中国移动 China Mobile

CAICT 中国信通院

AI原生基础设施实践指南

2026

中国移动通信集团公司数智化部
中国信通院云计算与大数据研究所

编制说明

本报告的编制得到了多家数智化领域企业支持，主要参与单位及人员如下：

中国移动通信集团公司数智化部：陈国、孙昊、马维晶、吴晶、于顺治、王婷、钟玮军、范坤、陈仁强、魏宝辉、吴坤、赵昱、杨葳、邵欢庆、廖伟达、柴壮、朱永北、童同、王泽琪、尹星宇、张盈辉、洪建辉、赵玉春、严俊、金可、韩宇峰、梁小涛、孟繁宇

中国信息通信研究院云计算与大数据研究所：姜春宇、马鹏玮、王超伦、刘渊、高月

中国移动通信集团江苏有限公司：樊野、李睿、张兰、杜曦、张念启、杨林、王上淇

中国移动通信集团上海有限公司：江济、温士帅、赵宇、王灏、顾王一

中国移动通信集团安徽有限公司：胡卫、周岚、李正、魏代祥、申凯、王浩然、李锐

中国移动通信集团山东有限公司：张振刚、潘东、黄涛、徐秀珊、赵强、袁明兴

亚信科技（中国）有限公司：英林海、安军、苗森、唐迎春、徐晨兴、张云翔

科大讯飞股份有限公司：余俊涛、朱广勇、黄达、李海明

阿里云计算有限公司：王亮、梁洪伟、随鑫、邹雨竹、李华东

腾讯云计算（北京）有限责任公司：郭枭、周锐

华为技术有限公司：孙守恒

火山引擎：宋雪思、林森、张春雷、汤程平

南京星邺汇捷网络科技有限公司：庞海东、兰清

合肥非度信息技术有限公司：杨咸福、吴峰、陈皓

前　言

随着数智化转型进入深水区，人工智能技术正在以前所未有的深度和广度渗透各行各业，不仅重构了生产要素的配置逻辑，更催生层出不穷的新型产业形态，驱动经济社会发展模式发生根本性变革。2025 年 8 月 26 日，国务院发布的《关于深入实施“人工智能+”行动的意见》提出“发展智能原生技术、产品和服务体系，培育智能原生企业，催生智能原生新业态”的总体要求，标志着我国数智化转型正迈向全面智能化阶段。AI 原生基础设施作为智能原生业态创新的必要条件，已成为数智化时代新质生产力的关键技术底座。

AI 原生基础设施（AI-Native Infrastructure）是从设计阶段即将规模化支撑 AI 原生应用作为核心理念，全栈适配 AI 特性的基础设施体系。其不仅仅是现有业态的“AI+”升级，而是从根本上重塑了价值获取、创造和交付方式，并实现技术自主可控、场景高效落地、生态开放协同。

新型基础设施适度超前建设的政策导向和“AI+”行动的持续推进使得产业对 AI 原生基础设施需求空前高涨。在此背景下，中国移动通信集团公司数智化部联合中国信通院云计算与大数据研究所及业界专家共同编制《AI 原生基础设施建设指南（2026）》，旨在洞察国家战略导向、聚焦产业实践与技术前沿，深度融合多方实践经验，为国央企数智化领域的规划者、建设者及 AI 原生基础设施产业全链条从业者，提供兼具前瞻性与实践性的参考指引。

目 录

图 目 录

表 目 录

一、 AI 原生基础设施兴起的时代背景

当前，我国正迎来人工智能产业化发展浪潮，AI 的规模化应用业已成为行业发展主旋律。国家持续加大相关政策供给力度，护航 AI 产业高质量发展。随着开源大模型 DeepSeek 等国产化新技术的涌现，企业引入 AI 技术的门槛显著降低，在拓展数智化转型实践纵深的基础上，为传统 IT 基础设施演进升级解锁了更多可能。

(一)政策牵引力

2017 年，国务院发布《新一代人工智能发展规划》，AI 作为国家“新质生产力”的关键载体，其重要性已上升至国家战略层面。此后各部委陆续出台相关政策，从教育、产业、科技、安全等方面完善 AI 战略布局。2025 年 8 月 26 日，国务院公布《关于深入实施“人工智能+”行动的意见》，提出“发展智能原生技术、产品和服务体系，培育智能原生企业，催生智能原生新业态”。“人工智能+”本质在于以 AI 技术作为核心驱动力，对经济社会全链条进行“重构式”融合，实现生产力跃迁和生产关系变革。AI 已不再是简单的效率辅助，而是像电力一样成为支撑所有行业的通用基础设施，重塑各个行业的底层逻辑，对 AI 的战略定位从“赋能工具”向“基础设施”转变。2025 年 10 月 28 日，《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》提出“全面实施‘人工智能+’行动，以人工智能引领科研范式变革，加强人工智能同产业发展、文

化建设、民生保障、社会治理相结合，抢占人工智能产业应用制高点，全方位赋能千行百业”。AI 已成为国家布局推动智能经济发展、构建智能社会的重要战略要素。

表 1 我国人工智能关键政策

政策文件	发布主体	发布时间	AI 相关布局
《新一代人工智能发展规划》	国务院	2017 年 07 月	提出“到 2030 年人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工智能创新中心，智能经济、智能社会取得明显成效”。
《高等学校人工智能创新行动计划》	教育部	2018 年 04 月	部署“优化高校人工智能领域科技创新体系”、“完善人工智能领域人才培养体系”、“推动高校人工智能领域科技成果转化与示范应用”。
《国家新一代人工智能创新发展试验区建设工作指引》	科技部	2020 年 10 月	布局“建设 20 个左右试验区，产出一批重大原创科技成果，创新一批切实有效的政策工具，形成一批人工智能与经济社会发展深度融合的典型模式，积累一批可复制可推广的经验做法，打造一批具有重大引领带动作用的人工智能创新高地”。
《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	科技部等六部门	2022 年 07 月	提出“以促进人工智能与实体经济深度融合为主线，以推动场景资源开放、提升场景创新能力为方向，强化主体培育、加大应用示范、创新体制机制、完善场景生态，加速人工智能技术攻关、产品开发和产业培育，探索人工智能发展新模式新路径，以人工智能高水平应用促进经济高质量发展”。

《生成式人工智能服务管理暂行办法》	网信办	2023 年 07 月	规范生成式 AI 服务发展，强调“坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管”。
《信息化和工业化融合 2025 年工作要点》	工信部	2025 年 07 月	提出“实施‘人工智能+制造’行动，支持企业在重点场景应用通用大模型、行业大模型和智能体”。
《关于深入实施“人工智能+”行动的意见》	国务院	2025 年 08 月	提出“发展智能原生技术、产品和服务体系，培育智能原生企业，催生智能原生新业态”。
《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》	中共中央委员会	2025 年 10 月	提出“全面实施‘人工智能+’行动，以人工智能引领科研范式变革，加强人工智能同产业发展、文化建设、民生保障、社会治理相结合，抢占人工智能产业应用制高点”。

来源：中国信息通信研究院,2025

（二）产业驱动力

2025 年，全球 AI 产业化进程势头正劲，技术产业融合速度持续加快。IDC 数据显示，2024 年全球人工智能的 IT 总投资规模为 3158 亿美元，并有望在 2028 年增至 8159 亿美元。同时，我国 AI 产业也进入规模化落地阶段，据中国信通院测算，2024 年我国人工智能产业规模已超 9000 亿元，同比增长 24%。AI 正以核心引擎之姿，引领发展逻辑从从“数据驱动”向“智能决策”的跨越，加速新质生产

力的形成。AI 规模化落地具有三个显著特征：一是 AI 应用从小模型应用转向大小模型协同，算力和数据需求量级跃升，推动 IT 基础资源供给规模持续增长；二是 AI 深度融合企业核心业务场景，对 IT 基础设施的性能和可靠性提出了更高要求；三是 AI 应用范围持续拓宽，AI 普惠化成为驱动产业数智化的重要趋势。由此可见，AI 规模化落地对企业 IT 基础设施提出了“既要大、又要快、还得省”的刚性需求，原有 IT 基础设施的转型升级刻不容缓。

（三）技术创新力

AI 技术历经多年高速发展，迄今仍处于持续迭代演进的上升期，其发展活力主要体现在算力、数据、模型和应用等四个方面：

算力方面，算力需求正驱动智算基础设施发生根本性变革，算力平台向超大规模异构融合与全局调度演进。为了支撑万亿参数的模型训练，智算集群正从万卡级规模向十万卡级迈进，这对集群网络拓扑（如非阻塞架构）、高速互联（800G/1.2T 光互联）和冷却技术（液冷普及率超 80%）提出了极高的要求。在硬件层面，CPU、GPU、NPU 乃至存算一体芯片的超异构融合成为突破算力与能效瓶颈的关键路径，例如 NVLink-C2C 等先进互连技术，可实现内存一致性共享，大幅降低数据搬运开销。在算力供给模式上，业界正在积极探索从单体中心支撑向社会化服务的转变，加速发展算力网络以实现对广域分布式、跨技术架构（通算、智算、超算）算力资源的统一标识、感知与智能调度。领先企业已率先通过构建“算网大脑”

与“四算合一”调度平台，实现全国性算力资源的一站式供给与任务级智能编排，使算力成为可即取即用的社会公共服务。在自主可控方面，构建软硬件一体化适配平台，兼顾多元 AI 芯片生态，加速国产算力从“可用”到“好用”的进程。

数据方面，高质量数据集不仅是被动输入的“原料”，更是主动驱动模型、应用能力演进的结构性资产。数据技术正经历三大范式跃迁：一是从“以存储为中心”转向“以语义智能为中心”，通过本体建模与知识图谱，实现多源异构数据的深度语义对齐，使数据具备可推理、可组合的智能属性；二是从“依赖真实采集”转向“按需制造”，依托生成式 AI 与物理仿真引擎，构建任务导向的合成数据工厂，实现高保真、高合规、高覆盖的虚拟语料自主生产；三是从“静态交付”转向“动态闭环供给”，融合自动化质量评测、细粒度数据血缘追踪与跨模态检索能力，形成可评估、可迭代、可追溯的高质量数据持续供给机制。这些创新共同构成了面向 AI 应用的新型数据基础设施，其核心目标不再是简单管理数据，而是生成智能、保障价值、激活要素。基于这一基础设施，不仅能有效应对数据隐私、稀缺性和成本等关键挑战，更开启了“数据制造”的新范式。为充分释放数据要素潜能，行业正积极构建数据要素流通基础设施，例如数据空间（Data Space）与数据联网（DSSN），结合隐私计算、区块链等技术，在确保安全合规与权益归属的前提下，促进跨域数据的安全可信流通与协同利用。

模型方面，大小模型产业共生，开源生态与闭源伙伴协同发展。聚焦通用智能的基础大模型、面向端侧和 IoT 的海量小模型，以及介于两者之间深耕行业知识的行业特色模型，共同构成了模型产业发展版图。开源生态（如 DeepSeek、Qwen）极大地加速了技术民主化，闭源模型则突破性能的束缚。当前，模型技术的创新已不仅追求规模扩张，也注重效率与实用性的提升。推理优化技术（如投机解码、注意力优化）致力于降低大模型服务成本。领先企业的 MaaS（模型即服务）平台通过模型、算力与研发过程的集成，提供从模型选型、微调到部署运维的一站式服务。

应用方面，智能体（Agent）作为新一代人机交互界面，已成为 AI 应用的主流形态。前沿技术正在从单智能体任务执行领域，迈向多智能体的高质量协同，通过角色分工、知识共享与竞争协作，解决复杂规划问题（如供应链优化、自动驾驶等）。支撑智能体规模化发展的基础设施日益成熟，正逐步演进为“AI 原生操作系统”，提供关键的系统级服务，覆盖算力调度、记忆存储（向量数据库）、工具调用（通过 MCP 协议）、网络通信（Agent-to-Agent）以及全生命周期的可观测性与运维等方面。这标志着 AI 基础设施的焦点，已从支撑模型训练，扩展到支撑模型的持续认知与行动。头部企业纷纷推出智能体研发平台（如 Microsoft Agent Framework、华为鸿蒙平台、字节扣子平台、阿里云百炼平台、中国移动聚智平台等）。未来，标准化协议与低代码平台将进一步推动智能体向普惠化、专业化发展，深度嵌入企业的经营管理与生产运营的核心场景。

二、 AI 原生基础设施发展脉络与架构

AI 原生概念于 2020 年被百度首次提及。大模型技术的爆发刺激了产业在 2023 年对 AI 原生理念的关注。亚信、清华大学和 Intel 联合在2024年将AI原生定义为“从设计之初即以 AGI（Artificial General Intelligence）能力为基础构建的数字化系统”。2025 年国务院《关于深入实施“人工智能+”行动的意见》中首次以官方文件的形式提及“智能原生”，标志着 AI 原生系统及应用正式进入规模化落地阶段。AI 原生系统及应用的高速发展对企业 IT 基础设施建设提出了新的要求。

