

面向智能体能力构建的 高质量数据集建设指南 (2026 年)

中国信息通信研究院人工智能研究所

联想（北京）有限公司

中国人工智能产业发展联盟数据委员会

2026年10月

版权声明

本报告版权属于中国信息通信研究院、联想（北京）有限公司、中国人工智能产业发展联盟，并受法律保护。转载、摘编或利用其他方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院、联想（北京）有限公司、中国人工智能产业发展联盟”。违反上述声明者，编者将追究其相关法律责任。

前 言

当前，人工智能正处于从“模型能力驱动”向“应用能力驱动”加速跃迁的关键阶段。以大模型为核心的新一代技术体系不断突破认知边界，推动人工智能由感知智能向认知智能与决策智能演进。在此背景下，智能体（AI Agent）作为连接基础模型能力与复杂应用场景的重要载体，正逐步成为推动人工智能规模化落地与产业深度融合的核心形态。围绕智能体构建的数据、算法与系统能力体系，正在重塑数字经济的发展范式，并对数据资源的组织方式、供给质量与治理机制提出了全新要求。

数据作为智能体能力形成的基础要素，其重要性已从“支撑模型训练的原材料”上升为“决定智能体性能边界与应用深度的关键资产”。与传统以静态样本为主的数据集不同，面向智能体能力构建的数据体系呈现出多模态融合、过程导向、场景驱动与持续演进等显著特征。高质量数据不仅需要具备准确性、完整性与一致性等基础属性，更需要在任务规划、推理决策、工具调用、多轮交互及安全对齐等关键能力维度上提供系统性支撑。然而，从现实情况看，当前数据供给在规模、质量、结构及标准化程度等方面仍难以充分匹配智能体快速发展的需求，数据孤岛、标注规范缺失、质量评估体系滞后以及安全合规挑战等问题依然突出。

在国家层面，围绕“人工智能+”行动部署及数据要素市场化配置改革，一系列政策文件持续推动高质量数据集建设与开放共享，强调以数据为纽带促进技术创新与产业应用协同发展。在产业层面，通

用平台与行业专用智能体双轨并进，对高质量、可复用、可持续演进的数据资源提出了更加精细化和场景化的需求。在技术层面，多智能体协同、本体知识建模、多模态感知及强化学习等前沿方向的发展，进一步强化了对结构化语义数据、过程数据与反馈数据的系统需求。

基于上述背景，构建面向智能体能力构建的高质量数据集建设体系，已成为提升人工智能核心竞争力的重要基础工程。本指南立足智能体能力发展需求，系统梳理数据类型、规模与质量要求，提出面向能力与场景的数据集建设总体框架与关键方法，构建覆盖数据采集、标注、评估与迭代优化的全生命周期管理路径。同时，结合行业实践与发展趋势，对数据合规、安全治理及生态协同机制进行了系统分析，力求为各类主体开展高质量数据集建设提供参考依据与实施路径。

目 录

一、智能体发展现状与趋势	1
(一) 智能体发展现状	1
(二) 智能体技术发展趋势对数据建设的牵引	4
(三) 智能体数据供给面临的关键问题与挑战	7
二、智能体能力构建的高质量数据集需求分析	9
(一) 面向智能体六大能力的数据需求分析	10
(二) 行业场景对数据的特殊需求	14
(三) 智能体高质量数据集需求清单	17
三、面向智能体能力构建的行业高质量数据集建设总体框架与关键方法	20
(一) 建设总体思路	20
(二) 技术路线	22
(三) 关键方法	23
四、面向智能体能力构建的行业数据集建设路径	34
(一) 数据集建设阶段框架	34
(二) 数据集建设的主体模式与协同机制	39
(三) 数据合规、安全与伦理治理	41
五、发展展望和建议	46
(一) 行业数据集发展趋势展望	46
(二) 推动行业高质量数据集发展的建议	48
附件 行业智能体高质量数据集建设代表性实践	54
(一) 企业运维领域：面向企业内部 IT 运维智能体的高质量数据集建设	54
(二) 医疗健康领域：专病、临床辅助决策、血管智能评估智能体高质量数据集建设	57
(三) 通信领域：网络运维智能体高质量数据集建设	59
(四) 知识产权领域：专利智能体高质量数据集建设	61
(五) 汽车领域：汽车金融智能体高质量数据集建设	63
(六) 产业园区领域：园区智能体高质量数据集建设	66
(七) 科研领域：模拟计算智能体高质量数据集建设	68

图 目 录

图 1 智能体任务规划数据流转逻辑 11

图 2 行业应用数据赋能逻辑 14

图 3 面向智能体能力构建的数据集建设技术路线 23

图 4 数据采集与生成闭环结构图 23

图 5 人工标注技术路线图 28

图 6 人机协同标注流程图 28

图 7 自动化标注技术路线图 29

图 8 综合治理框架 42

表 目 录

表 1 环境感知多源异构数据分类明细 11

表 2 工具调用核心数据明细 12

表 3 安全可控核心数据分类 13

表 4 智能体数据类型需求清单 17

表 5 智能体数据规模与质量需求清单 18

表 6 智能体标注体系需求清单 19

表 7 面向智能体能力构建的数据采集与生成关键方法 25

表 8 智能体能力域与标注任务映射重点 26

表 9 复杂场景专项标注方案建议 30

表 10 面向智能体能力构建的数据质量评价重点 31

表 11 数据质量自动评估能力分层 32

表 12 未来行业高质量数据集发展方向 48

一、智能体发展现状与趋势

当前，人工智能正从感知智能向认知智能、决策智能跃升，智能体（AI Agent）作为连接大模型能力与行业应用场景的核心载体，成为培育新质生产力的重要抓手。

（一）智能体发展现状

2024 年以来，全球智能体产业进入从技术验证向规模化商用迈进的加速期，逐步形成“模型层—框架层—应用层”协同演进的产业生态。高质量数据已成为决定各国智能体产业竞争力的关键要素。

1. 国际智能体发展态势与产业布局

产品能力持续迭代升级，驱动智能体走向自主执行。在产品能力方面，OpenAI 于 2025 年 5 月推出 Codex 编码智能体，主要在云端隔离沙盒中围绕代码仓库执行编码、测试、修复和 PR 生成；Anthropic 的 Claude Code 在自主编码调试领域实现突破性进展，Computer Use 功能支持桌面端长时程任务规划与执行；Google 发布 Gemini Enterprise，提供无代码 Agent Designer 与多智能体协作能力；Microsoft 基于 Copilot Studio 实现低代码多智能体编排，已有超 23 万家组织构建智能体应用；Amazon 推出 AgentCore，提供企业级多步自动化运行时环境与全栈评测能力。在工程化支撑方面，OpenAI 开源 Agents SDK 与 Apps SDK，Google 构建 Agent Gallery 发现机制与治理框架，有效降低了智能体的开发部署门槛。在开源生态方面，开源框架活跃度持续攀升，OpenClaw 作为 2026 年最受关注的开源自托管智能体项

目，GitHub 星标数已突破 30 万¹，通过技能系统架构实现能力模块化扩展，推动智能体技术向普惠化方向演进。

标准协议加速生态互联，夯实跨平台协同底座。异构系统间的互联互通成为制约智能体产业规模化发展的关键瓶颈，国际社区围绕多层通信协议与中立治理组织加速推进标准化建设。通信协议标准化是基础，Anthropic 发布的 MCP 协议实现大模型与外部数据源、工具的标准双向安全连接，已被 OpenAI、Google 等主流平台广泛采纳。开放社区共建是保障，将协议演进、标准维护与生态协同置于独立第三方框架下，避免由单一厂商主导生态规则。2025 年 12 月 Linux 基金会成立智能体人工智能基金会（AAIF），通过统一接口规范与互操作实践加速生态从封闭孤岛向开放互联演进，由 Anthropic、OpenAI 和 Block 等联合创始，Google、Microsoft、Amazon 等主要厂商作为会员加入²。

