美丽中国建设的意见》明确加快数字赋能，深化人工智能等数字技术应用，加快建立现代化生态环境监测体系。数智化转型作为现代化生态环境监测体系的动力引擎，是精准治污、科学治污、依法治污的重要手段与途径，将极大提高监管效率、保障监管效果，有助于推动构建美丽中国数字化治理体系，建设绿色智慧的数字生态文明。

(2) 落地方案

中国环境监测总站基于百度智能云千帆模型开发平台和应用开发平台构建全国生态环境监测领域行业大模型开发及应用平台，针对生态环境监测的数据生产、数据管理、数据应用环节中的关键场景开展探索，将逐步开展数据治理、行业知识库构建、算法工具研发及智能体（Agent）搭建，在环境质量研判、污染源监管、国家网质控运维等方面率先形成应用助手，推动 AI 赋能生态环境精准治理，充分释放监测数据在现代环境治理中的核心价值，助力环境监测业务技术创新能力智能化升级，提升业务管理、决策指挥和公共服务的效率和质量。



图 14 中国环境监测总站基于大模型平台的落地方案

(3) 实践效果

生态环境质量形势分析智能助手：为政府决策部门、科研机构及公众提供快速便捷的生态环境质量相关数据获取途径，帮助用户快速了解生态环境质量现状，为政策制定、环境保护措施制定提供依据。

环境质量监测运维质控助手：运维质控业务人员通过简单问答形式唤醒助手，实现运维质控流程中的任务分配、进度跟踪等环节智能控制，并实时向业务人员推送关键信息，提高流程效率。

生态环境监测技术知识问答助手：为生态环境监测领域的从业人员提供便捷、高效的技术知识获取途径，精准匹配技术知识，快速响应用户提问，提升从业人员的专业技能。



图 15 生态环境监测智能助手使用界面

(八) 能源行业场景实践案例

国家电网有限公司成立于 2002 年 12 月 29 日,2023 年收入 3.86 万亿元。中央直接管理的国有独资公司,是关系国民经济命脉和国家能源安全的特大型国有重点骨干企业。公司以投资、建设、运营电网为核心业务,承担着保障安全、经济、清洁、可持续电力供应的基本使命。公司经营区域覆盖我国 26 个省(自治区、直辖市),供电范围占国土面积的 88%,供电人口超过 11 亿。

1.需求痛点

数据：当前电力专业应用场景下,数据资源虽然丰富,但从海量数据中提炼有效信息,构建实时、高效、准确的语义理解与缺陷预测仍然是个艰巨的挑战。

应用生态：围绕电力专业业务系统能力提升和创新,如何确保

模型研发过程的有效性和可靠性，形成大模型与专业模型协同应用生态，有力促进模型在实际业务场景中的应用。

人际协同：提升电力设备运检效率，减少人为失误，确保电网安全稳定运行，需突破业务和技术融合应用的瓶颈。

下游服务：面向电网企业庞大的组织以及业务规模，如何降低下游模型研发门槛、提升电力专业数据和知识服务水平、加速模型应用于专业场景的效率，需开展行业级的人工智能平台工程化、体系化建设，解决领域内的挑战，为未来电力领域的发展提供技术支持和战略突破口。

2.落地方案

2020年，百度便与国家电网达成战略合作，进行电力行业人工智能创新，支持国家电网完成人工智能“两库一平台”智能基础设施（“两库”指模型库、样本库，“一平台”指包含运行环境和训练环境的人工智能平台）建设，通过打造AI应用，帮助国家电网实现业务提效等一系列突破性进展。

在此基础上，今年百度支持国家电网有限公司发布国内首个千亿级多模态电力行业大模型——光明电力大模型，为电网安全稳定运行、促进新能源消纳、做好供电服务提供“超级大脑”。百度智能云为国家电网提供了通用基础大模型与工具链平台、相关底层技术等重要支持，并在电网规划、电网运维、电网运行、客户服务等专业领域探索大模型应用。



图 16 电力行业大模型

1) 电网规划场景－可研文档评审智能体

针对场景的共性能力进行提炼，百度提供多轮改写、语义检索、语义排序等组件化能力，便于多场景灵活调用。

评审意见辅助生成：大模型智能化分析评审要点、将专家经验一定程度上标准化输出给大模型，然后智能生成评审意见。

2) 电网运维场景－设备运检智能体

针对多类型的主设备状态进行智能评估，融合分析分项诊断结果得到综合研判结论；结合历史故障案例、检修方案、操作规程等知识，应用大模型生成能力实现检修策略、检修计划、检修方案辅助生成，实现检修时长从一周缩短到一个小时，提升缺陷处置作业频率和效率。