(一)AI 原生基础设施发展历程

AI 原生基础设施概念是一个不断发展的过程，可分为萌芽期、探索期、发展期三个阶段，从引入 AI 的 IT 基础设施逐步迈向 AI 原生基础设施。

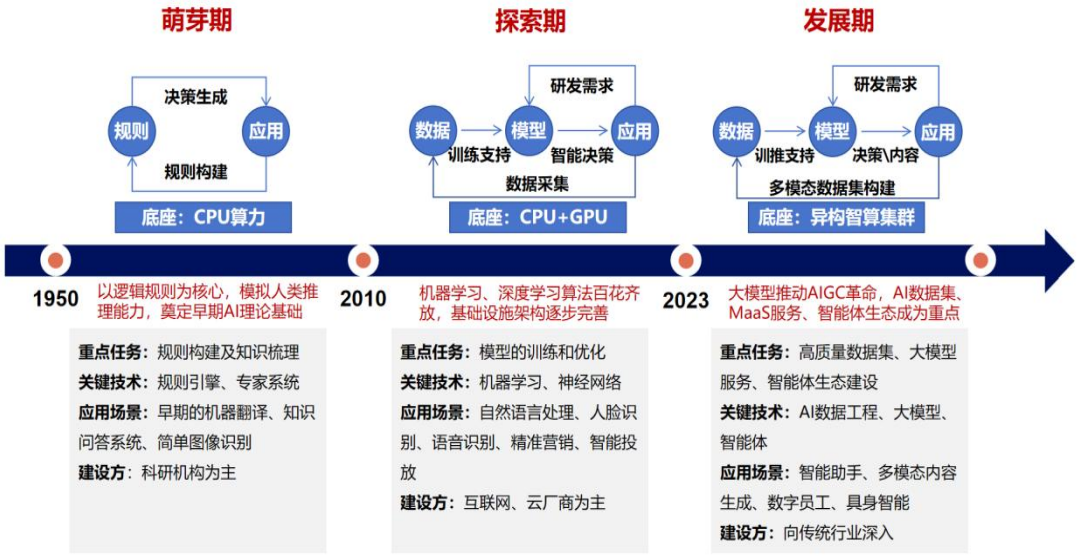

图 1 AI 原生基础设施的发展历程

萌芽期（1950 年至 2009 年）：著名的图灵测试诞生，标志着人类即将开启 AI 时代的新篇章。在此后长达 60 年的时间，AI 的应用与实践主要活跃在学术圈和实验室。这一时期，AI 算法依赖人工进行规则设计与开发，人们主要在 IT 系统中引入单个 AI 能力来解决问题，真正意义上面向 AI 的基础设施尚未形成。

探索期（2010 年至 2022 年）：大数据技术崛起，机器学习、深度学习大幅推动 AI 技术和产业升级。谷歌、亚马逊、阿里云、百度等头部云厂商积极发掘 AI 商业价值，并推出了 AI 相关平台及工具产品，满足机器学习、深度学习算法的研发、部署及服务需要，并同步开启了基于 AI 来设计基础设施的探索之路。

发展期（2023 年至今）：伴随 ChatGPT4.0 的问世，生成式大模型及相关应用爆发式增长，为企业数智化转型提供了无限可能。AI 与各类业务场景深度融合发展，产业对人工智能的需求空前。大模

型及智能体的技术革新，为基础设施的体系化重构迎来重要契机，全栈适配 AI 特性的基础设施已成为产业可预见的演进趋势。

(二)AI 原生基础设施定义

“AI 原生（AI-Native）”是指从设计之初就将 AI 考虑进来，实现产品、服务甚至整个业务模式围绕 AI 核心能力（理解、生成、推理、记忆）进行根本性创新的范式。

IT 基础设施是创建和部署应用所需的硬件和软件集合。狭义基础设施概念立足 IT 支撑人员，聚焦 IaaS 层软硬件，是覆盖算力、网络、计算框架的能力底座。广义基础设施则是从 IT 应用的最终用户视角出发，涵盖 IaaS、PaaS 及 SaaS 层能力，整合支撑应用所需的算力、存储、网络、数据、算法、工具等各类要素。本报告采用广义定义。

综上所述，AI 原生基础设施（AI-Native Infrastructure）是从设计阶段即将规模化支撑 AI 原生应用作为核心理念，全栈适配 AI 特性的基础设施体系，通过软硬件、网络、数据、算法等要素的深度协同，为 AI 原生应用的研发、部署、运行和管理提供全生命周期的能力支持。AI 原生基础设施以支撑丰富多样的 AI 原生应用为底层设计逻辑和核心驱动力，将大小模型协同的 AI 能力作为核心价值输出，赋能企业打造全新架构的 AI 原生应用系统，实现业务流程的全流程再造和系统性效率提升，进而催生出的全新的商业模式、组织形态和产业生态。AI 原生基础设施将承载一种全新的生产关系，

为人类从事业务价值创造活动开创了 AI 原生的工作场域，重塑了人机协同发掘-创造-验证-优化业务价值的新生产方式。区别于现有基础设施的“AI+”升级，它真正重塑了 AI 的业务赋能体系。

（三）AI 原生基础设施架构

AI 原生基础设施建设的总体目标是构筑面向智能应用的一体化开发、运行、支撑的软件平台，打通“算力调度—模型开发—智能体部署”全链路，助力 AI 应用一个入口访问、一套接口集成、一套用户体系登录、一套技术架构运转、一套数据标准流转、一套运维体系管理，构建统一技术架构、统一接口与标准、统一数据、统一用户、统一运营与运维体系。基于 AI 原生基础设施在 AI 应用全生命周期表现出的特征，其架构设计按照各类关键要素的服务场景可分为通智算基础资源、通智算调度引擎、沙箱、模型研发生产、数据供给、向量数据库、智能体引擎、AI 网关、AI 原生应用开发管理、AI 原生运维、AI 安全保障、数字可信等。

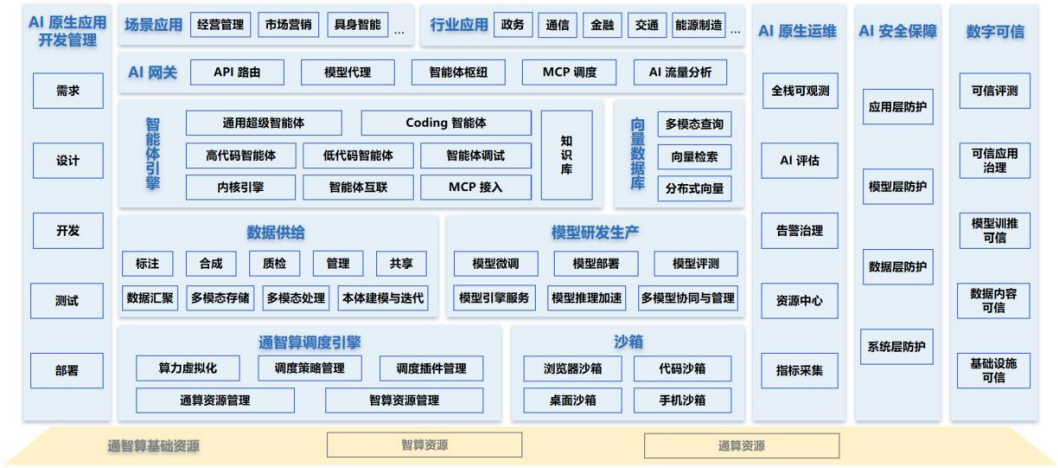

图 2 AI 原生基础设施总体架构

通智算基础资源：通智算基础资源是面向大模型与智能体时代的新型“算力+数据+网络”基础设施，在传统通用算力基础上，将智能算力也纳入了整体算力基础设施体系。

通智算调度引擎：通智算调度引擎作为 PaaS 层技术底座，承担着容器层异构算力资源调度,连接底层基础设施与上层应用的作用。引擎提供各类计算资源、网络资源和存储资源的统一动态调度分配，针对不同场景需求，支持多种调度策略配置与插件能力。

沙箱：沙箱是智能体运行时的关键组件，它使智能体能够安全、可靠地调用外部工具（如执行代码、操作浏览器），成为连接大模型智能体与外部世界的“安全操作手套”。

模型研发生产：模型研发生产模块提供覆盖模型微调、模型部署、模型评测、模型引擎服务、模型推理加速、多模型协同管理的工具链服务，支撑模型研发标准化、一体化、体系化的生产运营体系。

数据供给：数据供给模块是一个面向 AI 原生的综合性数据基础设施，集数据汇聚、存储、处理、标注、合成、质量评测、管理、共享、本体建模与迭代等能力于一体，构建全流程的高效数据供给体系。

向量数据库：向量数据库是 AI 原生应用的重要数据组件，承担高维向量的高效存储、检索与管理能力。依托既有的关系型数据库+向量引擎深度融合能力，以及混合查询基础架构，支撑智能知识检索、检索增强生成（RAG）等各种复杂应用场景。

智能体引擎：智能体引擎是 AI 原生基础设施中的核心上层建筑，其定位不仅是一个开发工具集，更应致力于打造企业级智能体操作系统（Agent OS）。智能体引擎汇聚全景 AI 能力，通过建设统一的技术标准与协议规范，屏蔽底层基础设施的复杂性，为上层应用提供标准化的“系统调用”接口。

AI 网关： AI 网关是 AI 原生基础设施构建的核心要件。其核心功能包括 API 路由、模型代理、智能体枢纽、MCP 调度、AI 流量分析等模块，具备丰富的集成和全生命周期治理能力。在最终用户、AI 应用和模型之间发挥枢纽作用，实现了 AI 应用研发生产要素的高效调配、为业务提供应用级精细化运营支撑。

AI 原生应用开发管理：将 AI 能力深度嵌入项目管理，覆盖需求、设计、开发、测试、部署的全过程，提供涵盖项目管理相关的智能化能力。从辅助研发升级为 AI 自主化操作，重塑研发交互方式，提升研发效能。

AI 原生运维： AI 原生运维是面向 AI 原生的全栈可观测运维体系，可精准监控模型行为、快速定位故障、科学评估输出质量，保障生产环境中模型的可靠运行，实现模型性能与业务需求的深度融合，确保 AI 应用高效稳定安全运行。

AI 安全保障： AI 安全保障是 AI 原生基础设施安全、可靠、可信运行的核心保障体系，助力系统应对提示词攻击、数据投毒等新型风险，确保 AI 行为可控、输出合规、运行可信。

数字可信：以数字可信体系为基石，构建覆盖“可信算力协同、可信数据供给、可信模型训推、可信应用治理”并以“可信测评体系”贯穿支撑的总体架构，形成面向 AI 原生的可信 AI 能力底座。

三、 AI 原生基础设施建设思路

(一)通智算基础资源

通智算基础资源是面向 AI 原生的新型“算力+数据+网络”基础设施，包括智算资源和通算资源在内的弹性资源池。

智算资源：是以 GPU、NPU、MLU 等异构加速卡为核心、通过高速 RDMA 网络与分布式存储池化形成的弹性 AI 算力资源,为 AI 模型的训练、微调和推理提供高效的支撑。

通算资源：是由 x86/ARM 服务器、大内存、SSD 组成的通用算力，负责处理向量数据库、消息队列、AI 训练数据预处理等通用负载。

通智算基础资源的管理优化是实现 AI 原生基础设施高效、低成本运营的关键。企业通过采取制定合理的资源规划方案、实现智算资源与通算资源的融合调度、建立全面的资源监控体系以及实施资源回收与再利用策略等措施，可以充分利用资源的弹性优势，满足业务快速变化的需求,最终实现像水电一样按需取用 AI 算力的能力。

(二)通智算调度引擎

在 AI 原生基础设施整体架构中,通智算一体化调度引擎发挥着容器层异构算力资源调度、连接基础资源与上层应用的作用。引擎提供各类算力资源、网络资源和存储资源的统一动态调度分配，针对不同场景需求，支持多种调度策略配置与插件能力。

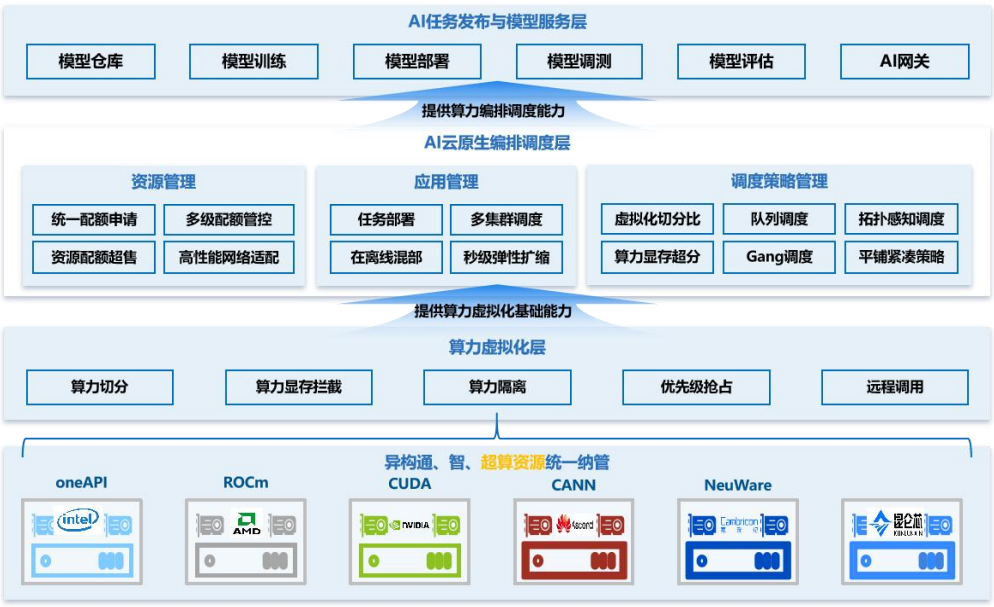

图 3 通智算一体化调度引擎架构

通算资源管理：可按照集群、主机、租户配额、系统配额等维度进行通算资源（CPU、内存等）的管理和监控。

智算资源管理：可按照集群、主机、设备卡、租户配额、系统配额等维度进行智算资源（GPU、NPU、MLU、显存等）的管理和监控。

算力虚拟化：基于各类算力资源的技术架构，通过软件定义资源池的方式，实现对异构通、智计算设备的适配支持与虚拟化调度。

调度策略管理：可按照集群和主机等维度对通、智算资源调度策略进行统一配置管理，如设置集群或主机计算资源超售比、虚拟化切分比、队列调度、网络拓扑感知调度、Gang 调度、平铺紧凑策略等，以支持更多场景下资源的高效利用。