企业部署持续走深走实，加速释放规模化价值。部署重心由 PoC 试点转向生产运行。美国厂商以云端平台和开发工具为依托，将智能体深度集成至办公生产力与企业级 workflows；日韩厂商聚焦电信网络自动化、搜索引擎智能化等本土优势场景，推动智能体与既有数字基础设施的深度融合；欧洲在合规治理框架下稳步推进，以监管沙盒和主权数据中心为保障有序探索高风险场景应用。根据 MarketsandMarkets 预测³，全球智能体市场将从 2024 年的 51 亿美元增长到 2030 年的 471 亿美元，年复合增长率达 44.8%。

¹ <https://github.com/openclaw/openclaw>

² <https://www.linuxfoundation.org/press/linux-foundation-announces-the-formation-of-the-agentic-ai-foundation>

³ <https://www.marketsandmarkets.com/Market-Reports/ai-agents-market-15761548.html>

2.国内智能体发展态势与典型应用场景

顶层设计布局持续强化，引领智能体发展方向。党中央、国务院高度重视人工智能发展，面对新一轮科技革命带来的战略机遇，围绕智能体技术创新与产业应用作出系列重要部署。2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》，从国家战略高度部署推动人工智能与科技、产业、消费、民生、治理等重点领域深度融合⁴。《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，抢占人工智能产业应用制高点，全方位赋能千行百业”。2026 年 7 月，国家互联网信息办公室发布《人工智能拟人化互动服务管理暂行办法》，就智能体训练数据管理、用户交互数据保护、未成年人信息合规等关键环节提出规范要求⁵，为智能体健康有序发展提供了制度保障。2026 年 1 月，工业和信息化部等八部门联合印发《“人工智能+制造”专项行动实施意见》，明确提出“推动 3 至 5 个通用大模型在制造业深度应用，打造 100 个工业领域高质量数据集，推广 500 个典型应用场景，培育 1000 个高水平工业智能体”⁶。截至 2025 年 6 月，全国已建设高质量数据集超 3.5 万个、总量超 400PB⁷，为智能体训练与应用夯实了数据要素基础。随着顶层设计与行业专项政策的持续完善，智能体在政策牵引下正加速由技术探索迈向规模化落地的新阶段。

4 https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html

5 https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm

6 <https://www.miit.gov.cn/>

7 <https://www.caict.ac.cn/>

应用格局加速成形，通用平台与行业专用双轨并进。国内智能体应用呈现通用平台与行业专用智能体双轨并进的发展格局。通用平台拓展应用广度、降低开发门槛，行业专用智能体深耕垂直场景、提升服务深度，二者协同推动智能体向千行百业渗透。**通用平台方面**，以字节扣子（Coze）、百度千帆 AppBuilder、阿里云百炼平台、华为 AI 原生应用引擎等为代表的科技企业围绕低代码/无代码开发、可视化 workflow 编排、插件生态、知识库集成等核心能力持续迭代。**行业专用智能体方面**，依托通用平台底座叠加行业大模型与高质量数据集定制而成，深度融合行业机理、业务流程与私有数据，在制造、电信、金融、医疗、政务等重点行业持续走深向实。

（二）智能体技术发展趋势对数据建设的牵引

1. 多智能体协同催生协作通信数据体系建设

多智能体协同通过标准化协议实现智能体间的发现、通信与任务委托，显著提升复杂系统任务的处理能力。在协同运行过程中，不仅形成新的交互模式，也沉淀出一批具有较高价值的新型数据资源，为智能体训练优化和能力评测提供重要支撑。**一是协同任务与决策数据**，多个智能体围绕复杂任务开展协同过程中，会产生任务拆解、角色分工、协同决策、结果汇总等数据。这类数据能够完整反映复杂任务的求解过程，是构建智能体规划能力、协作能力和复杂推理能力训练数据的重要来源。**二是协作轨迹数据**。协同过程中形成的任务执行路径、工具调用链路、决策过程记录、反馈修正过程等数据，能够系统反映智能体完成任务的行为模式和推理逻辑，是开展监督微调、强化学习

和智能体评测的重要数据基础。**三是共享知识与环境数据。**多智能体通常依托共享知识库、公共记忆空间或统一环境开展协作，由此形成知识调用记录、上下文关联关系、环境状态变化等数据资源，这类数据有助于构建面向复杂场景的知识增强数据集，提升智能体对动态环境的理解与适应能力。

2. 本体智能凸显结构化语义数据建设重要性

知识本体通过对领域概念、实体及其关系的形式化规约，为智能体提供结构化的推理基础，有效弥补统计模式驱动下大语言模型在语义一致性和逻辑推理方面的不足。当前，大模型与本体的深度融合已成为重要趋势，本体构建模式正从传统的纯人工构建向“大模型自动本体抽取+人工校验”的混合模式演进，由此衍生出对混合标注数据、校验样本及自动抽取结果修正数据的新型需求，进一步凸显了结构化语义数据体系建设的重要性。**一方面**，结构化知识标注数据的重要性大幅提升，构建高质量知识图谱需要大量的实体识别、关系抽取、事件标注和因果关系标注数据，在混合构建模式下，标注工作还需覆盖对大模型自动抽取结果的审核纠错，标注的专业性和复杂性显著提高。**另一方面**，知识图谱中任何实体缺失、关系错误或语义不一致都可能导致推理偏差，行业本体的设计与标准化建设、混合模式下自动抽取结果的质量验证与人工核校以及知识图谱的持续更新维护成为当前亟待解决的关键课题。

3. 环境感知与决策能力强化驱动多模态与过程数据建设

随着智能体从以文本理解为主向“感知—决策—行动”闭环能力

演进，其对环境信息的理解与基于上下文的决策能力不断增强，对数据形态与标注方式提出新的要求。一是多模态环境感知数据，智能体需要融合文本、图像、视频、语音等多源信息以理解复杂环境语境，推动多模态对齐数据、跨模态标注数据的大规模建设，并对不同模态之间的语义一致性提出更高要求。二是情境化决策与反馈数据，智能体在动态环境中进行任务规划与策略调整，需要构建包含任务上下文、决策路径及反馈结果的数据集，以支撑其策略优化与泛化能力提升。三是行动过程与工具使用数据，智能体通过调用工具、执行操作完成任务，其过程数据（如调用序列、参数配置、执行结果等）成为重要训练资源，但当前在采集规范、过程标注及效果评估方面仍较为缺乏。

4.行业适配驱动领域专属语料建设深化

行业深度适配通过融合企业私有数据、业务规则与专家经验，推动智能体从通用助手向行业专家演进，对领域专属数据的规模、质量与专业性提出更高要求。一是领域专业知识的数据化，将资深专家的经验、操作手册、故障案例库等隐性知识转化为高质量的问答对、思维链数据和知识图谱，形成智能体可利用的领域知识资产。二是业务流程数据的规范化，将企业的业务流转规则、审批逻辑转化为智能体可执行的标准作业程序数据，支撑复杂业务流程中的准确执行。三是合成数据在垂直领域的规模化应用，在金融等数据敏感或样本稀缺领域，利用仿真技术生成符合行业分布特征的高质量合成数据，有效弥补真实数据的不足。