3) 电网运行场景－配网负荷转供智能体

在故障发生后，利用大模型的任务编排能力和工具使用能力，结合已有机理模型工具，自动规划形成配电网故障处置的任务链；基于大模型的自主决策能力，评估分析机理模型生成的方案效果，依据分析结果自动调整机理模型的配置参数以实现转供方案的优化，

并通过智能体监视转供操作过程，评估恢复供电的成效。

4) 客户服务场景－供电服务方案生成

综合运用大模型能力，进行业务意图理解、多轮对话，实现智能化业务受理。运用大模型模式识别与自然语言理解能力，深度学习并模仿行业专家思维，基于电网一张图，综合电网资源、可靠运行等多维因素，依托智能体方案生成能力，自动提取并解析同类型方案，智能生成符合各项规范规程、高度定制化的供电方案答复单，打造供电方案策略库、案例库，实现供电方案的实时答复和线上勘察。

5) 物资场景－辅助招投标

综合运用大模型能力，进行业务意图理解、多轮对话，实现招标策略智能制定、招标文件比对、投标智能问答、投标文件自检、智能辅助评标业务智能化。运用大模型模式识别与自然语言理解能力，深度学习并应用招投标专家思维，依托智能体生成能力，自动提取并解析招标策略，智能生成符合的招标策略数据格式、实现投标文件智能检查，投标文件智能初评和详评，打造辅助招投标策略库、知识库，实现招采过程的智能化升级。

6) 日常办公场景——智能写作与办公辅助

在行政办公中，基于行业大模型与通用大模型能力，结合用户输入快速生成各类公文，如大纲生成、方案、总结、纪要、通知等，精准满足用户需求。大模型还提供文档比对与阅读助手功能，深度阅读并总结文档内容。同时，利用大模型的理解与 RAG 能力，实现

智能问答等。

3.实践效果

支撑国网贯彻国家发改委专项任务和国资委产业焕新行动要求，历时一年多正式发布千亿级多模态行业大模型——光明大模型，对于推动新型电力系统建设，加快形成新质生产力，更好保障国家能源安全、促进能源绿色转型具有重要意义。

1) 构建千亿级电力行业大模型：依托 Ernie3.5，融合电力专业知识，使用电力行业预训练数据 6G，增强数据 24G，行业 SFT 数据 2.6 万条，通用 SFT 数据 100 万条，构建电力行业大模型，在电力专项评测集上，整体效果较通用模型提升 12%

2) 构建企业级大模型开发与应用工具链：依托百度智能云千帆大模型平台，兼容国产服务器，实现 Ernie3.5 的训练与推理，融合百度 AI 应用开发工作台实现企业级智能体构建，支撑电网营销、办公、调度、设备、物资等 6 大业务方向的创新应用

(九) 百度智能云千帆大模型创新基地

百度智能云在全国多地建立大模型产业创新基地和中心，推动人工智能产业化。通过提供算法、算力和工具链，促进创新创业，培育企业，培养人才，助力传统产业智能化升级，同时普及人工智能知识，为区域经济发展和科技创新贡献力量。

1. 百度智能云千帆大模型产业（北京）创新基地

百度智能云千帆大模型平台在北京城区落地生根，既给石景山

区带来创新基石，又树立了百度与地方政府合作的全新模式。通过“一基地三引擎”服务模式，一是营造人工智能创新创业氛围：2024年初，已有两场百度智能云新品发布会在首钢园举办，吸引大量人工智能产业链企业关注和参与；百度将保障每年在石景山区举办4场优质论坛及交流活动。二是挖掘潜力企业开展培育赋能：通过前期调研，百度团队已经与多家属地企业达成基于百度智能云千帆平台共创AI原生应用的意向。三是人工智能人才引入与培养：通过百度AI品牌课程和AI领域教育资源，培训本地企业、开发者及相关人才，伴随企业进驻，带动人工智能大模型产业领域高层次人才。四是打造人工智能大模型示范应用：在制造、教育、医疗、能源、交通、金融、文旅和城市管理等行业领域，依托百度平台带动传统产业进行智能化改造升级。

2. 百度智能云千帆大模型（无锡）创新中心

百度首个全国私有化区域部署的大模型创新中心，它集成了算法算力中心、科研中心、智能制造工场、人才创培基地及完善的运营服务体系。依托百度飞桨深度学习平台，提供人工智能全流程工具链，推动区域智能制造全面升级。

3. 百度智能云（武汉）新质生产力产业基地

百度集团将以汉江湾人工智能产业园建设为契机，与硚口合作建设百度智能云（武汉）新质生产力产业基地，围绕算法、数据、模型、人才、生态等人工智能产业要素，打造大模型产业创新中心、人工智能数据标注中心及大模型人才实训培训中心，联合合作伙伴，

引入百度生态链企业落地研口，入驻基地，展开构建人工智能产业集群。