调度插件管理：支持多类型智算资源调度插件的一键安装和卸载，插件能够实现对集群智算资源的感知发现，并能够实时更新。

异构算力资源调度引擎建设主要分为算力虚拟化层、AI 云原生

编排调度层两层架构。算力虚拟化层基于各类型算力资源的内核与驱动技术架构，通过软件定义资源池的方式，在内核态与用户态进行 API 拦截，实现对异构通、智算设备的适配支持与虚拟化调度，为调度层提供算力虚拟化基础能力；AI 云原生编排调度层基于底层算力虚拟化统一纳管适配，进行统一资源管理、应用管理和调度策略管理等，为上层模型服务层提供算力编排调度能力。

（三）沙箱

沙箱是一种资源隔离与控制技术，它通过构建一个受限制、可监控的虚拟执行环境，使程序或代码能够在此环境中运行，而无法直接访问或影响真实的主机系统、网络、数据和其他应用程序。在 AI 原生应用架构中，智能体沙箱使智能体能够安全、可靠地调用外部工具（如执行代码、操作浏览器），成为连接大模型智能与外部世界的“安全操作手套”。沙箱作为数字世界的安全隔离与实验场，其核心内涵在于“隔离”与“控制”，将潜在的威胁如恶意代码、不稳定程序、未经测试的软件限制在可控范围内，是保障系统整体安全和稳定的关键基础设施。根据智能体的应用类别，沙箱可分为浏览器沙箱、代码沙箱、桌面沙箱、手机沙箱四个场景。

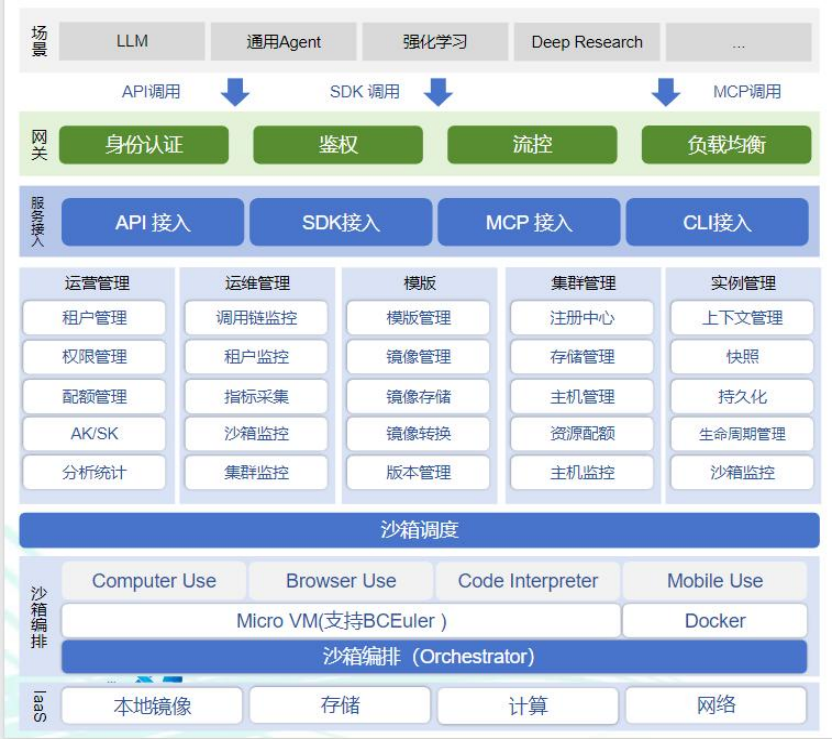

图 4 AI 沙箱架构

浏览器沙箱：将网页内容（如渲染进程、插件）的执行环境与浏览器内核及操作系统隔离，防止恶意网页危害用户设备。实施同源策略（SOP）、内容安全策略（CSP），控制资源加载与通信，可限制网页脚本对本地文件系统、设备硬件的直接访问。起到抵御网络钓鱼、恶意脚本、零日漏洞攻击，保护用户上网安全的作用。

代码沙箱：为动态代码（特别是用户提交或 AI 生成的不可信代码）提供安全的执行环境，严格控制其系统调用和资源使用。核心能力包括系统调用过滤及虚拟化，如通过 Seccomp、Namespaces 等机制限制或重定向文件、网络、进程的访问请求等；资源配额限制，如严格限制 CPU 时间、内存、磁盘 IO 及执行时间，防止拒绝服务攻击等；环境虚拟化支持创建 Python、Node.js 等临时纯净语言运行时，执行完毕后环境即被彻底销毁，以此满足在线代码评测、

第三方插件与脚本运行、AI 代码解释器（如 Code Interpreter）等场景的环境需求。

桌面沙箱：将整个桌面 AI 应用程序及相关数据封装在隔离环境中运行,防止 AI 应用潜在的恶意行为或不稳定因素影响宿主系统。核心能力包括文件系统虚拟化及重定向，使应用对“系统文件”的修改实际发生在隔离区，真实系统不受影响；预制可用软件，可根据客户需求，自定义工具软件，配合智能体完成任务的自动化。支持安全运行来源不可信的软件，测试不稳定或冲突的软件，保护企业终端安全。

手机沙箱：将手机应用及其数据封装在隔离环境中运行，防止应用潜在的恶意行为或不稳定因素影响手机宿主系统。智能体等 AI 应用对手机系统文件、用户数据的读取和修改操作，实际仅发生在沙箱的隔离区域内，真实的手机系统和个人数据不受任何影响。手机沙箱支持安全运行来源不可信的手机智能体应用，保护手机终端的系统稳定与用户隐私安全。

沙箱的功能架构包括隔离层、编排与管理层、安全策略引擎、沙箱亲和调度、生命周期管理、资源管理、可观测性与运维等能力模块。隔离层：基于 MicroVM 构建内核级隔离能力，提供更强隔离，适合多租户不可信代码；相比完整虚拟机消耗资源更小。编排与管理层：建设调度器管理 MicroVM 实例。安全策略引擎：集成或开发策略管理模块，支持动态加载安全策略如网络规则。沙箱亲和调度：提供根据调用参数、各节点资源使用情况、会话亲和需求等信息将

沙箱调用请求定位到合适的节点上。生命周期管理：实现实例的创建、暂停（快照）、恢复、销毁全生命周期自动化管理。资源管控：集成配额管理（CPU、内存、磁盘、网络），防止资源滥用。可观测性与运维：采集实例级别的资源使用率、网络连接、进程列表等指标，记录所有安全相关事件（如策略违反、逃逸尝试）和用户操作日志，建立沙箱基础镜像和宿主系统的安全补丁定期更新机制。

(四)模型研发生产

模型研发生产提供覆盖模型微调、模型部署、模型评测、模型引擎服务、模型推理加速、多模型协同管理的工具链，支撑模型研发标准化、一体化、体系化的生产运营体系。在 AI 原生基础设施的实践框架中，模型研发生产与服务是面向 Agentic AI 的核心引擎。

模型微调：是在预训练好的大模型基础上，使用目标任务的小规模数据集，对模型部分或全部参数进行小幅度更新的过程。可依托 LoRA 工具链压缩微调显存占用，结合 DPO、KTO、GP3 等偏好优化算法，精准提升模型效果。

模型部署：是将训练好的机器学习、深度学习模型从研发环境迁移到生产环境，使其能够接收输入、执行推理并输出结果的过程。核心目标是让模型稳定、高效、低成本为 AI 应用提供服务。

模型评测：是对 AI 模型的性能、效果、安全合规等维度进行系统性评估的过程。通过多模型可视化批量评测工具解决单模型评测低效问题，引入“裁判员模型”构建自动化、并行化的智能评估

体系。

模型引擎服务：是执行模型推理的核心组件，负责解析模型格式并高效运行推理计算，常用引擎如 ONNX Runtime、TensorRT、OpenVINO、TorchServe 等。

模型推理加速：通过量化、并行计算、剪枝等技术，削减模型计算量与内存占用，核心价值在于降低推理延迟、提升吞吐量。可基于 vLLM、SGLang 等技术构建多模型集成框架，通过专家并行、PD 分离等策略实现大模型推理效率的提升。

多模型协同管理：是在统一的管理框架下，对多个功能各异、部署环境不同的 AI 模型进行统筹调度、版本管控、资源分配与生命周期维护的管理模式。通过搭建标准化的模型注册中心与服务编排平台，实现多模型的按需调用、协同推理。

模型研发生产模块在建设过程中，可能会面对现存的“模型研发周期长、推理性能不足、工具碎片化、协同管理难”等痛点，需锚定业界主流训推技术，以“推理加速引擎、全栈多模态服务、精准模型研发、模型评测体系”为四大战略支点，构建覆盖“模型引擎-部署-微调-评测-管理”的全链路工具链，最终实现推理性能与模型质量的双重保障。

（五）数据供给

数据供给模块是一个面向 AI 原生的综合性数据基础设施，集数据汇聚、存储、处理、标注、合成、质量评测、管理、共享及跨模

态检索等能力于一体，构建覆盖“采—存—治—标—用—评—管”全流程的高效数据供给体系。

图 5 数据供给平台架构

数据汇聚：作为数据入口，支持多源异构数据的统一采集与回流，保障语料来源广泛、更新及时，为大模型训练提供持续且全面的原始数据基础。

多模态存储：面向图文音视频等异构数据类型，打造多模态湖仓一体、图数据库等存储能力，适配多模态数据存储，确保数据存储的高效与灵活。

多模态处理：集成数据清洗、去重、脱敏、价值观合规过滤及跨模态对齐等核心处理能力，通过可编排的数据管线对原始语料进行自动化治理，提升数据一致性、可用性与语义对齐度，为后续标注、合成与训练环节奠定高质量数据基础。

数据标注：集成智能标注、思维链标注等先进工具，紧密贴合大模型团队标注需求，在保障标注准确性的同时显著提升标注效率。

数据合成：基于大语言模型、扩散模型等生成技术，通过数据改写、数据蒸馏、GAN 合成、VAE 合成等手段，构建高质量、任务导向的合成语料，有效扩充原始语料的规模与多样性。

质量评测：建立覆盖完整性、干净性、专业性、多样性、安全性等维度的自动化评估体系，结合规则引擎、统计指标与模型打分，对数据进行细粒度质量量化与问题诊断，实现数据质量的闭环管控与持续优化。

数据管理：提供全生命周期的数据管理能力，包括元数据管理、数据血缘追踪、版本控制等，支持对海量语料的精细化分类、检索与运营，确保数据可管、可控、可追溯。

数据共享：通过数据 MCP 服务、统一数据目录等，打破部门与系统间的数据孤岛，在保障隐私合规与访问控制的前提下，实现跨团队、跨项目、跨平台的高效数据流通与协作复用。

本体建模与迭代：通过构建领域本体模型、自动抽取与对齐多源语义，形成结构清晰、语义一致的知识骨架，支撑数据到业务的精准投射与智能系统的可解释推理。

数据供给平台通过自动化流水线与智能工具链，支持多源异构数据（如文本、图像、音频、视频等）的统一接入与融合处理，为大模型训练提供高质量、多样性的语料资源。

（六）向量数据库

向量数据库是 AI 原生应用的重要数据组件，承担高维向量的高

效存储、检索与管理能力。依托既有的关系型数据库和向量引擎深度融合能力，以及混合查询基础架构，支撑智能知识检索、检索增强生成（RAG）等各种复杂应用场景。向量数据库主要功能包括多模态查询、向量检索、分布式向量等。

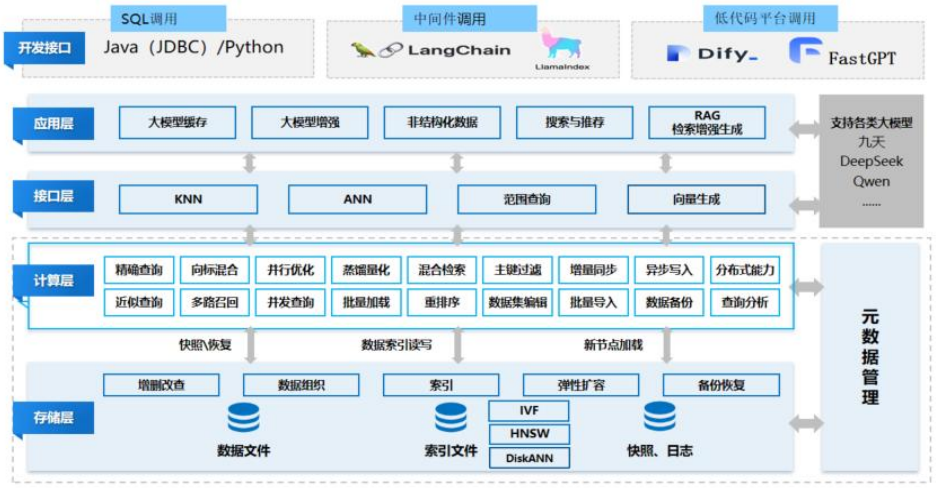

图 6 向量数据库架构

多模态查询：是在一个查询请求中无缝融合标量过滤、全文检索和向量检索等多种查询模式的能力。通过多模态查询可解决单一检索模式的局限性，实现“精准筛选”与“语义扩展”的平衡。在 RAG 场景中，多模态查询可确保系统既能理解用户提问的深层意图，又能严格遵守业务约束条件，返回相关且精准的知识片段，极大提升生成答案的准确性和可控性。多模态查询的关键技术指标包括查询延迟、查询吞吐量、召回率等。

