人工智能行业应用建设发展
参考架构

国家信息中心公共技术服务部

二〇二四年十一月

前言

党中央、国务院高度重视人工智能发展，习近平总书记强调，人工智能是新一轮科技革命和产业变革的重要驱动力量，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。贯彻党的二十届三中全会精神，落实2024年政府工作报告关于“深化大数据、人工智能等研发应用，开展‘人工智能+’行动”等工作部署，为充分发挥我国超大规模市场和丰富应用场景优势，以应用为牵引推动人工智能技术与行业深度融合，加快行业应用建设统一架构设计，集中力量打通从模型到应用的落地堵点，降低应用开发验证门槛，提高部署效率，加快形成人工智能规模化应用正成为市场关注的重点。

国家信息中心在深度调研人工智能产业发展现状基础上，聚焦人工智能行业应用发展关键问题，面向制造、交通、金融、医疗、消费等人工智能应用重点行业领域，研究提出《人工智能行业应用建设发展参考架构》报告（以下简称“报告”）。报告从算力基础、数据服务、模型服务、应用开发、运维平台、运营平台等六个方面，提出人工智能行业应用建设的共性能力和特性能力。通过构建一套技术架构统一、数据规范统一、标准体系统一的参考架构，摆脱企业服务模式不同带来的限制，降低供需边际成本，有效发挥规模效应，促进应用创新，激发市场活力，持续推动产业健康高效发展。

本报告旨在研究推进行业应用发展标准化的参考架构，期望为各行业主体明确人工智能应用建设发展的重点和目标，降低应用开发和复制的边际成本，促进人工智能技术的创新成果与产业深度融合，为加快推进人工智能行业应用规模化落地提供有益参考。

主要编写人员：

徐春学、张立峰、宦茂盛、刘飞飞、张润宇、唐露、赵昊楠、张楚妍、史宇萌、张驰

目录

一、人工智能行业应用总体进展

（一）国内外人工智能行业应用发展现状

1.全球人工智能技术演进日趋激烈

当前，全球人工智能产业正迎来蓬勃发展的黄金时期。基础技术不断实现突破，产业生态日益成熟，行业应用范围不断拓宽。与此同时，各国政府纷纷出台相关政策，规范行业发展，试图在全球竞争中抢占发展先机。

国外的模型能力、多模态技术、混合模型等基础技术持续演进。一是当前大模型能力快速提升。Open AI 最新发布的 o1 大模型采用思维链方式对复杂问题进行逐步分析，使得解决多层次的数学、科学和编码问题成为可能，该模型成为第一个与人类专家能力相当的模型。多模态人工智能技术已经能够综合处理文本、图像、音视频等多种类型的数据，提供更丰富和复杂的服务，例如多模态具身视觉语言模型 PaLM-E、文生图模型 DALL-E、多模态经典模型 CLIP 等。注意力机制与其他机制结合的模型能够在降低计算成本和内存占用的同时，保持甚至提高准确性。例如 Google DeepMind 的 Griffin 模型结合了线性递归和局部注意力，大大减少了训练时使用的标记数量，依然能保持与 Llama-2 相当的性能。二是基础技术体系加速收敛。经过十多年的发展沉淀和市场选择，美国 AI 产业逐渐形成统一收敛的局面，展现出强大竞争力。在算力（英伟达 GPU 占比 85%以上）、人工智能深度学习框架（PyTorch 占比 90%以上）、模型（GPT、Llama

系列）等层面，基于收敛的技术栈，吸纳了全球开发者的贡献，形成了强大的生态体系。

国外人工智能产业从芯片、并行计算引擎、人工智能深度学习框架、工具链、开源社区等方面形成了完整的产业链。芯片上，英伟达 GPU、谷歌张量处理单元（TPU）等为大规模并行计算提供支持，英特尔、AMD 等则布局多种人工智能专用芯片，满足数据中心、智能终端、自动驾驶等场景的需求。并行计算引擎上，以英伟达 CUDA 为代表，凭借其丰富的算子生态、强大的社区支持，成为主导技术。同时，CUDA EULA 限制条款要求不允许逆向工程，不允许在非英伟达硬件平台上进行转译运行，进一步巩固了其产业地位。人工智能深度学习框架上，形成了以 Meta 主导的 PyTorch 和 Google 主导的 TensorFlow 为代表的聚集效应。其中，PyTorch 是目前全球最流行的人工智能深度学习框架，并结合英伟达 GPU 芯片底层进行计算优化，占据了 90%以上的份额。开源社区上，国外开源社区聚集效应显著，以全球最大的开源 AI 社区 Hugging Face 为例，吸引了大量的开发者和企业参与，截止 2024 年 11 月平台收录超过 100 万个大小模型和超过 20 万个数据集，包括微软、谷歌在内的超过 15 万家机构使用。