（三）智能体数据供给面临的关键问题与挑战

1. 数据供给质量与规模尚难匹配应用需求

数据质量方面，智能体的推理与规划能力需要经过思维链标注的高质量数据支撑，然而互联网公开数据多为碎片化信息，缺乏逻辑严密的推理过程；构建此类数据集往往需要领域专家深度参与，数据清洗、标注等环节所需成本占比高，成效不够显性，价值转化周期长⁸。**数据模态方面**，市场对多模态数据、思维链数据等细分领域的需求呈井喷之势⁹，但能够精准对齐的跨模态数据极度匮乏，非结构化数据的清洗与语义对齐技术尚不成熟。此外，基础数据的开放汇聚难度较大，权属规则不明晰、供给激励机制不明确以及安全合规边界模糊，使得数据提供方的供给意愿难以有效提升¹⁰，无论公共数据授权运营还是跨主体的行业高质量数据集共建，实践中均面临较大的沟通协调困难，数据孤岛问题普遍存在，制约了行业智能体的能力提升。

2. 企业数据治理面临价值释放与安全平衡难题

在企业侧数据资源向数据资产转化的过程中，价值释放与安全保障间的平衡问题日益凸显。**数据资产转化方面**，企业拥有海量高价值的业务数据，但多以非结构化文档、日志等形式存在，企业内部管控和数据治理水平仍需提高，质量较差的数据将影响数据资产的质量及长期可用性¹¹。**数据安全防护方面**，企业在将私有数据接入大模型或第三方智能体平台时，面临商业机密泄露风险，完善的数据脱敏与分

⁸ <https://www.caict.ac.cn/kxyj/qwfb/ztbg/202511/P020251128616212191150.pdf>

⁹ https://www.nda.gov.cn/sjj/jgsz/jld/llh/llhldhd/0323/20260323202204680553721_pc.html

¹⁰ <https://www.caict.ac.cn/kxyj/qwfb/ztbg/202511/P020251128616212191150.pdf>

¹¹ https://pdf.dfcfw.com/pdf/H3_AP202410031640156502_1.pdf

级分类机制亟待建立。**权限治理方面**，智能体具备的自主决策能力给现有数据权限管理体系带来冲击，多智能体协作时责任边界模糊¹²，智能体在执行跨部门、跨系统任务时的数据访问边界如何动态界定，因数据质量问题导致的决策失误如何进行责任认定，当前企业治理框架中尚缺乏明确规范。此外，数据权属的使用范畴和权属界定尚不清晰，影响数据资源要素化、资产化及整个价值链的传导与利益分配¹³。

3. 个人数据权益保障机制亟待健全完善

随着智能体与用户交互的深度和频率不断提升，个人数据权益保护面临新的制度性挑战。**知情同意方面**，传统隐私协议难以覆盖智能体动态、持续的数据收集行为，用户对自身数据的流向和用途往往缺乏充分了解和有效掌控。《人工智能拟人化互动服务管理暂行办法》明确提出，“除法律、行政法规另有规定或者取得用户单独同意外，提供者不得将用户交互数据、用户敏感个人信息用于模型训练”，并要求“向用户提供删除交互数据的选项”¹⁴，体现了对用户数据权益保护的制度性回应。**数据确权方面**，用户在使用智能体过程中产生的高质量交互数据实质上为模型优化提供了重要贡献，但当前缺乏有效的个人数据确权与价值评估机制，用户无法分享数据带来的收益，长此以往可能影响用户贡献数据的积极性。

4. 标注规范与质量评估体系建设相对滞后

面向智能体的数据生产与管理实践中，统一的标注规范与系统化

¹² <https://aloudata.com/news/aloudata-noetl-automated-data-management-whitepaper>

¹³ https://pdf.dfcfw.com/pdf/H3_AP202410031640156502_1.pdf

¹⁴ https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm

的质量评估体系尚未成熟，制约了数据资源的高效利用。**标注规范方面**，智能体训练需要大量包含中间推理过程的数据，但行业内对思维链、反思过程以及多智能体交互逻辑的标注尚无统一标准，不同厂商采用的标注格式和质量要求各异，数据集之间难以兼容互通。**质量评估方面**，《高质量数据集建设指引》区分了静态质量与动态质量两个评估维度¹⁵，但现有评估实践主要关注格式规范性与准确性，缺乏针对数据逻辑性、完整性与安全性的深度评估指标，评估维度的系统性和全面性有待增强。此外，动态环境下数据的时效性管理难度可能导致智能体产生错误判断，高效的数据时效性检测与更新机制亟待建立。

5. 数据安全合规治理面临多维新挑战

在智能体广泛应用的背景下，数据安全与合规治理问题呈现出跨域、多维的新特征。**跨境数据流动方面**，跨国企业的智能体涉及数据的全球调度，各国对数据出境的监管要求日益严格，如何在满足业务协同需求的同时确保数据流动的合规性，是跨国智能体部署面临的重要挑战。**版权与伦理方面**，智能体生成内容的版权归属、训练数据中涉及的版权素材使用等问题尚存争议，相关法律法规和行业共识仍在形成过程中。在金融等高风险场景中，若训练数据包含偏见或错误知识，可能导致智能体产生不当决策，对高风险场景数据的安全审查与伦理管控提出了更高要求。

二、智能体能力构建的高质量数据集需求分析

与基础大模型和行业模型数据集主要服务于知识学习、语言理解

¹⁵ <https://www.caict.ac.cn/>

和专业能力训练不同，智能体数据集面向任务执行能力构建，重点支撑智能体在复杂环境中的感知、规划、决策、执行与反馈优化等能力形成。智能体数据集不仅包含任务相关知识和领域数据，还应包含任务目标、环境状态、工具调用过程、决策路径、执行结果、用户反馈等动态交互信息，能够完整反映智能体完成任务的全过程。因此，智能体数据集的核心特征不在于知识内容本身，而在于对“任务—决策—行动—反馈”闭环过程的记录与表达。从能力支撑角度看，基础大模型数据集主要解决“知道什么”的问题，行业模型数据集主要解决“懂什么”的问题，而智能体数据集则主要解决“如何完成任务”的问题，是支撑智能体自主执行和持续优化能力的重要数据基础。

（一）面向智能体六大能力的数据需求分析

1. 环境感知能力：多源异构上下文数据需求

环境感知能力是指智能体对其所处运行环境及任务上下文进行理解与建模的基础能力，其本质在于对多源信息的融合解析与语义对齐。随着智能体应用场景从单一文本处理向复杂任务执行演进，其感知对象已由单一数据模态扩展至覆盖物理环境、数字系统及交互语境的多维信息体系。从数据来源划分，环境感知数据可分为物理环境数据与数字空间数据两大类。物理环境数据主要来源于各类终端设备，用于刻画现实世界的状态信息；数字空间数据则来源于各类信息系统与平台，反映业务运行状态、用户行为特征及系统交互过程。此外，随着智能体与用户及其他系统的持续交互，交互过程本身也构成重要的上下文数据来源，对理解任务意图与动态调整策略具有关键作用。

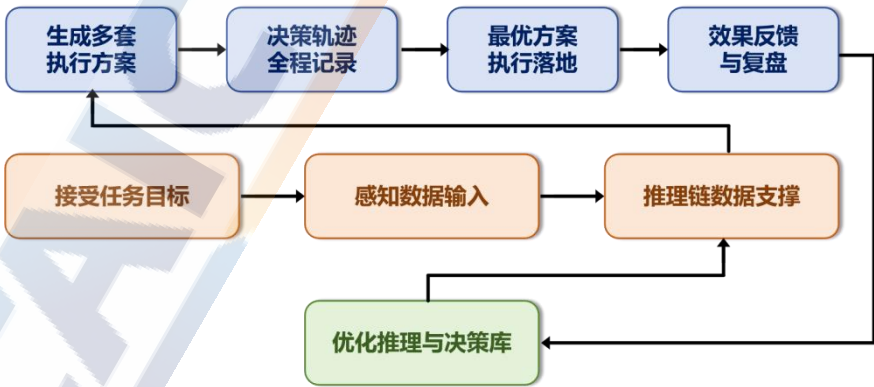
表 1 环境感知多源异构数据分类明细

数据来源	数据类型	典型数据内容	核心用途
终端设备	结构化/非结构化	图像视频、空间位置、设备状态等	物理环境状态感知、实体识别与状态监测
数字系统与业务平台	结构化/半结构化	系统日志、业务数据、运行指标、流程状态、权限信息等	数字环境建模、业务流程感知与系统状态监测
网络与外部信息源	全类型	公开数据、接口数据、网络状态、第三方服务信息等	外部信息补充、环境态势感知
交互与上下文数据	非结构化/半结构化	用户输入、多轮对话历史、任务上下文、操作记录等	意图识别、上下文理解与动态决策支撑