向量检索：是在海量高维向量数据集中快速找到与目标向量相似的多个向量数据的能力，其核心是高效的索引结构和搜索算法。向量检索是决定向量数据库性能的关键因素，直接决定了 RAG、推

荐系统等应用的响应速度和用户体验。高性能向量检索通常采用分层索引与量化技术，在内存中建立导航图（如 DiskANN,HNSW）实现向量数据的高速粗筛,在磁盘上存储精细向量数据以保证存储容量，使用乘积量化等技术压缩向量，减少 I/O 开销，实现内存与磁盘资源的平衡。衡量向量检索效率的关键技术指标通常包括 99 分位延迟、吞吐量、索引构建时间、召回率等。

分布式向量：是将向量数据集自动分片到多个物理节点上，并通过分布式查询引擎协调跨节点搜索任务的能力。可通过分布式向量解决单一节点的存储与算力瓶颈，实现系统的水平扩展，提升系统的高可用性、容错性和弹性伸缩能力，对于支撑企业 TB/PB 级知识库的 RAG 应用至关重要。分布式向量的关键技术包括数据分片与负载均衡策略、分布式查询优化、节点故障自动恢复等。

向量数据库作为 RAG 架构中的“长期记忆”或“知识库”，为 LLM 提供准确、相关的上下文信息，在 RAG 流程中扮演“语义理解与检索”角色。AI 原生基础设施中，向量数据库整体建设理念要注重融合架构、开发者友好两个方面：融合架构需摒弃“向量引擎+关系数据库”的松散耦合模式，在传统关系型数据库的高性能、高稳定性和数据强一致的基础上，增加支持多种类型的向量数据，与标量数据统一存储、统一管理、统一查询，从根本上优化混合查询的性能；开发者友好提供完善的 SQL/NoSQL 接口，深度集成主流 AI 开发生态（如 LangChain 等），降低使用门槛。

（七）智能体引擎

智能体引擎是 AI 原生基础设施中的核心上层建筑，其定位不仅是一个开发工具集，更应致力于打造企业级智能体操作系统（Agent OS）。引擎建设依托中台架构，汇聚全景 AI 能力，通过建设统一的技术标准与协议规范，屏蔽底层基础设施复杂性，为上层应用提供标准化的“系统调用”接口。其核心内涵在于实现从“模型驱动”向“智能体驱动”的范式转变，为各行业场景提供高效的应用开发能力与稳定的运行环境支撑，推动企业级 AI 原生应用规模化快速落地。引擎架构设计自下而上分为 Agent OS 内核层、智能体开发套件层和 AI 原生应用生态。

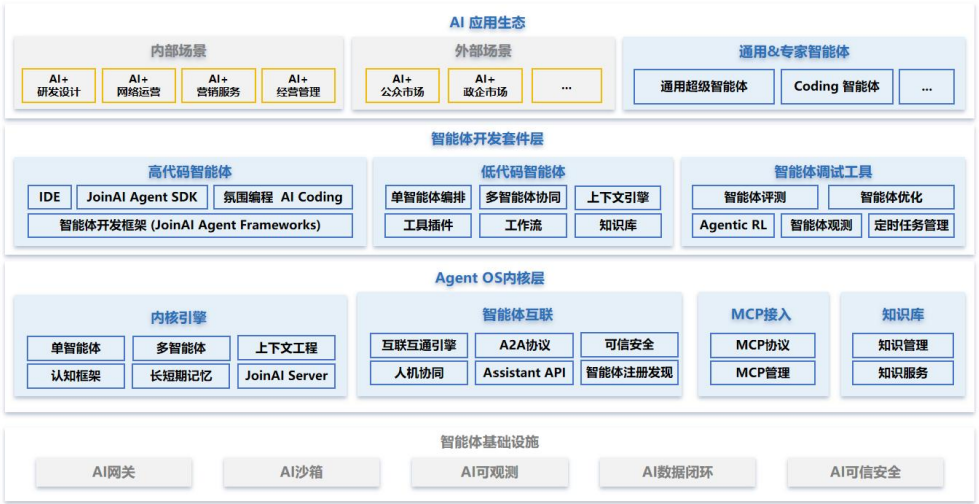

图 7 智能体引擎架构

Agent OS 内核层作为 Agent OS 的“心脏”，负责智能体的调度、通信、资源管理与核心认知，主要包括内核引擎、智能体互联、MCP 接入、知识库四大核心模块。

内核引擎：是智能体的核心控制单元，负责模拟人类的认知过程，集成认知框架，提供标准化的认知处理流，既支持单智能体的独立思考，也支持多智能体的复杂协作。引擎内置长短期记忆模块，使其具备跨会话的记忆保持能力，能够从历史交互中持续学习。引擎提供上下文工程能力，可动态管理模型上下文窗口，利用智能压缩与检索技术，确保在长窗口交互下关键信息不丢失，从而显著提升决策准确性。

智能体互联：定义了智能体社会的“通用语言”与交互规范，致力于实现不同平台、不同架构智能体之间的互操作，支持 A2A、REST、gRPC、流式消息等多种能力交互协议，屏蔽不同能力实现形态带来的差异，使智能体可与其他智能体进行能力协作。基于能力语义描述、运行状态和上下文约束，通过智能路由在多智能体节点之间动态选择最优调用路径，支撑跨模型、跨平台、跨区域的能力互联。此外，智能体互联能力定义了人机协同与 Assistant API 标准接口，支持 Human-in-the-loop 反馈机制，并通过可信安全与注册发现机制，实现智能体身份可信及动态寻址。

MCP 接入：作为规范 AI 模型在推理、协同过程中上下文信息传递与交互的标准协议，MCP 为各类工具的协同提供了统一协议参考。支持 MCP 服务的接入，使智能体可以直接发现和调用 MCP 能力。

知识库：作为智能体的“外部大脑”，支持检索增强生成技术，允许智能体挂载企业私有文档与结构化数据，为智能体的推理过程

提供领域专业知识，增强智能体在专业领域决策能力的同时也起到减少幻觉的作用。引擎支持知识管理能力，包括权限管理、版本管理、知识索引、异常监测等，并提供知识检索、查询及推理等服务。

智能体开发套件层（Development Kit）主要为不同层次的开发者提供高效的智能体开发工具,主要包括高代码智能体的深度定制、低代码智能体的快速构建与智能体调试三大场景。

高代码智能体：面向专业算法工程师与全栈开发者提供高代码智能体开发能力，具有定制灵活性。高代码智能体的开发通常依托智能体开发框架，内置多种智能体设计模式（如 ReAct、AutoGPT、COT 等），允许开发者利用代码精细控制智能体的每一个行为细节。可面向智能体开发人员提供标准化的 SDK（如 JoinAI Agent SDK），加速复杂应用落地。

低代码智能体：为降低开发门槛，面向业务分析与产品设计，提供低代码快速构建能力。支持通过可视化编排工具，以拖拉拽的方式定义单智能体逻辑或多智能体协作流程（SOP），配合上下文引擎，自动管理对话上下文，简化提示词工程的复杂度。提供组件化配置功能，允许用户通过图形化界面灵活配置工具插件、工作流与知识库，实现“搭积木”式的应用构建体验。

智能体调试：针对智能体开发中常见的“黑盒”难调试痛点，智能体引擎需提供一套完整的调试工具链。智能体调试模块提供基准测试集，支持自动评估智能体在准确性、安全性等方面的表现，

并给出优化建议。通过引入 Agentic RL 机制，利用环境反馈自动优化智能体决策策略。通过智能体观测工具实现执行轨迹的可视化回放，帮助使用者直观理解智能体的决策路径与逻辑漏洞。

应用生态层的关注重点是如何将智能体引擎的核心能力转化为面向最终用户的具体应用,实现智能体能力的最终交付与价值释放，包括通用超级智能体与 Coding 智能体等。

通用超级智能体：利用多智能体架构打造通用超级智能体，支持精确的多模型协同及上下文管理，具备丰富的工具集合，并针对不同业务场景持续优化算法集合，具备模糊推理/时间推理能力、多源信息查询及推理能力、Agent-Code 协同能力等。

Coding 智能体：专精于代码生成与软件工程任务，赋能编程场景。与普通辅助工具不同，Coding 智能体能够贯穿智能应用建设的全流程，实现从需求分析、智能体设计、智能体开发到智能体发布的全流程智能化，最大程度降低智能体应用构建门槛。

智能体引擎建设应遵循四大核心设计理念，以支撑“Agent OS”的愿景落地。首先是生态开放协同，依托中台架构汇聚全景 AI 工具与能力，构建开放式架构体系，向下兼容异构算力与模型，向上支撑用户个性化定制需求，打破技术孤岛。其次是企业级生产保障，区别于实验性框架，建立研发测试生产全流程的严格保障机制、多租户多环境隔离及 SRE 服务体系，确保智能体在复杂企业环境下的高可用性与稳定运行。同时打造极致开发体验，推行“积木式组合

搭建”与“高低代码混合开发”模式，支持用户通过全生命周期的一站式管理快速构建场景化应用。最后，推动核心智能引擎进化，全面升级智能体认知框架与群体协作能力,重点强化智能体在感知、规划、决策、行动、记忆与自主进化六大维度的能力，实现从“工具型”向“认知型”智能体的跃迁。

(八)AI 网关

AI 网关是 AI 原生基础设施构建的核心要件。其核心功能包括 API 路由、模型代理、智能体枢纽、MCP 调度、AI 流量分析等模块，具备丰富的服务集成和全生命周期治理能力。在最终用户、AI 应用和模型之间发挥枢纽作用,实现了 AI 应用研发生产要素的高效调配、为业务提供应用级精细化运营支撑。

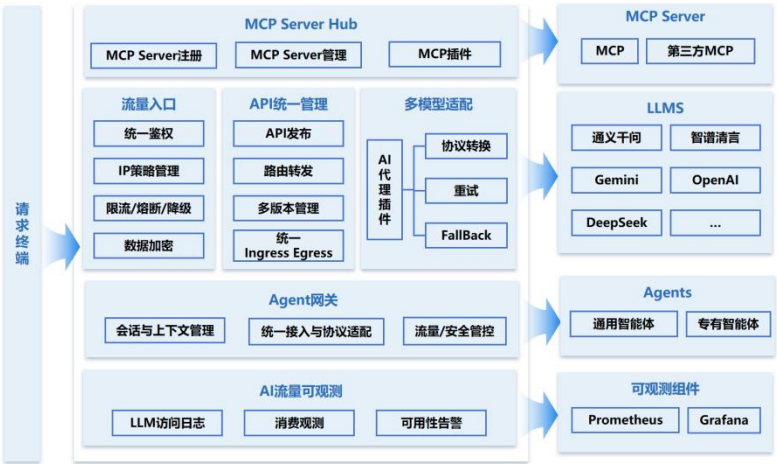

图 8 AI 网关架构

API 路由：是从外部客户端（如 AI 应用、终端用户、第三方系统等）指向 AI 系统内部服务（如大模型集群、MCP Server、AI 智能体业务层等）的请求流量入口，以及从系统内部返回给外部客户

端的响应流量入口。作为承接这类跨系统边界流量的统一接入点，API 路由提供统一鉴权、IP 策略管理、流量的限流、熔断和降级管理、数据加密、API 发布、路由转发、多版本管理等功能，为系统提供统一、全面的边界治理能力。

模型代理：是专门针对各类 AI 模型（如大语言模型、推理模型等）的核心代理模块。作为客户端应用与多源模型之间的中间层，模型代理封装不同模型的接口并提供统一接入方式，同时统筹模型调用的各类管控操作，是衔接应用与模型的关键枢纽。模型代理通过统一接口与协议，降低集成与迁移成本；通过重试机制，提升模型调用稳定性；通过 FallBack 机制，在主模型异常时可进行自动切换兜底，从而保障服务连续性。

智能体枢纽：是适配 AI 智能体交互场景的专项功能模块。作为智能体与模型、工具、外部系统及其他智能体的统一交互枢纽，基于 MCP、A2A 等专属协议完成协议适配与标准化接入，同时提供会话上下文管理、精细化安全权限管控、流量治理及全链路可观测能力。智能体枢纽可以屏蔽各层服务所使用的技术差异、降低多智能体集成成本，保障智能体协同的稳定性、安全性与可运维性。

MCP 调度：包括模型上下文协议（MCP）架构下的服务端托管模块，专门用于集中化管理多个 MCP Server，统筹协调 MCP Server 注册，统一调度请求，管控服务生命周期，从而保障系统稳定。

AI 流量分析：是针对 AI 模型调用场景的运营支撑能力，通过网关收集、整合并分析 AI 流量的全链路数据，涵盖调用指标、请求日

志、交互内容等关键信息，同时遵循 OpenInference 等 AI 专属规范进行数据标准化呈现，实现 AI 流量监控、问题追溯与行为分析等。

(九)AI 原生应用开发管理

AI 原生应用开发管理将 AI 能力深度嵌入项目管理，覆盖需求、设计、开发、测试、部署的全过程，提供涵盖项目管理相关的智能化能力。从辅助研发升级为 AI 自主化操作，重塑研发交互方式，提升研发效能。