国外人工智能技术应用在医疗医药、自动驾驶、高端制造等行业，并持续拓展应用场景。医疗医药行业，在疾病诊断、药物研发、基因组研究、智能健康监测等方面广泛应用。DeepMind 的AlphaFold系列模型专注于蛋白质结构预测，获得2024年诺贝

尔化学奖，这也是首次将该奖项颁给人工智能相关的研究。**自动驾驶行业**，英伟达的DRIVE平台可进行复杂的路径规划和实时决策，确保自动驾驶汽车的安全行驶。特斯拉FSD系统实现从感知到控制端到端自动驾驶技术在量产车型上的应用已成为现实。**高端制造行业**，在工业设计、需求预测、过程优化、供应链优化等领域广泛应用。DeepMind的AlphaChip利用了强化学习方法来设计芯片布局，能够在数小时内生成人工需要数周甚至数月的芯片设计工作，并应用于谷歌TPU芯片设计中。

2.我国具有独特的发展资源优势

习近平总书记强调：“中国具有社会主义市场经济的体制优势、超大规模市场的需求优势、产业体系配套完整的供给优势、大量高素质劳动者和企业家的人才优势，经济发展具备强劲的内生动力、韧性、潜力。”**中国的市场规模体现在消费者基数大和企业数量多。**中国拥有 14 亿多人口，中等收入群体超过 4 亿人，连续多年保持世界第二大商品消费市场、世界第一制造业大国、第一货物贸易大国地位，这为各种新技术、新产品提供了广阔的应用场景和市场空间。

中国在全球 AI 版图中的核心地位日益凸显，其增长潜力不容小觑。根据 2024 年《全球人工智能和生成式人工智能支出指南》，中国 AI 市场规模已由 2018 年的 84 亿美元增至 2022 年的 319 亿美元，预计于 2027 年将达到 1,150 亿美元，2022 年至 2027 年的复合年增长率为 29.2%。中国 AI 市场规模占全球 AI 市场规

模的百分比由2018年的11.9%上升至2022年的16.1%，预计于2027年将达到20.6%。

我国人工智能应用层企业完备，且具备良好的发展势头。根据前瞻产业研究院《2024年中国人工智能行业全景图谱》，在2200家人工智能骨干企业中，提供基础硬件设备和数据服务的企业仅有53家，从事包括核心算法、开发平台等在内的关键技术的研发的技术层企业有273家，涉及人工智能技术的集成和场景应用的应用层企业占比高达85.18%，达到1873家。

人工智能技术在互联网领域中的应用尤为活跃，并且正积极应用于实体经济产业中。一是各行业人工智能应用分多个波次发展。互联网头部企业作为智能化先锋，主导AI技术的研发和应用。运营商、金融、汽车、手机和政府服务类处于第二梯队，正积极训练、部署大模型，将AI技术应用于核心场景。电力、医疗、制造、交通和零售等行业处于数字化阶段，积极探索AI应用的落地。而餐饮、建筑和农牧业等仍处于起步阶段。二是人工智能技术正在加速与传统实体产业的融合，特别是在制造业、医疗、交通等领域。根据《中国互联网发展报告2024》，中国已建成近万家数字化车间和智能工厂，其中90%以上的示范工厂应用了人工智能和数字孪生技术。

3.行业应用正成为人工智能竞争的焦点

技术与产业竞争的焦点已经从单一的计算量和模型参数量转变为对高质量数据集规模的重视。基于规模理论（Scaling Law）

的发现，在人工智能领域，模型性能的关键影响因素包括计算量、模型参数量和数据集规模。其中，计算量取决于算力水平，模型参数量决定模型大小，而数据集规模作为人工智能系统的原材料，其规模和质量直接影响模型训练的效果和应用的广泛性。国内外人工智能技术的应用正在不断落地，尤其是在数据密集型行业如金融、医疗健康等领域。这是因为高质量、多样化的数据对于训练有效的模型至关重要，能够显著提升人工智能应用落地效果。