来源：中国信息通信研究院

2.任务规划能力：推理链与决策轨迹数据需求

推理链数据侧重底层逻辑支撑，是智能体开展规划的核心依据，覆盖各类规则逻辑、因果关系、条件判断、约束限制、领域公理、常识范式、历史案例经验、标准作业流程、行业规范等内容，兼顾演绎推理、归纳推理、类比推理、反事实推理等多种推理形式。这类数据能够支撑智能体将总目标拆解为细分步骤，由前置条件推导后续动作，由约束条件筛选可行方案。



来源：中国信息通信研究院

图 1 智能体任务规划数据流转逻辑

3.工具调用能力：工具使用与 API 交互数据需求

工具使用数据覆盖工具全维度信息，核心包括各类工具的核心功能定位、适用场景范围、使用限制条件、标准操作流程、参数配置规范、权限准入要求、优缺点对比、使用注意事项、多工具协同配合逻辑、工具切换衔接规则、调用效果判定标准等内容。API 交互数据聚焦接口调用的标准化、稳定化流程，是实现智能体与外部系统联通的关键，包含接口协议类型、数据传输格式、请求参数规范、返回字段含义、状态码对应说明、异常报错处理方式、超时机制设置、身份鉴权规则、调用频率限制、数据加密要求、数据传输安全规范等细则。除基础信息数据外，工具调用还需要历史调用效果数据、最佳调用时机数据、调用失败修复方案、多工具串联配合数据等。

表 2 工具调用核心数据明细

数据类别	核心数据项	质量要求	作用
工具基础数据	功能、场景、流程、权限、参数	全面、准确、规范	指导工具选用与基础操作
API 交互数据	协议、格式、参数、鉴权、异常处理	标准、统一、稳定	保障接口调用通畅安全
调用历史数据	调用时机、效果、失败案例、优化方案	完整、可追溯、可迭代	优化调用策略，提升效率

来源：中国信息通信研究院

4.多轮交互能力：上下文管理与对话数据需求

对话交互数据包含口语化表达、书面化表达、方言俚语、缩略语、隐含表达、情绪化话术等多种形式，充分贴合真实沟通场景。除原始语料外，还需要高质量标注数据帮助智能体精准识别交互意图、判断情绪倾向、理解深层潜台词、抓取核心诉求，避免误解、误读。上下

文管理数据承担着信息留存、关联、提取的功能，完整记录全程交互内容，确保在长时间、多回合交互中不丢失核心信息、不出现前后矛盾、不偏离核心诉求。

5.安全可控能力： 风险场景与对齐数据需求

风险场景数据覆盖全类型、全维度风险样本，构建立体化风险识别库。全面的风险数据能够帮助智能体快速识别风险隐患、提前预警异常、主动规避危险，从源头减少安全事故发生。价值对齐数据侧重行为约束与规范引导，是确保智能体行为符合人类预期、合规伦理的核心，约束智能体的决策、执行、交互行为，让其始终在合规、合理、安全的框架内运行，不触碰红线、不偏离轨道。

表 3 安全可控核心数据分类

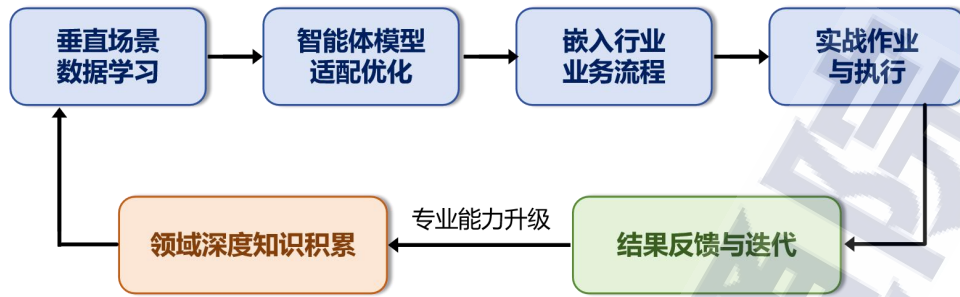
数据模块	核心内容	应用目标
风险样本数据	违规案例、漏洞信息、风险指标、禁区条款	风险识别、预警、规避
价值对齐数据	行为规范、伦理准则、合规要求、权责标准	约束行为，贴合人类预期
应急处置数据	止损流程、修复方案、审计规则、整改办法	异常处置，降低损失

来源：中国信息通信研究院

6.行业应用能力： 领域知识与场景数据需求

领域深度知识数据是智能体具备专业素养的核心，覆盖行业研发、生产、销售、服务、管理全链条知识点，帮助智能体建立完整的专业知识体系，适配行业专业作业要求。垂直场景实操数据贴合行业真实落地需求，具备极强的针对性、实用性与地域性，包括行业真实业务数据、细分用户画像、个性化需求特征、设备运行参数、故障案例库、

执行结果数据、场景约束条件、行业特殊要求等。



来源：中国信息通信研究院

图 2 行业应用数据赋能逻辑

（二）行业场景对数据的特殊需求

不同行业的业务流程、专业知识、监管要求、风险容忍度及价值创造模式存在显著差异，导致智能体在全生命周期对数据产生多元化、精细化甚至矛盾性的特殊需求。

1. 领域“行话”与“暗知识”¹⁶数据需求

在行业智能体能力构建过程中，通用语料难以覆盖专业场景，领域知识尤其是隐性经验类知识的数据化需求日益增加。一是结构化领域知识图谱，金融、法律、工业制造等行业核心是高度结构化、逻辑严密的知识体系。这类数据需求强调实体、关系、属性的精确性、一致性与完整性，通常来源于权威的教科书、行业标准、专利库、专业文献数据库（如 PubMed）、企业内部的规范文档与知识库等。二是非结构化领域经验与“暗知识”，大量行业知识以非结构化、经验化的形式存在，是教科书上不会记载的“行话”、“技巧最佳实践”和“失败案例”。这类数据通常在历史工单、维修记录、专家访谈录音、内

¹⁶ 段云峰,张欢,马田瑶,等.智能体互联网的概念、实践与演进[J].通信世界,2026,(03):11-16.