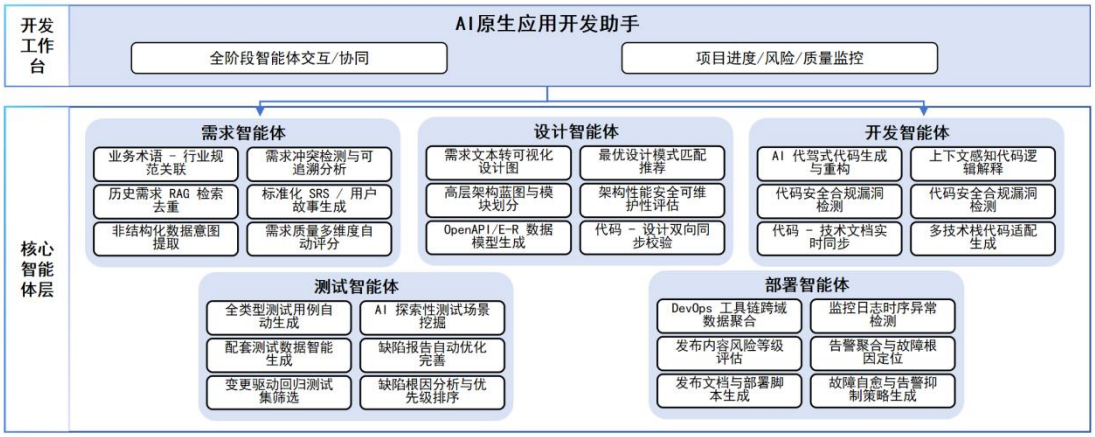

图 9 AI 原生应用开发管理

需求：深度融合意图识别与需求知识图谱技术，将业务术语与行业规范建立关联。需求智能体利用检索增强生成（RAG）检索历史数据，并运用冲突检测与可追溯性分析技术，从用户反馈、会议记录等非结构化数据中自动提取意图，生成标准化的 SRS 或用户故事。内置的质量智能评估引擎可实时对需求的完备性、一致性与可测试性进行自动评分，实现高质量的需求挖掘。

设计：基于多模态能力与知识驱动决策，通过代码-设计双向

同步技术，将文本需求转化为 Mermaid 或 PlantUML 可视化设计图。设计智能体能根据规格自动推演高层架构蓝图,生成模块划分建议、OpenAPI 接口定义和 E-R 数据模型等。结合设计模式匹配能力，设计智能体可推荐最优方案，并针对性能、安全及可维护性的架构提供智能评估，实现从需求到设计的无缝转化。

开发：开发阶段采用“AI 代驾”模式，依托代码领域大模型，在保留人类决策权的前提下提供行级补全、函数生成和代码重构等服务。开发智能体具备安全合规性检查能力，可自动识别漏洞并同步更新文档，实现代码与文档的实时一致。作为 IDE 中的智能伴侣，开发智能体具备上下文感知能力，能够进行代码解释、错误精准定位及调试辅助，大幅提升编码效率与安全性。

测试：基于风险驱动设计与全链路追溯技术，测试智能体结合代码覆盖率分析，可自动生成单元、功能及集成测试用例与数据，并根据代码变更范围智能筛选回归测试集。支持 AI 驱动的探索性测试，在执行测试后，可自动优化缺陷报告（填充环境信息、复现步骤），并利用缺陷根因定位技术对 Bug 进行分析与优先级排序，极大缩短修复周期。

部署：通过跨领域数据聚合技术，部署智能体能深度理解 DevOps 工具链（Git、CI/CD）数据，根据变更内容评估风险等级，自动汇编发布说明书及标准化部署脚本，并通过接入监控日志，利用时序数据异常检测识别故障模式,实现智能告警聚合与根因分析，并主动生成故障自愈建议与告警抑制策略，保障系统高可用性。

(十)AI 原生运维

AI 原生运维是面向 AI 原生的全栈可观测运维体系，可精准监控模型行为、快速定位故障、科学评估输出质量，保障生产环境中模型的可靠运行，实现模型性能与业务需求的深度融合，确保 AI 应用高效稳定安全运行。AI 原生运维体系涵盖全栈可观测、AI 评估、告警治理、资源中心、指标采集等核心能力。

全栈可观测：以探针埋点技术为基础，具备零代码接入、token 成本分析、端到端链路追踪能力，可呈现智能体内部、推理引擎内部工作流的详细执行过程，包含调用 LLM 和 MCP server，以及输入输出情况。通过串联用户终端、AI 网关、模型应用、模型服务、数据存储、通智算基础资源等多个层级，采用全路径还原、多维度关联、上下文透传等技术，将 AI 应用内部流程变得“可感知”，实现全链路 LLM Trace 串联。

AI 评估：通过构建多维度、全流程的自动化评估能力，实现大模型输出质量的自动化验证，降低人工审核成本。一套完整的评估体系涵盖性能指标评测、鲁棒性与泛化能力测试、偏见与公平性评估审查、可理解性与可解释性分析、合规与伦理风险筛查、持续监测机制和决策框架的构建等。结合裁判模型、用户反馈、人工标注等多元评估手段，支持文本简洁度、上下文正确性、事实准确性等

关键质量维度的量化指标动态追踪，实现 AI 评估的多范式全景。

告警治理：用 AI 代替人工经验，实现告警管理的“精准、智能、自动、预测”。通过语义理解、上下文关联、历史数据学习等技术实现从“海量告警”到“有效告警”的关键筛选，通过大模型的深度推理、多维度数据关联、威胁情报整合、动态规则优化等方式，实现专家级研判。

资源中心：通过统一建模（对象标准化定义）、统一接入（资产统一纳管）、统一调和（资源数据一致），将 AI Native 全域异构资源（包括智算资源、通算资源、存储资源、网络资源）纳入资源中心进行集中管理，实现全域可视、可管、可用。

指标采集：提供 AI 应用与服务的无侵入、低成本、高质量的指标采集能力，以字节码增强技术（Java 语言）、monkey patch 机制（Python 语言）、插桩技术（Go 语言）等方式实现智能体多框架埋点，注入可观测数据采集逻辑。

AI 原生运维体系以全栈可观测为核心，打通从用户终端到基础设施的完整链路，实现从问题发现到决策修复的全生命周期管理。通过全栈可观测、智能化闭环和标准化协同三大核心架构，为 AI 应用提供高可靠、高性能的运行保障，通过数据驱动与智能协同，构建面向 AI 原生时代的运维新范式。

（十一）AI 安全保障

AI 安全保障是 AI 原生基础设施安全、可靠、可信运行的核心保障体系，助力系统应对提示词攻击、数据投毒等新型风险，确保 AI 行为可控、输出合规、运行可信。

应用层防护：聚焦输入安全，通过意图识别、频率控制、资源熔断等技术防御恶意诱导、炸弹指令及第三方污染数据攻击，确保交互过程可控。

模型层防护：确立合规底线，结合多模态内容审核、敏感信息动态脱敏、数字水印嵌入确保输出内容安全可溯，并构建提示词攻击防御、恶意文件检测、URL 拦截等多重威胁防御机制，同时结合越狱检测、幻觉抑制、反爬机制保障模型健康。

数据层防护：贯穿采集至销毁全流程，在采集阶段实施分类分级、脱敏去毒，传输阶段采用 VPC 加密、TLS 协议及最小化解密策略，存储阶段实现隔离加密，访问阶段执行最小权限控制，处理阶段进行实时过滤，删除阶段确保完全清理。

系统层防护：夯实基础设施安全，通过主机安全客户端、端口管控强化基础环境，利用安全沙盒隔离容器，借助镜像扫描与签名校验保障供应链安全，并基于防火墙与零信任网络实现内外网流量管控。

各层能力通过统一安全运营中心进行集中监控、智能分析与协同响应，形成闭环安全管理，整体实现从被动防护到主动风险评估的

安全范式转变。一些关键指标包括攻击识别准确率超过 99.5%、审核延迟小于 200ms、PII 识别覆盖度达 20 类以上等。

（十二）数字可信

面向 AI 原生背景，协同治理加速演进，基础设施的核心瓶颈正在从“可用”转向“可信”，迫切需要在“算力—数据—模型—应用”全生命周期系统化融入可信能力，为此，需以数字可信体系为基石，构建覆盖“可信算力协同、可信数据供给、可信模型训推、可信应用治理”并以“可信测评体系”贯穿支撑的总体架构，形成面向 AI 原生的可信体系。

图 10 数字可信架构

基础设施可信：在基础设施层结合区块链、隐私计算等技术，构建链计算平台，提供开放的可信隐私计算服务，解决算力资源分散、利用率偏低以及隐私计算难以规模化的问题，为 AI 训推、智能体沙箱提供执行环境，为数据预处理与训练、模型保护与共享、智

能体隐私信息处理提供安全可信的计算支撑。

数据内容可信：在数据内容层基于数字可信构建覆盖“可信采集、可信标注、可信清洗、可信处理、可信流通、可信审计”的数据治理可信能力集，通过数字可信基础设施对数据采集、加工、流转等关键环节进行全程可信记录，配合加密存储、分级分类与多方安全计算等机制，形成“可用不可见”的数据可信流通能力，为模型训推提供安全可控的数据加工环境，构筑高效、安全、可控的可信数据能力。

模型训推可信：在模型训推层构建面向模型训推全流程的风险识别与拦截工具，打造覆盖“事前预防—事中追踪—事后审核”的全链路安全工具集，针对模型训练阶段可能出现的数据投毒、恶意样本混入、异常分布等问题实施风险识别与拦截，对模型推理阶段的恶意提示词注入、对抗攻击、异常调用行为等风险进行持续监测与证据留存，实现“风险可感知、可管控、可追溯”的闭环管理，打造 AI 可信训练场、隐私训推体系和风险可视化工具，为模型安全提供系统化保障。

可信应用治理：在可信应用治理层面向大规模、分布式的智能交互场景，通过融合区块链、隐私计算、可信身份等关键技术，打牢虚实共生、智能协同的数字可信能力，构建“身份可认证、记忆可留存、行为可追溯”信任协同体系，进而打造集“数字身份、数据资产、可信流通”于一体的“记忆银行”。一方面面向智能体场

景，构建“可信身份、可信行为、可信决策、可信互联”能力体系，为各类智能体提供统一身份认证、行为记录与审计、决策过程可解释、跨系统可信互联等支撑，确保人机协同与多智能体协同在可信环境下运行；另一方面面向 AIGC 场景，打造以 AIGC 水印、AI 对抗、伪造检测与 AIGC 版权治理为核心的内容安全能力，切实保障系统与用户安全。

可信评测：面向 AI 安全的可信测评搭建覆盖多维指标的一体化测评体系，形成数据质量测评、大模型可信测评、智能体可信测评与 AI 应用测评等能力组合，同时构建自动化测试工具与高质量、可复用的测评数据集，实现对数据可靠性、模型安全性等关键指标的持续评估与对比分析，为安全可信体系提供可量化、可验证的技术依据，支撑“发现问题—评估影响—采取措施—复测验证”的治理闭环。

可信能力通过上述架构嵌入基础设施、数据内容、模型训推和应用治理各环节。基础设施层提供可验证的算力底座，数据内容层夯实可信数据根基，模型训推层构建“事前—事中—事后”安全闭环，可信应用治理层保障智能体与 AIGC 等在可控边界内安全运行，AI 安全可信测评为整体运行提供量化评估与持续改进支撑。通过多层协同与测评贯通，构建结构清晰、可审计可追溯的可信体系，为 AI 原生基础设施发展提供安全与信任保障。

四、 行业实践案例

(一)通信行业：中国移动 AI 原生基础设施实践

案例背景：落实国家“AI+”战略行动意见要求，中国移动集团全面贯彻数智化战略，积极开展“AI+”行动计划，集团数智化部牵头，联合集团相关业务部门、下属全量省公司和专业公司的业务团队、技术团队，协同建设创新 AI 原生基础设施，旨在实现三个目标。目标一，面向云边混合架构提供可靠算力、数据资源供给。建设和优化异构算力池（CPU/GPU/NPU 协同），实现异构算力一体化调度，同时结合低时延网络，打造 AI 驱动的可服务型智能存储引擎和高质量数据供给平台，实现基础资源的自我管理与优化。目标二，优化 AI 开发及运行全生命周期效率。重塑 AI 开发工具链，实现 AI 原生应用的数据预处理、模型训练、推理部署、智能体开发、一体化自动运维端到端加速和可视可管可控。目标三，强化 AI 运营运维和安全合规治理。通过 AI 原生的统一运维、安全保障和数字可信手段，建立健全分级运营运维和审计治理体制机制，借助体系化、规模化、标准化、智能化优势，构建全链路治理体系，为数智化发展奠定坚实基础。

技术架构：中国移动聚智智能体平台基于丰富的 AI 原生基础设施能力,构建了以Agent OS 为核心的一体化智能体整体解决方案。涵盖多场景应用构建工坊、高/低代码开发套件、集运行时服务托管

与协议支撑的智能体操作系统及全栈基础设施支撑，实现智能体的高效开发、稳定运行与规模应用自研多智能体协同框架，构建稳定通用的开发与管理底座，核心能力超级智能体引擎（JoinAI Agent）10 月登顶 GAIA 全球榜单。高低代码结合，低代码快速编排验证，高代码灵活构建部署，支持自动集成、调试与优化。打通 AaaS+生态，汇聚多样 AI 工具能力，助力智能体应用快速搭建。高可用架构+容灾备份，配合加密与分级管理，保障系统安全稳定 7×24 运行。全链路监控与可视化分析，实时洞察运行状态与业务价值，助力智能体持续优化。