目前国内也呈现出在行业赋能效果提升上发力的趋势。**一是从模型开发到更关注高价值场景的开发。**模型开发和优化需要投入大量资源，随着基础大模型能力的显著跃升，企业的工作重点正在从模型开发和调优转移到高价值场景的应用开发上。企业可利用已有的基础模型、行业模型、小模型和机理模型，专注行业场景的应用需求。**二是从关注单点技术指标到考虑整体成熟度。**人工智能技术体系垂直整合度高，企业进行技术选择时，正从聚焦个别技术指标评估转向软硬协同整体系统的成熟度和适用性的评估。大规模集群能力、大参数模型训练能力以及算网存一体调优能力等，成为评价技术成熟度的关键。采用主流成熟技术可以保护企业投资，确保技术体系能长期满足需求。**三是从追求短期技术优势到重视长期可持续。**在当前国际产业竞争大环境条件下，确保产业链和供应链的可持续性至关重要，企业追求短期技术优势，长期看可能影响系统的持续演进和升级能力。

（二）人工智能行业应用建设发展模式

当前，人工智能行业应用的建设通常存在以下两种典型模式，这些模式基于业务建设目的的不同而区分。建设的行业应用以服务企业自身为目标的自研创新模式以及以服务行业内其他企业为目标的平台赋能模式。

1.自研创新模式

自研创新模式重点关注内部业务应用需求，需要解决与现有业务系统融合和功能复杂定制的问题。其特点如下：

一是聚焦内部业务需求。以满足自身业务需求为主要目标，通过分析自身业务流程中的痛点和机会，利用人工智能技术进行针对性的优化和创新。**二是与现有业务系统融合度高**。人工智能业务应用需深度融入企业现有的业务系统，并服务于企业自身关键业务。**三是功能定制复杂程度高**。企业对系统功能定制复杂度要求高，以更好地满足其独特的业务需求和工作流程。

2.平台赋能模式

平台赋能模式则需要关注市场拓展、业务运营和商业是否成功，以保障业务可持续发展。**一是关注市场拓展与商业可持续**。将人工智能相关的资源或应用作为一种服务提供给外部客户，例如提供行业数据集和行业模型，为行业客户提供数据分析和预测服务等。**二是关注技术创新与合作**。通过与外部伙伴的合作，不断改进算法和模型，并形成有竞争力的生态系统。**三是关注品牌建设与行业影响力**。树立行业领导者形象，参与和主导在行业协

会组织的技术研讨和标准制定工作，推动人工智能技术在行业中的规范应用和发展。

3.两种模式的关系

平台赋能模式一般由企业自研创新模式进一步拓展而来。不同的业务发展模式并不相互割裂，很多企业从自研创新模式开始，通过进一步建立面向外部服务的运营实体，并叠加新的系统能力，实现内部成熟能力的自建他用，从而创造新的业务和市场空间。

（三）推进统一参考架构设计是发展关键

1.行业应用发展受到多方面挑战

当前人工智能行业应用产业的发展，仍然受到一些关键要素的制约，导致模型和数据孤岛、应用开发和复制边际成本高、自主可控不足等情况，不利于发挥我国大市场优势，聚合形成产业优势，包括技术、成本和安全三方面。

（1）技术方面，不同厂商和平台之间无法实现互联互通。

在人工智能行业应用的发展过程中，技术体系的碎片化成为主要挑战。各大厂商的算力、开发框架、模型到应用等方面形成独立技术体系，一是技术不兼容，不同厂商和机构开发的AI技术难以互相兼容和融合，导致技术孤岛现象，限制了技术的广泛应用和集成，使得不同系统和平台之间的协同工作变得困难。二是数据流通不畅，由于缺乏统一的数据标准和格式，数据在不同系统间的流通受到阻碍。三是缺乏互操性，模型和算法的互操作性差，不同AI模型的基础原理各不相同，缺乏互操作性，限制

了不同模型之间的协同发展。**四是技术发展不均衡，**不同领域的AI技术发展速度和质量参差不齐。**五是技术扩散受阻，**不同子系统之间的行业标准不一致，发展速度不协调，对技术的扩散造成负面影响，统一架构的缺失限制了技术突破后的快速扩散和应用。

（2）成本方面，各公司自成体系导致行业整体研发、运营和推广成本高。

一是研发运营成本高。各个公司需要独立开发和维护自己的技术体系，导致研发和运营成本增加。这种成本的增加可能会阻碍小型企业进入市场，减少竞争和创新。**二是应用推广复制成本高。**由于人工智能行业应用当前缺乏统一的标准，导致人工智能行业应用的开发和推广仍需要大量的适配和定制，使用和推广模式也尚未形成固定机制，使得创新成果从开发完成到落地推广复制成本仍比较高，不利于成果转化。