部会议纪要、优秀案例的实操记录中，高度场景化、隐含性强，且质量极度依赖领域专家的标注与提炼。**三是领域任务链与流程数据**，行业任务往往涉及多步骤、有条件分支，需要智能体学习大量历史成功任务的过程数据。这类过程追踪数据对于训练智能体进行复杂规划、工具使用和反思至关重要。

2. 多模态时空融合数据需求

在智能体处理问题从单一模态向复杂场景演进的过程中，多模态融合及其跨模态语义对齐能力对数据提出了更高要求。**一是核心业务模态数据**，不同行业的核心数据模态差异巨大。例如工业与制造业中的核心数据包括设备读数、生产线控制信号等时序数据、质检图像、监控视频等视觉数据、设备异响等声学数据。**二是跨模态对齐与关联数据**，行业智能体要做出精准决策，往往需要理解不同模态数据之间的语义关联。这就要求数据集不仅包含多模态数据，更要提供丰富的跨模态关联标注，如图文对、视频片段与事件描述、设备运行数据与状态标签等。**三是时空上下文数据**，物流、农业、能源等行业具有强烈的时空属性。数据集需要包含精确的时间戳和地理位置信息。

3. 实时性、连续性与概念漂移数据需求

一是实时、准实时数据流，在决策窗口极短的场景，智能体需要基于源源不断的实时数据流进行毫秒级或秒级响应，要求数据供给管道具备高吞吐、低延迟的特性。数据集不仅包含历史样本，更应包含模拟真实数据流的、带有时间戳的连续序列数据，用于训练智能体的在线学习和快速适应能力。**二是连续性状态数据**，在化工、冶金等过

程工业或长期健康管理中，智能体需要关注系统或个人的连续状态演变。单一时刻的快照数据价值有限，有价值的是长期的、连续监测的数据轨迹。这类数据需求强调长期、稳定、无间断的采集能力。**三是概念漂移与分布外数据**，行业数据集需要能反映周期性、趋势性变化，并可能需要包含分布外（OOD）或极端情况的样本，以增强智能体的鲁棒性和适应能力。

4.安全、合规、伦理与隐私保障

在行业数据深度应用过程中，安全合规与伦理约束成为数据供给不可回避的基础性要求。**一是数据脱敏与匿名化**，数据预处理流程应符合相关法律法规及行业特殊规定，并且脱敏过程不能过度破坏数据中对于任务有用的统计模式和关联关系。高质量的脱敏数据本身就是一种关键需求。**二是可审计与可解释的数据链路**，模型应具备可解释性，训练数据来源应清晰、处理过程应可审计。数据集需要完备的元数据，记录每个数据样本的来源、采集时间、处理步骤、标注人员、版本信息等，形成完整的“数据谱系”。**三是公平性与偏见控制**，数据在涉及敏感属性的维度上（如性别、年龄、地域）具有代表性和平衡性，或者至少需要对数据中存在的偏见进行测量、标注和说明，以便在训练中采取去偏见技术。**四是主权与本地化要求**，在政务、国防及一些关键基础设施领域，数据可能存在境内存储、不得出境的强制要求，这催生了在封闭、安全环境下构建和利用高质量数据集的需求，即“数据不出域，模型可进线”。

5.场景价值可衡量性与强化学习信号

为了支撑智能体在复杂任务中的持续优化与自主决策能力，需要构建可量化的反馈信号体系与高质量训练样本。**一是稀疏且延迟的业务回报信号**，在很多复杂场景中，智能体一系列动作的最终好坏（业务结果）反馈是稀疏且延迟的，需要通过领域模拟器生成数据，或对历史成功及失败案例进行深度归因分析，构建稠密的中间奖励信号。**二是高质量演示数据与人类反馈**，包括记录专家完成任务的全过程轨迹数据（感知—决策—执行序列）。更进一步，可以收集人类对智能体生成结果（如多个诊断建议、多个文案方案）的偏好排序或评分数据，即人类反馈数据，用于对模型进行更精细的微调对齐。**三是仿真与合成数据**，利用高保真行业仿真环境生成合成数据，或利用生成式 AI 技术（如 Diffusion 模型）生成符合物理规律的合成数据（如带有各种缺陷的工业品图像），是满足此类数据需求的重要途径。其核心需求是仿真的保真度和合成数据的多样性、有效性。

（三）智能体高质量数据集需求清单

1.数据类型需求清单

智能体的能力构建依赖于多元化、分层次的数据输入，不同类型数据在不同训练阶段发挥关键作用。

表 4 智能体数据类型需求清单

数据大类	数据子类	典型数据形态	核心应用场景	质量要求
结构化数据	业务数据	数据库表、业务表单、交易记录	业务状态感知、决策依据	完整性、准确性、时效性
	知识图谱	实体关系三元组、属性图谱	认知推理、知识查询	准确性、一致性、完整性
	配置数据	参数配置、规则定义、权限配置	系统配置、策略执行	规范性、准确性、版本管理

表 4 智能体数据类型需求清单（续）

数据大类	数据子类	典型数据形态	核心应用场景	质量要求
结构化数据	日志数据	系统日志、操作日志、审计日志	行为分析、故障诊断	完整性、时序性、可追溯
半结构化数据	接口数据	JSON/XML 格式的 API 请求/响应	工具调用、系统交互	标准化、一致性、兼容性
	标记数据	HTML、Markdown、配置文件	内容解析、格式转换	规范性、可读性、兼容性
	标注数据	带标签的样本数据	模型训练、能力评估	准确性、一致性、覆盖度
非结构化数据	文本数据	自然语言文档、对话记录、专业文献	语言理解、知识获取	准确性、多样性、领域覆盖
	图像数据	照片、截图、扫描件、医学影像	视觉感知、目标识别	清晰度、标注精度、场景覆盖
	音频数据	语音录音、环境声音、音乐	语音识别、声纹分析	音质、降噪、标注准确性
	视频数据	监控视频、操作录像、教学视频	行为分析、过程追踪	分辨率、帧率、时序标注
	时序数据	设备状态和传感数据、指标监控、信号波形	状态监测、趋势预测	采样频率、连续性、同步性

来源：中国信息通信研究院

2.数据规模与质量需求清单

智能体应用对数据的规模与质量提出了差异化要求，不同训练阶段、不同能力维度对数据的需求存在显著差异。

表 5 智能体数据规模与质量需求清单

维度	具体指标	预训练	监督微调	强化学习	评估测试
数据规模	样本量级	千亿级 token/百万级样本	万级至十万级样本	十万级交互轨迹	千级至万级测试用例
	覆盖领域	跨领域通用语料	目标任务领域	应用场景覆盖	能力维度覆盖
	数据更新	持续增量更新	定期批次更新	在线实时更新	版本迭代更新
数据质量	准确性	基础准确，允许噪声	高准确，错误率<1%	高准确，偏好一致	绝对准确，可验证
	完整性	字段完整率>95%	字段完整率>99%	轨迹完整性>99%	用例完整性100%
	一致性	跨源一致性>90%	标注一致性>95%	偏好一致性>95%	标准一致性100%

表 5 智能体数据规模与质量需求清单（续）

维度	具体指标	预训练	监督微调	强化学习	评估测试
数据质量	时效性	近 5 年数据为主	近 2 年数据为主	实时或准实时	反映最新能力
	多样性	跨领域、跨模态	场景、任务多样	策略、路径多样	边界、极端覆盖
	代表性	覆盖主流场景	覆盖目标任务	覆盖决策空间	覆盖能力边界
特殊要求	偏见控制	降低历史偏见	避免标注偏见	防止偏好偏见	检测能力偏见
	可解释性	来源可追溯	标注可解释	偏好有依据	结果可验证
	安全性	过滤有害内容	强化安全对齐	防止奖励攻击	检测安全风险

来源：中国信息通信研究院

3.标注体系需求清单

为实现高质量、可泛化的智能体行为，标准化、多维度的数据标注体系至关重要。表 6 系统梳理了智能体训练所需的数据标注体系。

表 6 智能体标注体系需求清单

标注维度	核心要素	具体要求	应用场景
标注类型	分类标注	类别定义清晰、边界明确、覆盖全面	意图识别、情感分类、对象分类
	实体标注	实体类型定义、边界识别、属性标注	命名实体识别、关系抽取
	关系标注	关系类型定义、主体客体识别、关系方向	知识图谱构建、关系推理
	序列标注	标签体系、标注粒度、上下文关联	分词、词性标注、语义角色标注
	轨迹标注	步骤拆解、决策点标注、反馈标注	任务规划、强化学习训练
	偏好标注	比较维度、排序标准、一致性检验	RLHF 训练、价值对齐
标注规范	标注指南	定义清晰、示例丰富、边界说明	统一标注标准、降低歧义
	质量标准	准确率、一致性、完整性指标	质量评估与验收
	边界案例	典型边界、处理方式、参考示例	处理模糊情况、统一处理标准
标注流程	任务分发	任务拆分、难度分级、人员匹配	提升标注效率、合理分配资源