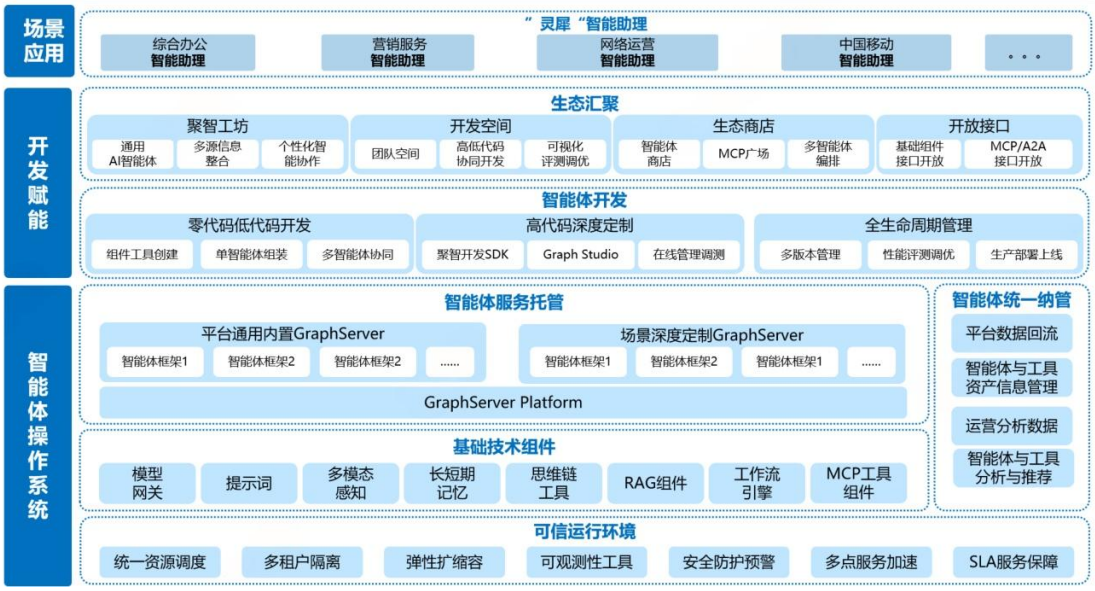

图 11 中国移动聚智智能体平台架构

应用成效：截至 2025 年 12 月，平台已面向中国移动体系内超过 90%的省专单位持续提供服务，已研发智能体数千个、研发和引入工具近万个，打造数百个标杆智能体应用，覆盖经营管理、网格运营、研发设计、营销服务、通用赋能等 5 个领域数十个业务场景，输出多套可复用的通用智能应用解决方案,有效推动 AI 原生应用在

一线业务中的融合应用与价值转化，平台累计调用量超百万次。同时，面向广大生态伙伴汇聚共性通用数智能力超 400 项，辐射精准营销、业务查询、运维告警、娱乐交互等三十余类业务场景，逐步推动 MCP、智能体等数智能力赋能全场景应用。

（二）通信行业：中国移动某省灵犀助手实践

案例背景：AI 正加速渗透至各行各业，驱动企业服务模式和生产方式的深刻变革。作为数字经济的重要支撑力量，运营商不仅承担着提供通信连接的基础职能，更需要在确保网络与业务稳定运行的基础上，不断提升客户体验与营销效率。面对业务场景的快速演化与用户需求的个性化、多元化，传统依赖人工操作、系统割裂的服务模式，已无法适应数字化竞争的新形势。

技术架构：灵犀助手是中国移动基于 Agentic AI 架构，使用九天大模型为客户经理打造的智能化工作伙伴，贯穿客户经理的一天，覆盖前置准备、现场推进与后续提升全流程，为业务办理与服务执行提供连续、精准的智能支撑。对客户经理而言，它既是数据参谋，也是业务助手，更是学习教练，让日常工作更加高效顺畅。前置准备阶段能快速响应自然语言指令，实时提供经营数据、提醒信息，支持一键查看商机等待办事项并快速生成组网方案与报价，为当日客户拜访和业务洽谈铺垫；现场推进时段提供贯穿式辅助，沟通时可即时查询产品信息与营销案例，办理业务时能用自然语言指令完成相关操作且支持智能填参,服务环节能对话查询订单进展、

协助处理投诉报障，保障服务交付；持续提升阶段通过学习陪练能力，助力客户经理开展模拟练习、知识学习和考试复盘，强化业务理解与沟通技巧，实现专业能力持续提升。

图 12 中国移动灵犀助手功能

应用成效：中国移动灵犀助手围绕客户经理的一天工作为主题进行全流程智能化重构。早晨准备拜访期间，灵犀助手准确帮助客户经理迅速获取各类集团经营信息与资料，查询效率提升 3 倍。集团报告生成从原先的天级缩短至分钟级，累计生成数百份集团洞察报告，使客户经理在开始一天的拜访前即可掌握客户画像、业务状态与关键任务，大幅提升工作准备的充分度。业务推进期间，灵犀助手将核心办理流程实现智能化加速，单业务办理耗时效率提升 60%。商机处理、合同推进、订单查询均可通过多智能体协同快速完成。目前灵犀助手用户规模已达数万人，每月调用量超过上百万次，覆盖超过 200 家省级战略客户与近 10 万家地市客户，已成为客户经理日常工作中最核心的智能化生产力工具。

（三）通信行业：中国移动某省大模型网关实践

案例背景：随着集团“AI+”战略深入推进，通信行业智能化转型进入规模化落地阶段，但大模型应用过程中面临多重痛点：企业内部需同时调用 Qwen 系列、DeepSeek 系列、九天等多类型大模型，业务侧适配成本高；不同模型特性差异大，场景化选择难度大；存在提示注入、敏感信息泄露等安全风险，合规管控压力突出，难以满足业务并发需求。为解决“适配烦、选择难、安全不可控、响应时间长”等问题，上海移动建设大模型网关项目，打造统一、安全、高效的大模型服务中枢，支撑通信行业及政企客户智能化升级。

技术架构：项目采用“统一汇聚-智能管控-数据驱动”的设计理念，构建“能力增强-监控运营-市场服务-数据分析”四层架构。核心层以大模型网关为枢纽，接入集团磐智算力部署的 Qwen、DeepSeek 等模型及本地算力部署的各类模型，通过统一 OpenAI 协议实现多模型热插拔；增强层优化 Token 配额管理、智能路由、语义缓存等核心能力，支持按租户并发控制、负载均衡及熔断降级；监控层搭建全链路可观测体系，覆盖健康度、响应时延等多维度指标；服务层打造大模型市场，提供自助订阅、在线体验、版本对比功能。

图 13 中国移动大模型网关架构

应用成效：项目实现多模型统一管控与高效调度，业务侧大模型适配联调效率提升 90%，大幅减少新模型版本接入工作量，年人均成本节省超过 10%。服务稳定性显著增强，生产环境异常请求发现时长从小时级压缩至分钟级，效率提升 95.8%，异常请求拦截率达 95%，故障发现率超 70%。响应性能大幅优化，通过语义缓存加速同类请求响应，结合智能路由均衡负载，大模型计算成本显著降低。

（四）通信行业：中国移动某用户运营智能体实践

案例背景：中国移动某省 10086 热线每月产生数以百万计个来话，其中少量来话将生成投诉工单，对于未立单的通话，存在大量的问题盲区。传统的人工抽查方式，合规检查覆盖率低，主观依赖性较大，溯源分析成本较高。需要采用智能化的手段对热线来话进行全面“回放”、溯源、修复。

技术架构：中国移动“AI+回声”项目依托智能体平台，打造

“AI+回声”用户运营智能体，解决投诉的源头治理问题。基于“全量”10086 来话，重塑“听音、回声、溯源、根因修复”四大流程。从用户诉求识别、用户态度感知等维度，及时识别风险，输出服务推诿、违规收费等数百个标签，开展满意度修复、优化营销设计、辅助业务决策,实现了投诉治理“抓早、抓全、抓小”。

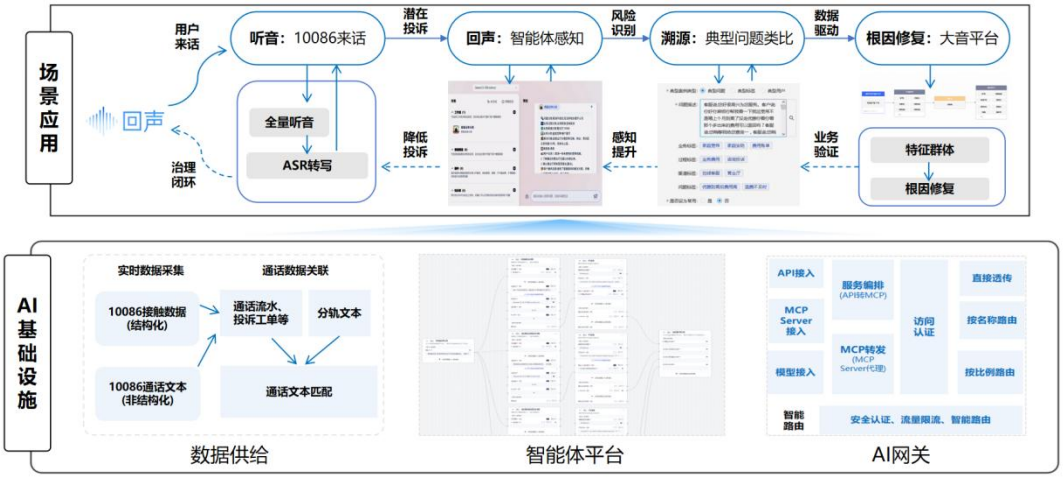

图 14 中国移动用户运营智能体系统架构

应用成效：上线以来，成果月均调用量达数百万次。投诉处理用后即评满意度同比改善 4.84pp，重复投诉率 1.88pp，同比改善 40%。该成果荣获中国互联网协会“2025 年智能体创新应用”典型案例、中国移动党组“领题破题、合力攻坚”优秀实践项目表彰等荣誉。

（五）通信行业：中国移动某省智能体助手实践

案例背景：在当前国家将 AI 提升至战略高度、推动“AI+”与经济社会深度融合的背景下，各行业智能化转型步伐加速。政策驱动之下，企业对 AI 基础设施的需求日益迫切，亟需通过集约化、平

台化的智能能力供给,实现业务创新与效率提升。以“九天大模型”为代表的基座模型正成为构建行业“智慧大脑”的核心引擎。然而，当前许多企业在报告生成等关键环节仍面临响应滞后、个性化不足与数据孤岛等现实矛盾，传统作业模式难以满足分钟级响应的业务需求，制约了智能应用的规模化落地与价值释放。

技术架构：项目遵循“分层解耦、智能驱动”的 AI 原生基础设施设计理念，构建了一套支撑企业级智能应用高效开发与运行的平台化体系。系统以通智算基础资源为底座，向上通过通智算调度引擎实现资源的智能编排与弹性供给。在此基础上，构建了涵盖模型研发生产、多模态数据处理、智能体开发与管理、以及全栈可观测与安全的全链路支撑能力。功能架构可分为资源与部署层、模型与数据层、开发与运维层：资源与部署层依托算力虚拟化与通智算混合调度，实现资源的高效利用与任务的灵活部署，支撑高并发、低延迟的 AI 服务;模型与数据层通过模型引擎服务与多模态处理存储，提供从模型推理加速到知识库构建的统一能力，为上层应用提供智能内核；开发与运维层基于智能体开发工具链和 AI 原生运维体系，实现智能应用的快速构建、迭代与可靠运行，并通过 AI 网关与全栈可观测保障系统可控、可信。本项目采用“模型即服务”与“智能体即应用”的架构模式，通过标准化的 API、MCP 协议及向量检索等技术，打通从数据、模型到场景应用的完整闭环，实现了智能化能力的集约化供给与敏捷化创新。

图 15 中国移动智能助手架构

应用成效：项目成功实现了效率的指数级提升，将传统需数小时乃至数日的报告编制流程缩短至分钟级，使营销人员能够通过简易指令“一键生成”结构完整的初版报告。此举不仅将团队从繁琐劳动中解放，转而聚焦于高价值创意与策略工作，更打造了应对市场变化的敏捷响应能力，显著增强了企业决策效能。成果已产生广泛的内部影响与行业示范效应,相关功能已上线多个核心业务系统，累计调用量数万次，其成熟的子智能体模式已成功推广至其他兄弟省份，充分验证了解决方案的可复制性与巨大推广价值。

（六）政务行业：某地方政府政务智算平台建设

案例背景：在国家“数字政府”建设提速、“十四五”政务信息化规划指引下,省级政务云面临政务服务智能化升级的迫切需求。传统政务 IT 系统存在算力资源分散、AI 基础设施国产化率低、大模型应用链路割裂等问题，难以支撑“一网通办”“智能审批”等场景的智能化演进。同时，政务数据安全与国产化替代政策驱动下，亟需构建统一、开放、安全的智算底座，实现从算力供给到模型服

务的全栈能力升级，推动政务服务从“数字化”向“智能化”平滑转型。

技术架构：项目基于“分层解耦、一云多芯、全链路覆盖”设计理念，构建“基础设施-IaaS-PaaS-MaaS”四层架构。基础设施层部署国产 GPU 服务器、高性能 RDMA 网络、分布式文件存储，实现异构算力资源池化；IaaS 层通过飞天云计算操作系统统一调度，提供“训练+推理”高性能算力服务，支持单机/多机 GPU 租用；PaaS 层整合大数据、数据库、中间件等能力，为模型生产提供底层支撑；MaaS 层覆盖模型训练、微调、部署、测评全链路，开放 60+种 GPU 资源规格及大模型服务目录，支持 API/插件等多样化调用。架构设计以国产化为核心，通过“硬件开放兼容+软件统一调度”，实现算力、算法、模型服务的全流程贯通。