（3）安全方面，技术体系不兼容不利于监管和业务连续性。

监管方面，缺乏统一架构使得制定全面、系统性的法律体系变得困难，无法全面覆盖人工智能快速发展带来的数据隐私侵犯、算法歧视、知识产权争议等复杂法律问题。**业务连续性方面，**安全是人工智能行业应用的重要保障，部分国外技术存在“断供”风险，一旦国外技术受限，构建在国外生态之上的人工智能应用将无法持续更新，甚至无法使用。外部依赖导致技术开发和应用能力供给韧性不足，对相关应用安全和业务连续性造成挑战。

2.统一架构有利于加速应用规模化落地

企业在采用统一架构的基础上，能够自由遴选并整合多家顶尖的大模型及技术供应商，确保在业务发展过程中，能够灵活地在现有平台上引入新的合作伙伴。这种集成不同厂商 AI 应用的能力，不仅降低了技术门槛，还有效缩减了成本，并优化了资源配置。最终，这些措施共同推动了行业内的创新合作，为企业的长期发展注入了活力。具体表现在，企业可以更加灵活地选择不同的大模型厂商，根据业务需求和成本效益分析，选择最适合的 AI 解决方案；当业务需求发生变化时，企业可以快速地在技术上实现新增其他厂商的应用，以适应市场和业务的快速变化，而无需担心技术兼容性问题；能够更快地部署新的 AI 应用和服务，加速产品上市时间，提高对市场变化的响应速度；能够轻松集成来自不同厂商的 AI 应用，实现跨平台的数据和功能整合，提升业务流程的连贯性和效率；减少了因技术不兼容导致的重复开发和维护成本；使得企业能够更好地管理和分配技术资源，集中精力在业务创新和价值创造上，而不是技术整合和兼容性问题上。此外，统一架构鼓励不同厂商之间的合作，共享数据和资源，促进行业内的创新和竞争。对于新进入市场的企业来说，统一架构降低了技术门槛，使得他们能够更快地融入市场，利用现有的技术和资源。

二、人工智能行业应用建设发展统一参考架构

（一）统一参考架构的内涵与特性

构建统一的人工智能参考架构对于推进行业应用至关重要。通过标准化的架构，可有效降低不同系统间集成的难度和边际成本，促进算力、数据、模型和应用的协同。此外，统一架构有助于推动产业链的发展，为产学研合作解决行业共性问题提供同一套“施工图”，实现技术互通、市场整合和行业人工智能的长远发展。

1.统一参考架构的内涵

人工智能行业应用参考架构主要包括两个方面：系统架构和能力模块。

（1）系统架构

人工智能行业应用系统架构，指从下至上的系统层次“框架”。即为人工智能系统设计通用统一架构，它包括了一系列基础组件和依赖关系，以支持不同的人工智能应用和服务。旨在指导建设统一平台，使得不同的AI应用能够更容易地集成和交互。

（2）能力模块

人工智能行业应用能力模块，指填充框架的“内容物”。即能力模块，能力模块可跨不同应用和领域重复使用的标准化组件或服务，通常围绕模型服务和数据治理两大关键要素，具备数据服务、模型服务、运维管理、运营管理等能力，形成统一标准，向上为应用开发提供统一服务，向下适配算力基础。其中，能力模块分为“共性能力”和“特性能力”。

2.统一参考架构的特性

行业人工智能参考架构的特性，体现在其普适性和可演进等方面，确保人工智能技术易于应用和推广。**普适性，**体现在参考架构的统一框架能够适用于多个行业和领域，不受特定应用场景的限制。**可演进，**参考架构统一框架随着技术的发展和业务需求的变化进行扩展和适应，且易于维护升级。

（二）统一参考架构组成

1.总体架构

为寻求基于当前产业发展实际的“最大公约数”，围绕系统架构，梳理模型和数据驱动的共性技术基础能力建设，总体提出共性支撑能力技术要求，实现人工智能产业发展和行业智能升级协同。如图1所示，按照人工智能应用的通用技术架构，**统一参考架构分为算力基础、数据服务、模型服务、应用开发、运维平台、运营平台等6个主要部分**，其中，算力基础包含3部分组成，数据服务包含4个共性能力和4个特性能力，模型服务包含4个共性能力，应用开发包含2个特性能力，运营和运维平台均为特性能力。

图 1 人工智能行业应用统一参考架构

参考架构共性能力以模型和数据两大关键要素提出共性支撑能力技术要求，特性能力以应用开发、数据处理、运营运维等依据业务的不同各具特点的要素提出。具体的，**共性能力，**是以数据和模型为核心的共性能力，是各种不同人工智能行业应用系统所需要具备的通用功能，且需符合一定的一致性技术要求。**特性能力，**指应用开发、运维平台和运营平台，以及数据服务中的数据接入、数据模型、数据存储和数据分析。特性能力是参考架构的组成部分，但对其实现方式不作统一要求，应根据建设实际情况适配实现方式。