表 6 智能体标注体系需求清单（续）

标注维度	核心要素	具体要求	应用场景
标注流程	多人标注	冗余度设置、一致性检验、仲裁机制	提升标注质量、处理分歧
	质量审核	抽检比例、审核标准、问题反馈	质量控制、持续改进
	迭代优化	问题汇总、规范更新、培训复训	持续改进标注质量
质量指标	准确率	标注结果与参考答案的一致率	衡量标注正确性
	一致性	多人标注的一致率、重标一致率	衡量标注稳定性
	完整性	必填字段的填充率、轨迹的完整度	衡量标注完整度
	时效性	标注任务的完成时效、更新频率	衡量标注效率
工具支撑	标注平台	多模态支持、协同标注、版本管理	提升标注效率
	质检工具	自动质检、异常检测、统计分析	辅助质量控制
	培训系统	标注教程、练习测试、能力评估	标注人员培训

来源：中国信息通信研究院

三、面向智能体能力构建的行业高质量数据集建设总体框架与关键方法

（一）建设总体思路

1.指导原则

- （1）**场景驱动**：紧密围绕行业典型应用场景，建立场景与数据映射关系，明确各场景下的数据类型、规模及质量要求。
- （2）**能力导向**：以智能体核心能力为牵引，建立能力与数据映射关系，确保数据集有效支撑各项能力训练与优化。
- （3）**任务牵引**：聚焦决策、推理、协作等任务类型，形成任务与数据映射关系，明确数据输入输出规范、标注标准及质量指标。
- （4）**可信评估**：将验证与确认机制系统性融入数据全生命周期，

确保数据集作为“动态可信认知资产”的可验性与可靠性。

（5）安全合规：严格遵循《数据安全法》《个人信息保护法》等法律法规，贯彻国标、行标要求，确保数据的一致性与可复用性。

（6）质量贯穿：建立“采集—清洗—标注—评估—优化”闭环管理机制，防范数据污染、标注不一致、测试集泄露等隐患。

（7）可持续演进：保障数据多维度多样性与代表性，建立版本管理与动态更新机制，推动数据集持续优化。

2.建设目标

一是夯实基础能力，实现核心达标。建成符合国家及行业标准的高质量数据集，确保其在数据说明文档规范性、数据内在质量和模型应用效能三个核心维度均达标。针对既定模型类型形成关键数据集的生产能力，满足模型微调与基准测试的基础需求。

二是完善体系覆盖，构建闭环机制。实现对多模态、多任务、多模型研发阶段的体系化数据支持，全面覆盖指令微调、人类偏好、合成数据、问答对、专业化标注数据等多样化数据类型，以满足模型预训练、有监督微调、强化学习及性能评估等各环节的差异化需求。建立并运行“建设—评价—优化—迭代”的数据质量闭环管理机制，使数据集能够依据模型反馈与应用场景变化进行持续的动态优化与版本更迭，显著提升其对复杂业务场景的适应性与支撑效能。

三是推动生态化发展与价值深度对齐。推动数据集体系从内部闭环向开放生态演进，形成数据自驱动、自演进的可持续发展能力，使之能够自适应地跟随模型技术、行业规范及应用需求的演进而迭代。

在价值层面，不仅要求数据内容严格符合人类伦理与价值观，全面规避歧视性、偏见性及违法不良信息，更将深度集成思维链、逻辑推理等深层信息，旨在从根本上提升模型的认知推理能力、可解释性与安全性，最终打造安全、可信、可靠且具备持续生命力的 AI 数据基础设施。

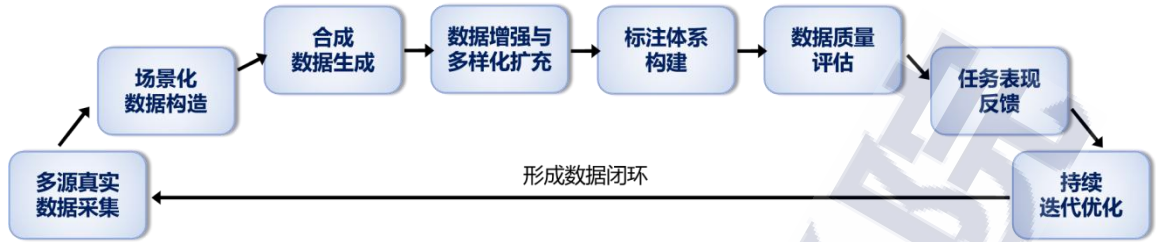
（二）技术路线

面向智能体能力构建的数据集建设，应从传统的样本汇集模式转向面向能力形成和场景落地的数据组织模式。与面向单一模型训练的数据集相比，智能体数据集除静态知识和结果标签外，还应覆盖任务上下文、状态变化、工具调用、执行反馈和异常处置等动态信息，以支撑感知、理解、规划、决策、执行和反馈全链路能力训练。

总体上，应坚持“能力牵引、场景驱动、真实优先、质量闭环、持续迭代”的原则，形成覆盖数据获取、结构化加工、质量校验、效果反馈和版本优化的闭环体系。建设重点不在于单一环节扩张规模，而在于围绕目标任务形成可持续、可评估、可复制的数据供给能力，使数据集能够同时支撑训练优化、评测验证和应用落地。

从总体路径看，面向智能体能力构建的数据集建设宜遵循“多源真实采集夯基、场景化构造组织、合成生成补缺、增强扩充扩边、标注体系赋能、质量评估校验、任务反馈回流、版本迭代升级”的递进逻辑。该逻辑强调数据建设必须与智能体任务目标、行业场景约束和业务运行机制紧密耦合，避免出现数据规模较大但能力支撑不足、样本数量较多但场景适配不足、建设流程完整但反馈闭环缺失等问题。

数据集建设总体技术路线如图 3 所示。



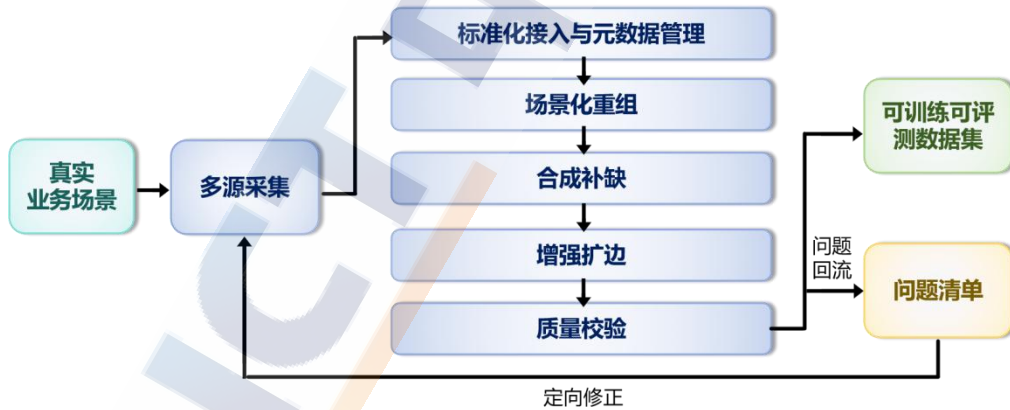
来源：中国信息通信研究院

图 3 面向智能体能力构建的数据集建设技术路线

（三）关键方法

1. 数据采集与生成方法

智能体数据采集与生成应服务于任务学习、场景适应和能力优化，以多源真实数据采集夯实真实性基础，以场景化构造提升任务组织能力，以合成数据生成弥补覆盖缺口，以数据增强与多样化扩充提升分布广度和边界能力。数据采集与生成闭环结构如图 4 所示。