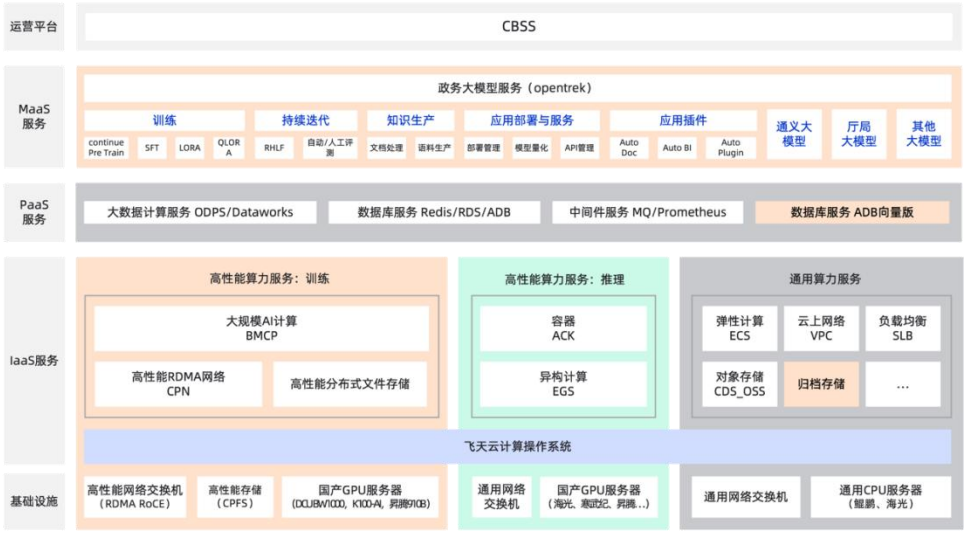

图 16 某地方政府政务智算平台架构

应用成效：项目构建了全国产化、多芯融合的政务智算平台，实现 10 种国产 GPU 大规模融合调度，GPU 资源规格覆盖 60+种，支持单机/多机算力按需供给。MaaS 服务覆盖模型训练、微调、推理

全链条，助力“政策生产辅助”“政务知识库”“智能审批”等场景快速落地，政务知识生产效率提升 200%，模型推理延迟降低 70%。作为省级政务云智能化标杆，该项目推动政务领域国产化算力替代进程，为全国“一网统管”“AI+政务”提供可复用的架构范式，入选省级数字化改革典型案例，形成政务智能化转型的“省级样板”。

(七)政务行业：某省会人工智能政务大模型平台建设

案例背景：近年来，该市政务领域数字化建设成果显著，但一定程度上存在“技术门槛高、重复投资”等问题，为此中国移动联合该市数据局筹建人工智能政务大模型平台，通过整合数据、算法和算力等基础设施资源，破解技术瓶颈与协同难题，有效支撑“一网通办”“一网协同”“一网统管”等场景的深化建设，推动南京政务服务智能化、集约化。

技术架构：通过整合模型管理、知识管理、智能体调度及能力共享等环节，构建“1+4+3+1+N”架构的 AI 通用能力平台，并与多个外部系统对接，实现政务服务的精准赋能。平台包括：1 套标准规范（涵盖大模型应用数据、安全与构建规范）；4 大中心（模型管理中心、知识中心、智能体中心、AI 能力中心）；3 个 AI 能力工具（OCR 识别、语音识别、合成服务及联网搜索服务）；1 套安全防护体系（覆盖大模型部署、输入输出和应用全链路）；以及基于平台开发的 N 个标杆应用（提供技术咨询、数据治理、代码编写等服务）。目前已与多个外部系统对接，包括政务云算力平台、“我的

南京政务版”用户体系、城市之眼视觉算法模型及其他现有智能体应用。

图 17 南京市人工智能政务大模型平台架构

应用成效：整合数据、算法和算力等 AI 基础设施资源，通过集约化建设减少重复投资，实现“一次投资，全市共用”的城市级 Maas 平台的构建。打造华东区首个城市级政务大模型平台标杆，形成可复制、可推广的“南京模式”，推动政务服务从“分散建设”向“统筹共享”转型升级。

（八）制造行业：某头部车企智能客服系统建设

案例背景：该企业是中国领先的合资汽车制造商，在面向车主的咨询服务场景中，原有客服系统面临效率不足问题。大量基础性问题依赖人工客服解答，导致运营成本高且响应速度慢。

技术架构：基于腾讯云智能体开发平台的全链路自研 LLM+RAG 技术，将车型手册、维修指南等非结构化数据转化为可检索知识库。运用 OCR 大模型、多模态理解模型提升知识检索精准度，支持用户

上传故障图片自动识别关键信息。平台集成到 APP、小程序、官网等渠道，通过挖掘历史客服对话记录补充知识库，降低维护工作量。

图 18 某车企智能客服系统架构

应用成效：通过大模型能力提高 C 端咨询的机器人独立解决率，减少人工接待会话量，节省客服中心人力成本。智能客服实现 24 小时服务，使人工客服更专注于复杂问题处理，提升客户满意度。智能客服机器人独立解决率从 37%提升至 84%，月均自动解决客户咨询问题 1.7 万次。问答准确率达到 84%，大模型出图率 70%，覆盖车辆使用指导、故障诊断等高频率场景。

（九）制造行业：某新能源车企 AI 数据专家

案例背景：随着智能电动汽车行业向高阶智驾与智能座舱深度演进，车企对多模态数据的高效管理与 AI 模型迭代提出更高要求。某汽车企业作为集团旗下高端智能新能源品牌，聚焦高阶智驾模型迭代与用户体验优化，需处理海量车端、雷达、点云等多模态原始

数据（如车辆运行监控、零部件管理、智驾模型训练等场景）。然而，传统数据链路存在召回率低、并发阻塞、架构冗余等痛点，难以支撑模型快速迭代需求。行业智能化趋势下，亟需构建高效的 AI 基础设施，实现数据的低延迟检索、统一管理与模型精调加速。

技术架构：项目采用“多模态数据湖+Data Agent 智能分析”双轨技术架构，覆盖数据管理与业务应用全链路。针对智驾数据场景，采用“向量+标量”混合检索方案解决长周期数据遗漏问题；通过“向量检索链路重构+资源动态控制+存储层优化”三重策略提升并发稳定性；结合“MPP 架构+湖仓一体”统一点查与统计数据，减少存储冗余。面向业务场景（如备件管理、车辆监控），以豆包大模型为核心，构建自然语言交互的智能数据平台，支持数据连接、智能问数（业务问题转 SQL）、深度分析等能力，实现业务人员自助查询与个性化分析。

图 19 某汽车企业 AI 数据专家技术架构

应用成效：项目显著提升数据效率与模型迭代能力：多模态数据召回率从小于 45%跃迁至 90%以上，查询性能提升 20 倍，存储成本与运维复杂度降低，闭环周期从 4 周缩短至 5 天。业务应用方面：

Data Agent 实现问题排查时间从 3 天缩短至 1 分钟，风险拦截提前 72 小时，业务沟通成本减少 10%-20%，交付成本降低 5%，支撑业务人员自助分析需求。本次项目案例验证了火山引擎 AI 基础设施在智能汽车领域的适配性，为行业提供“数据湖+智能分析”双轮驱动的参考范式，助力打造智驾技术壁垒与用户粘性。

（十）制造行业：某具身智能公司平台建设

案例背景：AI 与机器人技术的深度融合，智能体与大模型让机器人从被动编程走向主动决策，能够执行复杂的多阶段推理任务。技术路线方面，分层式一度为主流；伴随数据采集、模型泛化与推理响应问题逐步解决，端到端路线在未来有望成为主流。训练与落地仍面临数据与模型能力的多重挑战（感知/执行/学习/自适应、硬件性能、验证方法等），需依托高质量数据与开源/联盟建设，加速仿真与真实数据的融合应用。随着具身机器人大面积铺开与新场景叠加，常态化多场景模型训练与大规模推理成为刚需，面临诸多挑战，如线下机房资源不足，无法支持高密度训练任务，资源利用率与运维稳定性对初创企业挑战大，具身智能仿真与训练环境复杂，基于裸机开发不便于环境、任务等 AI 资产的统一管理等。

技术架构：项目以“AI 算力+大模型工具链”为核心，构建分层、解耦、可扩展的具身智能研发与训练体系，支撑多场景的快速迭代与规模化推理。基础设施/算力层：兼容 CUDA 生态的弹性算力池，常用算子与框架直接适配；一机多卡与大显存配置提升训练吞吐；

支持按成员/角色/任务优先级多层次管理与调度；提供多级 Quota 进行更精细化资源分配。平台与治理层(PAI 全栈具身智能工具链)：统一用户角色与权限；多维度任务监控与告警（资源组/任务级别数十种指标，失败电话/短信/邮件告警）；任务编排与公私有镜像管理（丰富官方镜像/自定义镜像），一键部署具身智能热门模型；快速搭建训练/仿真环境，开箱即用的 IsaacLab、IsaacSim。数据与仿真层：面向具身智能场景的仿真数据大规模生成与回放；支持遥操作、动捕与大模型驱动的主流训练方法；统一管理环境、任务等 AI 资产，提升数据闭环效率与验证稳定性。模型层：端侧小模型+云端大模型协同（云端进行复杂推理与多场景泛化，端侧保障实时性与安全性），提升对开放场景下复杂、动态、连续任务处理能力。

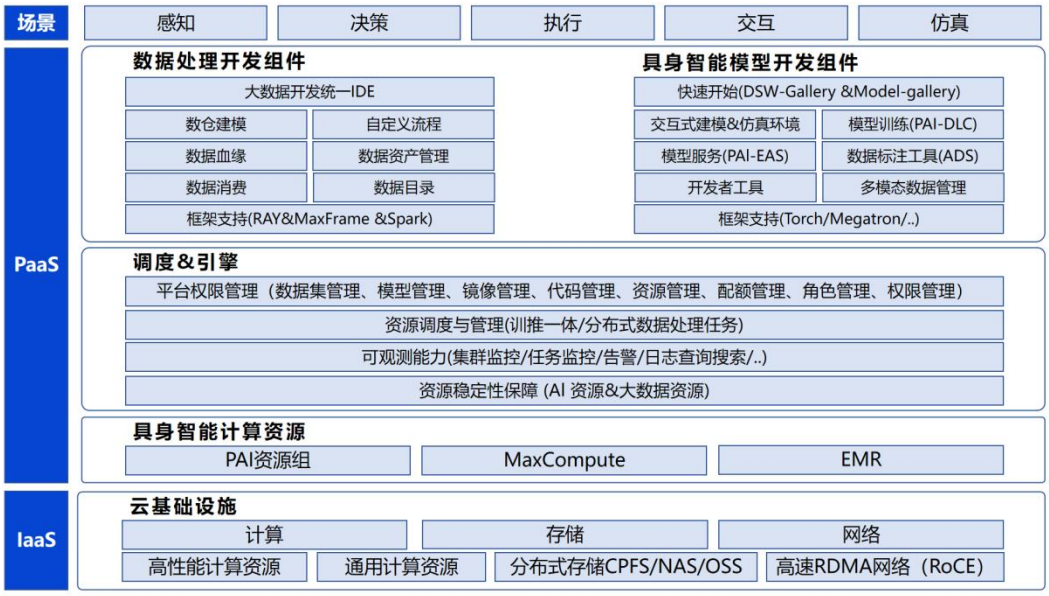

图 20 某具身智能智算平台架构

应用成效：兼容 CUDA 生态，几乎无迁移成本，训练任务迁移用时约 4 小时；常用算子与框架直接跑通；一机多卡+大显存配置带来训练速度提升，多场景模型训练效率显著提高。采用“端侧小模型+

云端大模型”协同，机器人特情问题下降约 60%，复杂多阶段任务推理能力与泛化能力增强。PAI 全栈具身智能工具链实现统一角色/权限/资源治理，多级 Quota 与优先级调度保障关键任务；数十项监控与多渠道告警提升运维稳定性；IsaacLab/IsaacSim 开箱即用，快速搭建训练仿真环境，降低工程难度、提升科研与开发效率。环境、任务等 AI 资产实现统一管理；仿真数据可大范围应用，助力在巡扫、安检、配送等新场景的持续扩展与规模化推理。

(十一)金融行业：某国有大型商业银行数智化建设

案例背景：金融行业的业务线繁复，涵盖了对公信贷、信用卡、个人金融、对公金融、普惠金融等多个业务领域。目前在客户营销、产品创新、业务运营等方面都亟需 AI 技术加速企业数智化转型，提高金融服务的效率性及便捷性。传统 AI 生成模式，单个业务模型应用于单个业务场景，碎片化严重，且模型参数量小，优化训练对业务效果提升有限。因此，亟需通过大模型技术解决传统 AI 技术业务局限性与问题，同时业务模式的高质量发展、底层技术平台及基础设施也需配套提升。

技术架构：某国有大型商业银行通过 AI 原生基础设施能力建设来满足不同阶段和不同需求的 AI 应用开发和部署。其中算力基础平台将计算、网络及存储集成为可统一调度的资源池，服务于大模型训练和推理；通过统一的云平台，进行一站式资源纳管与灵活部署。云底层主要负责 AI 算力管理，完成训练、推理任务的调度、监控，

以及算力资源的管理、调度和回收；AI 平台引入 ModelArts，MindX 等应用使能组件，让 AI 开发更便捷、更高效。同时支持 MindSpore、PyTorch 等多种业界主流 AI 框架。AI 平台同时还具备统一的智算运维平台，通过一键式巡检、故障诊断和实时性能监控工具，降低故障频次，快速故障恢复，全面提升智算资源池的有效利用率。

图 21 某国有大型商业银行 AI 技术体系架构

应用成效：通过 AI 原生基础设施能力建设，该行的金融服务业务收益得到了显著的提升。在远程银行上，通过自然语言大模型及相关 AI 技术的应用，能够自动化生成前序坐席与客户的沟通主旨摘要，防止有效信息丢失。并且可以通过在线跟听，动态预测客户意图，实时分类业务场景，自动进行资料搜索，并归纳总结形成推荐的答复话术提取通话关键词条，提质增效助力客户满意度提升。金融市场中，通过大模型重塑金融市场总分行业务流程，在价格磋商过程中，通过运用大模型智能识别交易话术，生成交易意向单达成交易，并在交易中实时完成客户审查并生成分析报告。