其中要重点指出的是，**共性能力，不对实现做具体约束；特性能力里不对功能做具体约束。**即共性能力中的功能建议以任意开发形式建设，并满足本报告中的技术要求；特性能力中的功能设计仅作示例，具体以实际业务需求为准，建成系统中存在该能

力的相关功能即可。

2.算力基础

算力基础包含三方面的组成，基础设施、算力资源管理平台和AI开发平台。

一是算力基础设施。包括了AI芯片、网络、存储、并行计算引擎等，作为数字基础设施的核心，为企业提供数据存储、计算和应用服务，是企业数字化转型的基石。

二是算力资源管理平台。用于监控、调度和管理算力资源的系统。算力资源管理平台是企业数字化转型中的关键工具，帮助企业优化资源配置，提高运营效率。

三是AI开发平台。包括AI深度学习框架、软件仓库、算法库、模型开发、训练微调、部署等功能。其提供了一套完整的工具和框架，使得开发人员可以更加便捷地进行AI应用的开发。

3.数据服务

数据服务能力包含4个共性和4个特性能力，数据服务需重点关注采集、处理、共享、数据集管理，形成统一的数据服务。

其中，4个共性能力分别是，①数据工程工具链：应关注训练数据质量，具备数据增强、数据评估、数据合成、数据清洗等能力；② 数据采集模块：由于质量和可用性是作为模型输入，使用模型进行推理和应用的关键，具备通信网关、数据标准规范、智能终端操作系统、数据安全隔离等能力；③行业数据空间：由于数据可信交换是基地对外提供公共服务的基础能力，应具备可

信交换、融合共享、安全策略、空间管理等能力；④数据集服务：由于聚合和发展行业内高质量数据是不断迭代大模型的基础，具备数据集发布管理、加密封装、权限认证、数据集规范管理等能力。

4个特性能力分别是，①数据接入，可以通过多种方式实现，包括开放API接口、数据导入、数据源接入和数据埋点等方式；②数据模型，任意数据模型均可实现，网状数据模型和层次数据模型等；③数据存储，支持各类数据库的存储；④数据分析，支持各领域数据分析模型和展现形式。

4. 模型服务

模型服务包含4个共性能力，模型方面需重点关注基础模型、模型训练、模型使用、资产沉淀，形成统一的模型服务。①模型组合：应关注模型的性能、可扩展性和适应性，因此具备几个主流国产基础大模型的组合能力，来发展和训练行业大模型；②模型工程工具链：应关注工具链功能完整性、开放性和易用性，具备训练引擎、压缩引擎、对齐引擎、治理引擎等几个能力；③Agent 工程工具链：应关注便捷化、智能化的编排规划对应用“最后一公里”的支撑，具备编排引擎、规划引擎、工具引擎、优化引擎等几个能力；④AI 资产管理：应重点关注 AI 资产的可沉淀、可交换和可复用，具备对模型、Agent 资产做好标准化管理和安全访问控制的能力。

5. 应用开发

应用开发包含 2 个特性能力。①行业应用引擎，支撑各行业各领域业务的引擎，包括搜索检索、GIS 等根据具体业务应用需要；②应用场景，围绕高价值应用场景图谱选择业务需要的应用开发功能。

6.运维平台

运维平台（特性能力），是指面向人工智能应用系统需求，确保系统的稳定和安全。

7.运营平台

运营平台（特性能力），是指面向服务用户提供用户管理、用户服务等相应功能。基于不同的业务目标，侧重点不同，具体有服务内部业务和服务外部客户两类。

三、统一参考架构的建设

（一）统一参考架构技术要求

通过加强共性支撑能力建设，解决行业共性问题，培育开放包容可持续激发各方合作活力的共赢生态，支撑行业人工智能高质量发展。

1.算力基础

构建统一算力基础，包含算网存基础设施、算力资源管理平台、AI开发平台。应构建统一算力基础，避免产业发展受阻，应构建支持万卡算力集群调度演进能力的软硬件基础设施，统一多厂商大模型使能接口，支持模型厂商大模型开发和训练推理。同时实现算网存高效协同，面向文本、图像、音频和视频等多模态

数据，提供高可用的对象存储、分布式数据库、文件系统、向量数据库等存储基础设施，提供统一的访问和管理接口，提升数据处理的效率和准确性，实现各类数据资源存储、管理、访问的技术规范统一，具体应实现以下9点技术能力。