来源：中国信息通信研究院

图 4 数据采集与生成闭环结构图

多源真实数据采集旨在形成可追溯的全过程数据轨迹，应覆盖任务目标、用户输入、环境状态、系统响应、工具调用、执行动作和反馈结果等关键环节。一是围绕典型业务场景开展全链路采集，优先覆

盖高频核心场景、关键决策节点和典型异常情形。**二是**加强多源异构数据协同采集，统筹文本、语音、图像、视频、业务日志、操作记录和接口返回结果等数据类型，并在时间、事件和任务维度建立关联关系。**三是**同步建立标准化采集规范，统一数据格式、字段口径、元数据结构、时间对齐规则和安全要求。**四是**将合规管理前置到采集环节，对敏感信息、个人信息和关键业务信息落实脱敏、授权和审计要求。

场景化数据构造是连接业务需求与能力训练的重要环节，核心是围绕典型任务，将目标、角色、流程、工具、约束和评价机制组织为可复用的任务场景单元。场景化构造应遵循如下流程：从典型业务流程中识别高价值任务，围绕高频、高价值、强约束场景进行数据组织；完整表达场景要素，除任务输入和结果外，还应覆盖环境条件、中间状态、约束变化、异常扰动和反馈闭环；兼顾标准场景、复杂场景和长尾场景，重点补齐多约束冲突、资源受限、流程中断和异常输入等复杂条件；推动场景模板化沉淀，将典型经验抽象为任务模板、流程模板和约束模板，提高跨场景复用能力。

合成数据生成是补齐真实数据不足的重要手段，主要目标是定向补足样本稀缺、长尾分布、高风险和难采集场景的数据缺口。在方法选择上，可采用大模型辅助生成和仿真环境采集两类路径。大模型适用于多轮对话扩展、任务改写、角色模拟、异常样本构造和长尾表达补齐，仿真环境适用于高风险、高成本、难复现和难完整记录场景，可通过参数化控制批量生成状态变化、执行动作、异常事件和反馈结果等样本。同时，应充分关注合成数据可能带来的真实性不足、分布

偏移、偏见放大、场景失真等风险。合成数据应坚持“补充而非替代”的原则，避免过度依赖生成数据导致数据同质化和模型泛化能力下降。对于生成结果，应建立真实性验证、质量评估和人工审核机制，确保其与实际场景保持一致，提高数据集的可靠性和可用性。

数据增强与多样化扩充的目标是提升样本覆盖范围、改善分布结构并增强智能体泛化能力。数据增强应围绕输入表达、任务路径、环境扰动和边界条件开展系统扩展。文本数据可通过表达改写、约束重述、风格变化和歧义构造等方式增强；语音、图像和视频数据可通过噪声扰动、视角变化、背景变化和时序扰动等方式增强；结构化数据和日志数据可通过字段扰动、缺失模拟和异常注入等方式扩充。与此同时，应针对多路径任务、复杂决策任务和长尾场景开展专项扩充，推动数据建设由平均覆盖转向关键能力覆盖。增强策略不宜一次性固定，应结合模型训练结果和场景评测反馈持续优化。

表 7 面向智能体能力构建的数据采集与生成关键方法

方法	建设目标	主要方式	适用重点
多源真实数据采集	夯实真实性基础，沉淀真实任务过程数据	业务系统采集、交互记录汇聚、多模态采集、日志与状态采集	高频核心场景、真实交互流程、关键任务链条
场景化数据构造	提升任务组织能力和场景适配性	任务拆解、流程抽象、场景模板构造、约束建模	典型业务任务、复杂流程场景、多步骤交互任务
合成数据生成	弥补覆盖缺口，扩展长尾与高风险场景	大模型辅助生成、仿真环境采集、规则驱动生成	难采集场景、低频长尾任务、高风险或极端情形
数据增强与多样化扩充	提升覆盖范围和分布多样性	表达变体扩展、模态扰动增强、任务路径扩展、长尾样本补齐	多模态输入场景、复杂决策任务、分布不均衡数据集

来源：中国信息通信研究院

2.标注体系构建方法

（1）面向智能体能力的标注框架设计

标注框架设计应建立“能力域—任务类型—标注对象—标签结构”的映射关系，使标注结果直接服务于训练、评测和优化。总体上，宜采用“基础层、语义层、行为层、评价层”的四层标注架构。基础层重点标注文本边界、图像区域、时间片段等基础信息；语义层重点标注意图、实体、属性和关系；行为层重点标注推理步骤、任务规划、工具调用链和状态转移；评价层重点标注正确性、完整性、合规性、风险性和偏好信息。

在建设模式上，宜采用“通用框架+场景扩展”的设计模式。在通用层统一标注本体、标签主表和编码规范，在行业层扩展专业术语和业务规则，在项目层细化边界样本处理规则和特殊任务要求。

表 8 智能体能力域与标注任务映射重点

能力域	典型标注任务	典型标注对象	标注重点
感知理解	意图识别、实体抽取、槽位填充、多模态对齐	用户意图、实体、属性槽位、图文对应关系	语义边界清晰、标签定义一致
推理规划	推理步骤标注、目标分解、约束识别、路径规划	推理链、中间结论、子任务树、约束条件	过程完整、逻辑合理
工具调用	工具选择、参数映射、调用顺序、异常处理	工具名称、调用参数、调用链、回退策略	依据明确、参数准确、容错可追溯
记忆管理	上下文关联、历史检索、记忆更新	指代关系、历史片段、知识条目	上下文引用准确、记忆边界清晰
多智能体协作	角色分工、消息传递、任务交接、冲突处理	角色定义、交互消息、协作结果	协作过程清晰、责任可追踪
结果评价	正确性标注、偏好排序、风险标注、满意度标注	回答结果、候选答案、评分标签	正确性、可用性、安全性并重

来源：中国信息通信研究院

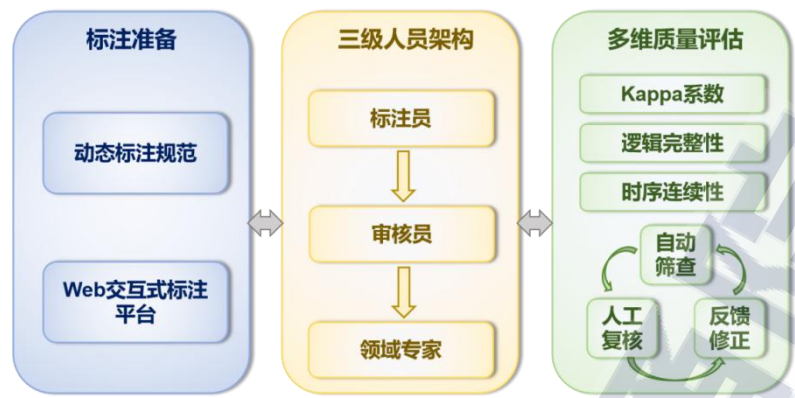
（2）人工标注：规范制定与质量管控

人工标注在复杂语义理解、专业知识判断、边界样本识别和高风险内容审核等场景中仍具有不可替代性。尤其在行业知识密集、判断标准复杂、错误代价较高的任务中，人工标注不仅承担标签赋值工作，更承担规则解释、边界裁定和质量兜底职责。