(十二)金融行业：某头部证券公司 AI 原生交易 APP

案例背景：证券行业作为金融领域的核心板块，面临智能化转型的迫切需求：一方面，监管对金融服务的严谨性、实时性与合规性提出更高要求，通用大模型在专业信源整合、复杂决策支持等场景存在短板；另一方面，投资者对“买什么、何时买、怎么买”的精准服务需求激增，传统“内嵌 AI 模块”的局部改造模式已难以满足全场景智能化需求。作为国内 TOP5 券商，为构建 AI 时代核心竞争优势，亟需解决模型效果与性能/成本/安全的平衡、外部信源补充、复杂智能体规模化落地等挑战，探索“AI 原生”券商转型路径。

技术架构：项目以“全场景智能”为设计理念，构建“金融级技术栈+多模型协同+安全合规”的一体化架构火山引擎提供金融级云底座，通过 RDMA 网络与算力调度方案实现高性能训练加速，并支持云下本地集群（保障客户信息安全）与云端弹性资源（支撑公域信息处理）协同。模型层面采用“自研金融大模型+豆包大模型”多模型协同策略，前者负责专业内容生成与判断，后者处理互联网碎片化信息提炼；同时接入中焯行情、Wind 等专业插件，形成“外部信息飞轮”。项目通过部署大模型安全防火墙，对输入输出双向审核拦截违规信息；通过 HiAgent 平台实现智能体全生命周期管理，支持复杂业务逻辑拖拉拽开发。

图 22 某国内头部证券机构 AI 原生交易 APP 技术体系架构

案例成效：项目成功打造行业首个 AI 原生交易 APP，重构证券服务逻辑。用户规模与体验方面，“股市助手”用户超 150 万，位列股票垂类第一；上线后，通过智能选股、盯盘、语音下单等功能，显著提升投资者决策效率与交互体验。建设完成后的统一大模型应用服务平台，已覆盖智能投行、投研、投顾、客服等数十个业务领域，支撑 2000+智能体运行；公有云模型调用量年初至今增长数十倍，成本与性能优势凸显。此项目为证券业提供首个可复用的 AI 原生转型样本，2026 年多家大型券商将基于此模式推出专属 AI 化 APP，推动行业从“问答式”向“执行式”AI 应用升级。

（十三）能源行业：某能源央企海能 MaaS 平台建设

案例背景：当下国家层面将 AI 发展提升至战略高度，同时油气产业高速迭代，正处于业务场景与 AI 深度融合探索关键阶段。集团

内部基于管理效能提升与多元化业务场景落地的双重需求，对构建“一站式人工智能数字底座”提出了明确要求。

技术架构：“海能”平台整体遵循“算力-模型-平台-应用”四层建设逻辑。算力层是平台的“动力中心”，形成了兼顾通用计算与智能计算的强大算力池。模型层是平台的“智慧引擎”，实现“大模型赋能通用能力、小模型攻坚专项需求”的协同效应。平台层是承上启下的“核心枢纽”，智算平台提供“智运”、“智训”、“智管”等模型训推全链路服务，智能体平台快速灵活适配复杂业务场景。应用层是价值落地的“终端窗口”，涵盖问医助手、智能问答、文档写作、智能翻译、智能会议等通用应用，服务高频使用场景。深度服务集团“人工智能+ ”行动。

图 23 “海能”MaaS 平台架构

应用成效：“海能”平台构筑“大模型底座+智能体引擎+行业知识库”三位一体架构，支持中海油研究设计、现场作业、生产运营、贸易销售、科学研究、管理提升 6 大业务域 20+业务场景与 AI 模型的深度融合，打造 AI 原生基础设施交付样板间，推动构建“复用、共享、迭代”的油气行业 AI 应用开发生态。

（十四）能源行业：某央企统一 MaaS 平台建设

案例背景：该企业在前期大模型底座建设中由于缺乏统一的开发环境与组件平台，导致应用开发效率低下、组件重复建设、资源管理粗放，严重制约 AI 能力规模化应用。为实现管网统一大模型服务平台建设，让软件资源（大模型能力）和硬件资源（算力网络等基础设施）与场景应用紧密且高效结合，为应用开发者打造便捷的应用开发框架，同时满足模型管理及各项模型服务统一接口需求，构建统一的 MaaS 平台已成为必然需求。

技术架构：通过提供“智训、智运、智管”大模型训练推理全链路服务能力，整合国家管网分散的、单一场景下的各种数据分析算法与模型，形成统一的、适用于各种场景下的 AI 分析底座，并与现有机理模型、局部场景下数据驱动“小模型”协同融合，构建“一站式”AI 分析平台，全面为智慧管网提供核心模型支撑，持续提升管网各业务产品建设落地能力。

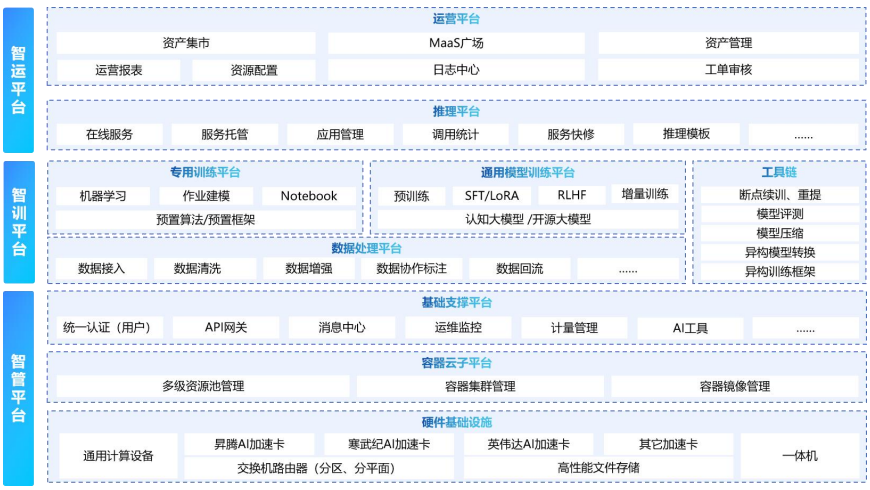

图 24 统一 MaaS 平台架构

应用成效：MaaS 平台为国家管网打造坚实的 AI 技术底座，打通从模型生产到应用落地的关键链路，实现“五统一”能力构建：统一算力资源调度、统一模型训练开发工具、统一云边端模型协同管理、统一大模型与小模型资产管理、统一创新应用生态运营。目前平台已覆盖智能办公、综合管理、作业生产、财经应用等多业务场景，20+模型服务已落地应用，并将持续整合 算力、数据与模型资源，有效推动管网 AI+创新体系建设。

（十五）交通行业：某航空公司 AI 中台建设

案例背景：某航空公司积极响应国家“AI+”专项行动及政策号召，推动智能化转型。当前，航空业面临公文处理流程冗长、人工校对效率低、简历筛选精度不足等痛点，亟需 AI 基础设施优化管理效能。通过引入 AI 技术，该航空公司旨在提升跨部门协同效率，强化人才管理智能化，降低运营成本，同时规范公文流程并保障数据安全。行业智能化趋势与政策红利叠加，促使 AI 成为航空业数字化转型的核心驱动力。

技术架构：本系统构建于 GPU 算力与分布式存储基础设施之上，集成通用与专用大模型形成 AI 能力核心。其关键的 AICT 集成平台，通过统一纳管模型、构建企业知识库与完善用户体系，为 AI 应用提供全链路闭环管理与开箱即用的运营赋能。系统基于微服务架构，以“平台复用+场景定制”模式，快速支撑智能文稿生成、AI 校稿及简历筛选等具体功能，确保系统的灵活性、安全性与可扩展性。

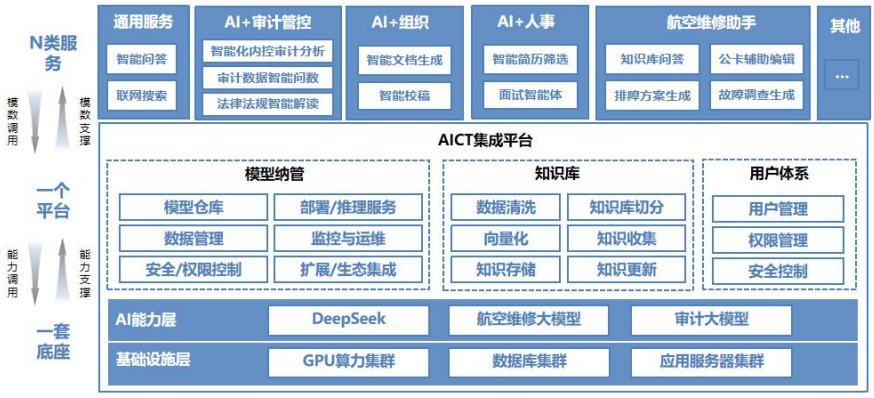

图 25 某航空公司 AI 中台架构

应用成效：通过 AI 中台能力建设，公文处理效率提升 40%，审批周期缩短 30%；简历筛选准确率跃升 50%，招聘周期压缩至两周，人力成本降低 40%；校稿差错率下降 90%，公文质量显著提升。年节约办公成本超百万元，并通过释放 HR 人力至核心业务，优化人才管理质量，增强企业竞争力。

（十六）医疗行业：某三甲医院大模型平台建设

案例背景：在国家《“十四五”国民健康规划》等政策推动下，三甲医院面临提升诊疗效率、优化患者服务、精细化运营的核心需求。传统 IT 架构难以支撑 AI 场景的快速落地，存在算力资源分散、模型训练周期长、应用场景开发碎片化等痛点。随着医疗 AI 技术成熟，医院亟需构建统一的 AI 基础设施，实现从算力调度、模型开发到场景落地的一体化能力，为智慧医疗服务提供底层支撑。

技术架构：项目以“AI 智算算力+大模型工具链”为核心，构建分层解耦的架构：基础设施层整合了高性能计算集群，实现“一云多算”统一调度；联动医院数据中台，提供标准化医疗知识库。AI

中台层部署了“百炼专属版”（含 QWEN-72B 基础模型），通过 SFT（监督微调）、LoRA（低秩适应）、RLHF（基于人类反馈的强化学习）等工具链，支持模型快速训练、推理与迭代。应用层以医疗智能体为载体，优先落地“智能问数”（医院运营数据分析），逐步扩展智能客服、分导诊、临床辅助决策等场景，实现模块化扩展与低代码部署。项目整体设计理念强调“高性能、可扩展”，通过算力虚拟化、模型轻量化技术，保障多场景并发响应与弹性伸缩。

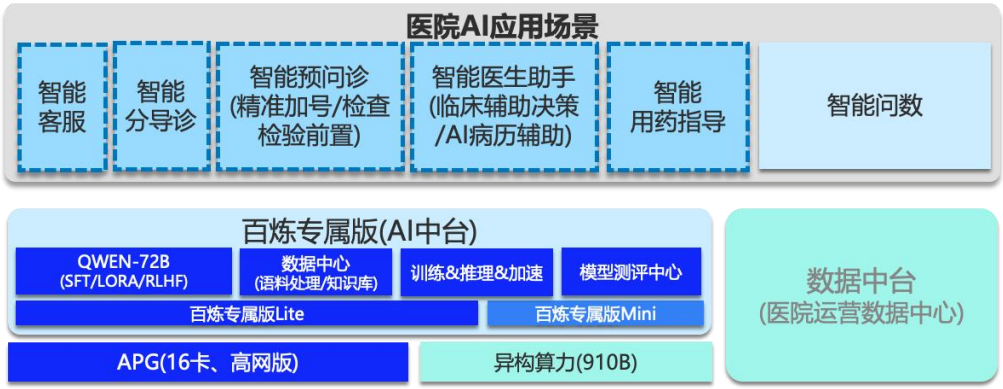

图 26 某三甲医院智算大模型平台架构

应用成效：率先实现“智能问数”场景上线，医院管理者可秒级获取运营数据报表，决策效率提升 60%；同步推动智能客服、分导诊等 5 大场景开发，患者就诊流程平均缩短 20%。构建统一 AI 算力池，资源利用率提升 50%；百炼专属版工具链将医疗模型训练周期从月级压缩至天级，降低开发成本 70%。形成“云+AI 中台+医疗智能体”的可复用模式，为智慧医院建设提供标准化范本，获行业权威期刊案例收录，推动医疗 AI 向“一站式、规模化”落地演进。

五、 总结与展望

当今世界数智产业风起云涌,正以核心引擎之势驱动数字经济的创新变革与高质量增长。AI 技术正在加速向千行百业渗透，不仅重塑产业发展逻辑,更催生并推动着 AI 原生基础设施的规模化建设与落地实践。本文系统性梳理了 AI 原生基础设施的建设框架和关键要点，以期为行业各方的探索与实践提供参考。

展望未来，随着 AI 原生基础设施的全面普及与深度应用，数据智能产业必将迸发更旺盛的创新活力。在国产化软硬件体系持续迭代完善，开闭源模型生态百花齐放，行业高质量数据集建设持续稳步推进的多重利好下，AI 原生基础设施将加速向医政务、制造、金融、交通、能源等传统行业的核心业务流程，为各行业的数智化转型注入强劲动能。其建设与发展,不仅将有力支撑我国“人工智能+”行动的高质量落地，更驱动企业实现深层次业务创新，重塑产业生产关系与竞争格局，助力我国在全球数智化发展浪潮中抢占战略制高点，擘画数字经济与实体经济深度融合的崭新篇章。

联系方式:

中国信通院云计算与大数据研究所

地 址:北京市海淀区花园北路52号

邮 编:100191

邮 箱:wangchaolun@caict.ac.cn