一是算力可靠高效。包括算力集群规模及可用度、支持算网协同通信优化、支持训推共池及算力切分。

二是训推架构统一。支持训推同架构、预置主流的分布式训练框架和分布式并行推理。

三是具备算网存一体化监控功能。包括统一的运维管理工具和跨域故障快速修复。

四是提供存储基础设施。提供高可用的对象存储、分布式数据库、文件系统、向量数据库等，以支持大规模数据的存储和访问。提供高效的数据检索和访问机制，包括支持多种查询语言和API。

五是具备数据资源管理功能。提供数据资源管理功能，支持团队资源管理模式，包括成员权限控制、任务分配和进度跟踪。实时监控数据服务的性能，并提供详细的报告和分析。

六是具备安全管理功能。实现工作空间之间的信息隔离，以保护数据的独立性和安全性。实施高级安全措施，包括数据加密、访问控制和审计日志，以保护数据不被未授权访问。

七是具备全栈技术能力。提供芯片、并行计算引擎、算子开发工具、人工智能深度学习框架等，满足具备完整知识产权或立足于国内主导的开源社区的条件，并预置主流算子加速库。

八是统一数据访问接口。设计统一的访问和管理接口，使得不同模态的数据能够以一致的方式被访问和操作。确保数据在存储、处理和传输过程中的一致性，包括数据的完整性和准确性。

九是统一模型服务接口。为不同模型提供规范化的模型运行时接口、模型加载与执行接口、算子编译和执行接口、性能 Profiling 接口、故障运维管理接口等，支撑多种基础模型的高效稳定运行。提供规范化的模型权重和镜像封装接口，支撑基础模型的便捷部署。

2.数据服务

（1）数据工程工具

数据工程工具包括数据评估、数据合成、数据清洗、数据存储、数据接入。针对大模型的训练数据准备，提供面向开发者易用、低门槛的开发环境，提供数据增强、数据评估、数据合成、数据清洗等能力，提升数据开发的效率。自动或半自动地创建新的特征，对于图像、音频等媒体数据，使用旋转、翻转、缩放或添加噪声等方式生成额外的训练样本以提升模型的泛化能力，提高数据的质量和丰富性。

（2）数据采集管理

统一数据采集传输标准，从源头治理数据质量。建设单位应统一行业智能终端的操作系统、IOT 及终端设备的传输协议、数据格式及数据隔离安全隐私保障机制，从而降低采集侧数据治理的成本。并且应包含端侧要求、边侧要求、传输要求、安全要求。

（3）数据集管理

一是应构建高质量的行业数据集，向模型训练提供高质量、不断迭代的数据集服务，包括标准化的通用数据集和特定行业的数据集，支持数据集的加密封装和发布，并与模型开发工具链集成，对生态伙伴开放，以降低模型训练和评测门槛，从而实现模型的快速开发和持续迭代。**二是应包含数据集发布、安全流通、数据追溯功能**，建设统一的数据服务市场，提供服务权限认证、记录完整的服务订阅关系与调用日志。**三是数据服务集成**，实现数据集与模型开发工具链的自动化集成能力。

（4）数据交换空间

提供数据交换空间，保证数据开放共享。为构建互信、互利的生态系统，充分发挥行业数据价值，需提供空间管理、可信交换、融合共享、安全策略等能力，确保数据要素在企业间、企业内不同主体间跨边界可控流转，实现数据价值的最大化。建设多种类型的数据主权控制能力，例如限定访问次数、使用时长、传输方式、限定用户群等，确保多方数据交换可信。面向模型平台，支持常用数据格式的对接能力，包括且不限于JSONL、XML等常用AI数据格式对接。具备统一的数据清算中心，集中管理多方

身份认证以及数据交换记录。具体要求包括构造多方信任的生态环境、打造可组合的使用控制策略和安全审计追溯。

（5）企业数据空间

数据服务应有效支撑面向企业应用的数据管理。一是满足多型态的数据服务和管理，确保数据的一致性、可访问性和质量，同时优化数据的利用，满足场景对各型态数据服务的支持。二是需支持多模态数据集成融合、数据分析和处理、数据质量控制、数据安全和隐私保护、实施高级安全措施、团队协作和工作空间管理、数据共享和隔离。

（6）4个特性能力

特性能力包含数据接入、数据模型、数据存储和数据分析等，须根据项目实际实现个性化需求。其中数据接入提供文件、数据库、API 等多种数据接入方式；数据模型提供物模型、整合层管理等能力；数据存储提供多模态数据的存储访问管理；数据分析提供数据可视化、自助分析等能力。

3.模型服务

完善模型工具链支持体系，提升一体化开发效能。在模型训练、微调、推理部署过程中，为向开发者提供易用、低门槛的开发环境，项目应当提供统一接口、功能完整、服务化的工具链体系，实现训练引擎、压缩引擎、对齐引擎、治理引擎，支持多模态训练数据准备、标注、数据集配比等能力，降低工具间的适配成本，从而提升开发效率。

（1）多模型组合

引入多种主流基础大模型和小模型。支持行业模型快速开发和组合，以满足不同场景下行业模型的快速开发与应用。多模型能力体现在模型多样、模型模态多样、行业大模型可随基础大模型能力提升而不断进化、具备基础大模型的组合能力。

（2）模型工程工具

完善模型工具链支持体系，提升一体化开发效能。在模型训练、微调、推理部署过程中，为向开发者提供易用、低门槛的开发环境，应当提供统一接口、功能完整、服务化的工具链体系，实现训练引擎、压缩引擎、对齐引擎、治理引擎，支持多模态训练数据准备、标注、数据集配比等能力，降低工具间的适配成本，从而提升开发效率。

（3）Agent工程工具

具备自主决策、多模态数据处理、记忆机制、推理与规划、外部能力集成、多代理协作能力。支持共享的行业任务规划引擎、规划引擎、工具引擎、优化引擎等，降低行业智能体的开发难度，加快开发周期，快速将大模型能力传导至应用侧。

（4）AI资产管理

包括模型管理、组件管理、应用和 Agent 管理、空间管理、监控和报告等能力。全面维护关键人工智能资产，包括模型、Agent 资产等，统一资产类型、格式，支持统一资产管理和访问，

实现人工智能资产的可复用性，支持相关人工智能资产的迭代演进。

（5）统一多模型推理服务

建设统一推理服务标准，确保大模型快速服务行业应用。应包含推理部署、推理监控、支持提供在线推理服务和批量推理服务，满足各种时延要求和成本要求的推理场景、镜像文件和模型文件加载接口等。为了加速推理服务的快速部署和行业复制，应加强大模型推理服务的标准化，支持多厂商模型组合和兼容，统一模型文件格式和性能标准，并将模型服务封装为服务化 API，从而实现模型的高效加载、管理和调度，实现模型的快速部署应用。推理服务应统一 Restful 访问接口，支持多种推理框架，同时支持多种模型和 Agent 服务。

4.应用开发

包括应用工具链和数据反馈平台。一是提供行业应用引擎和行业应用组件等应用工具链，并将相关能力封装成标准化应用和解决方案，支撑高价值、高负载的业务场景高效开发、训练和推理。二是经用户授权的脱敏样本数据可反馈回平台，优化数据集和样本，并形成数据飞轮，推动模型持续迭代。

5.运维平台

提供系统管理维护、高可用、灾备、监控优化的算网存统一运维能力，集应用、数据、模型、算力等自上而下的安全保障能力，满足安全可信等级保护要求，保障平台平稳运行。

6.运营平台

面向企业内外部用户，构建一个应用产品共享平台，支持智能应用订阅、数字资产管理等服务，提供统一多租户和权限管理。为用户提供独立的专属空间和权限管理，实现数据、模型、应用、作业流程的自主管理和运营。用户可从平台获取资源、模型服务或标准化应用，并在用户侧部署和运行。建立端边云协同推理机制，解决边端算力不足问题。建立全流程用户数据保护机制，保障数字资产的可信交换。

（二）基于统一参考架构的应用建设

1.各服务模式下的建设能力

不同类型的人工智能应用都可使用统一参考架构进行建设，包括算力基础、数据服务、模型服务等，同时在应用开发、运营平台和运维平台的功能建设方面，存在一定的差异和重点，如下图所示：

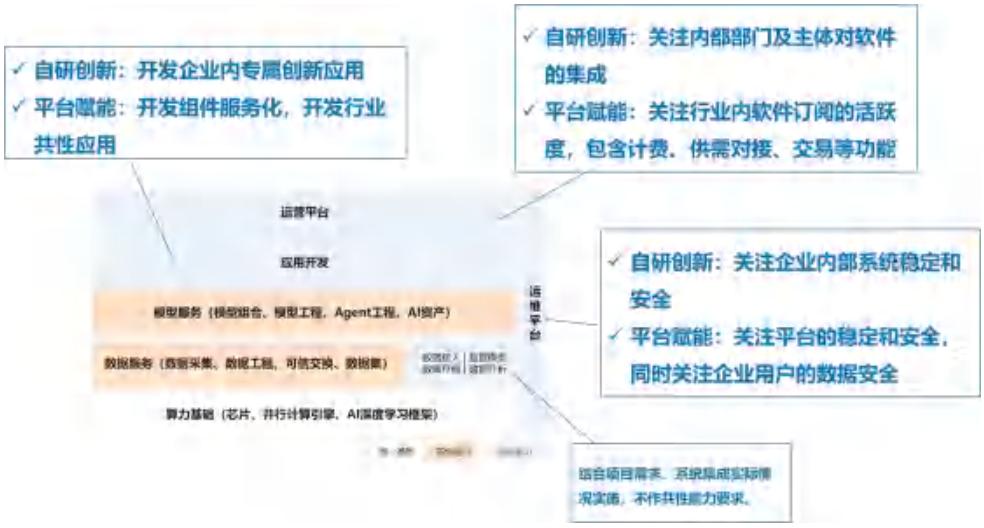

图 2 自研创新模式和平台赋能模式建设能力差异

在应用开发方面，自研创新模式关注开发企业内专属应用，而平台赋能模式则要求开发组件服务化，主要用于开发行业共性应用。在运营平台方面，自研创新模式主要关注内部部门及主体对软件的集成，而平台赋能模式关注行业内软件订阅的活跃度，包含计费、供需对接、交易等功能。在运维平台功能上，自研创新模式关注企业内部系统稳定和安全，而平台赋能模式关注平台的稳定和安全，同时关注企业用户的数据安全。在数据接入、数据模型、数据存储、数据分析等特性数据服务模块，结合项目需求、系统集成实际情况实施。

2.自研创新模式下的技术架构

自研创新模式的参考架构如下图所示：

图 3自研创新模式下的架构

对照统一参考架构，在自研创新模式下，各建设单位除共性

能力建设一样之外，运营平台、应用开发、运维平台和数据交换空间四个方面的特性主要表现在：①运营平台更关注企业内部应用部署、订阅；②应用开发更偏向行业应用引擎引入，满足特定场景需求；③运维平台重点关注企业内部稳定和安全；④数据接入、数据模型、数据存储、数据分析等模块重点在于结合企业现有系统、数据和数据应用的实际情况实施。

3.平台赋能模式下的技术架构

平台赋能模式的统一参考架构如下图所示：

图 4平台赋能模式下的架构

对照统一参考架构，在平台赋能模式下，各建设单位除共性能力建设一样之外，运营平台、应用开发、运维平台和数据交换空间四个方面的特性主要表现在：①运营平台更关注市场管理、合作管理、资产、运营、营销等；②应用开发更偏向开发组件服务化，开发行业共性应用；③运维平台重点关注平台的稳定和安

全，同时关注企业用户的数据安全；④数据接入、数据模型、数据存储、数据分析等模块应结合具体行业的实际情况，建设相关数据标准，并统一进行标准化建设。

四、总结与展望

人工智能应用建设参考架构将在金融、医疗、制造、交通、教育、互联网和智能家居等多个“平台多、系统复杂、业务变化快、成本把控严格、追求效率”的领域展现出巨大潜力，通过优化业务流程、提高效率和降低成本，推动这些行业的创新和变革。未来，人工智能统一参考架构的应用将加速以下方面的发展。

一是激发统一市场活力，有效发挥规模效应。通过加强共性技术的支撑，推动产业聚集，形成人工智能统一大市场，充分利用我国庞大的市场和丰富的数据资源，激发市场的活力，发挥规模效应，最终带动产业整体发展。

二是降低供需边际成本，激发创新应用活力。通过整合底层资源，提供统一的服务，规范开发范式，降低开发和创新的成本，充分激发AI的创新活力，培育开放、包容、活跃的共赢生态。同时，通过统一共性架构标准，拓展推广渠道，降低应用推广的成本，加速AI技术的应用落地。

三是立足可持续技术框架，促进产业持续发展。推动人工智能芯片、深度学习框架、基础大模型等方面技术的应用，打造基于可持续软硬件环境的有竞争力的产品和技术方案，持续培育行业生态，打造安全可靠、竞争力强的现代化产业体系，确保业务

安全和产业可持续发展。

人工智能行业应用的建设与发展是一项长期而艰巨的任务，需要各方共同努力，我们期待与社会各界积极推动统一参考架构在各领域的广泛采用，降低行业训练模型的成本和应用门槛，加速人工智能在行业中的规模化落地，助力高质量发展！