标注全过程规范方面，建立分级分类的组织机制，合理设置基础标注员、专业标注员、审核员和领域专家等角色。应通过多人独立标注、交叉复核、抽样检查和争议样本升级处理等方式，持续提升标签一致率、逻辑一致率和格式合规率。对涉及敏感信息和高风险数据的任务，还应同步落实脱敏展示、权限控制和操作留痕要求。**标注稳定性方面**，宜建立“培训、试标、校准、抽检、复核、回流”的连续机制。培训阶段应通过正反例和边界案例统一理解口径；试标阶段应发现规范盲区和执行歧义；校准阶段应通过专家点评和集中修订统一判断标准；抽检和复核阶段应识别高频错误和系统性偏差；回流阶段应将争议样本和错误样本沉淀为规范补充材料，推动规则持续完善。**标注质量评价方面**，综合关注边界一致率、逻辑完整性、任务完成度、工时有效性和复核修正率等指标。

（3）人机协同标注：模型辅助与人工审核

人机协同标注典型流程为“模型预标注、置信度评估、分层流转、人工审核、结果汇聚、模型回流”。

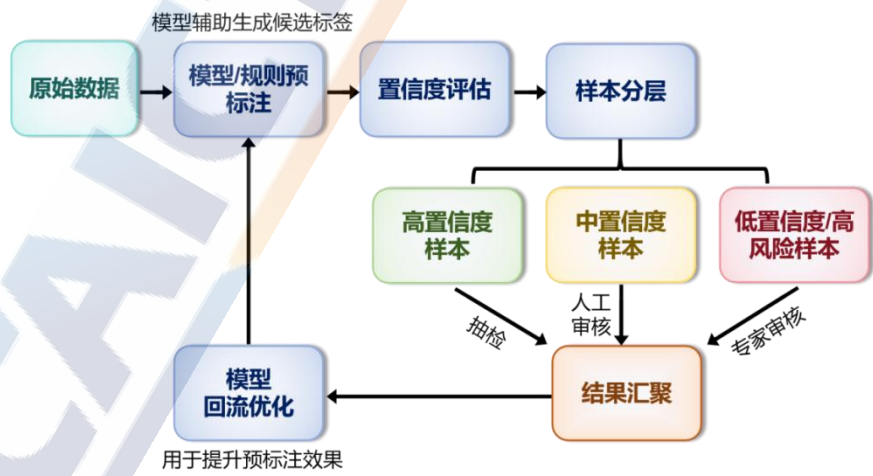


来源：中国信息通信研究院

图 5 人工标注技术路线图

在该模式下，预标注模型先生成候选标签、候选推理过程或候选工具调用链，再依据置信度和风险等级进行分流。高置信度样本宜进入抽检池，中等置信度样本宜由专业标注人员审核修正，低置信度和高风险样本应提交专家复核。重点在于持续提高预标注模型的任务适配能力，并防止人工审核对候选结果形成依赖，确保审核以规范为准。

在人机协同场景下，人工审核必须坚持以标注规范和任务要求为准，通过盲审抽样、多人复核和一致性检查等方式，持续评估模型预标注对人工判断的影响，防止“候选即答案”的隐性问题。



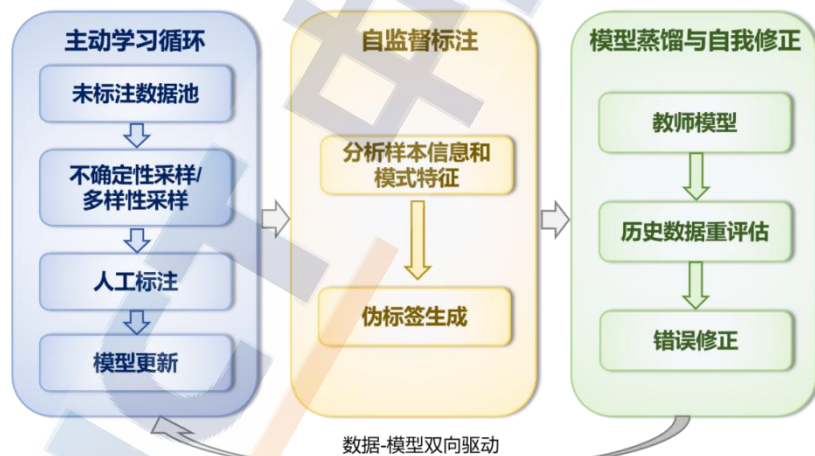
来源：中国信息通信研究院

图 6 人机协同标注流程图

（4）自动化标注：主动学习与持续优化

自动化标注核心是利用规则模板、知识映射、弱监督和主动学习等方式扩大标注规模，同时通过人工抽检和模型迭代保持质量可控，适用于数据规模大、更新频率高且部分任务规则相对稳定的场景。

在自动赋标阶段，可针对规则较稳定、结构相对明确的任务，利用规则模板、领域词库、知识库映射和历史高质量样本训练得到的预标注模型，对原始数据进行批量赋标，生成候选标签或候选结构化结果。对于日志解析、字段抽取、结构化映射、标准问答归类等任务，自动化方式通常具有较高适用性。对于复杂语义任务，则可采用“规则+模型”的混合方式，在提升覆盖度的同时降低纯规则方法的僵化风险。



来源：中国信息通信研究院

图 7 自动化标注技术路线图

（5）针对推理链、工具调用等复杂场景的标注方案

推理链、多步规划、工具调用、检索增强和多智能体协作等复杂任务是智能体能力建设的重点领域，应采用专项标注方案。

推理链标注应重点记录推理步骤、中间结论、依据来源和质量评价，增强求解路径可解释性。工具调用标注应覆盖工具选择依据、参数映射关系、调用顺序、执行结果和异常回退策略，增强调用准确性与容错性。多步任务规划标注应明确总体目标、子任务划分、执行顺序、状态转移和阶段性验收标准。检索增强场景应重点标注检索相关度、证据支撑关系和回答忠实度。偏好优化场景应重点标注候选答案排序、评分标签和偏好原因。多智能体协作场景应重点标注角色职责、消息传递、任务交接和冲突解决过程。

表 9 复杂场景专项标注方案建议

场景类型	关键标注对象	核心标注维度	标注方式建议
推理链标注	推理步骤、中间结论、推理依据、最终答案	步骤拆解、依据引用、逻辑完整性、结论正确性	以人工标注和人机协同为主
工具调用标注	工具名称、参数字段、调用顺序、返回结果、异常回退路径	工具选择依据、参数映射、调用依赖、容错机制	采用规则加模型预标注加人工审核
多步任务规划标注	总体目标、子任务、约束条件、执行路径、状态转移节点	目标分解、任务依赖、执行顺序、阶段校验	以人工标注为主，结合模型辅助
检索增强标注	查询问题、候选片段、证据内容、生成答案	检索相关度、证据支撑关系、答案忠实度	模型辅助召回与人工审核结合
偏好优化标注	多候选回答、排序结果、评分标签、偏好原因	回答优劣比较、帮助性、安全性、风格一致性	以人工排序与打分为主
多智能体协作标注	角色定义、交互消息、分工节点、协作结果	角色职责、信息传递、任务交接、冲突解决	以人工标注为主，结合流程日志抽取

来源：中国信息通信研究院

3.数据质量评估体系

数据质量评估体系是保障智能体数据集可用、可信和可持续演进的重要基础。智能体数据集质量评估不应停留于格式、完整性和准确

性等基础指标，还应重点评估数据对任务理解、逻辑推理、工具调用、多轮交互、异常处置和持续进化等能力的支撑程度。

（1）面向智能体能力的质量评价指标体系

质量评价指标体系宜由通用基础质量指标和智能体能力适配指标共同构成。指标设计应突出场景适配性和能力针对性，不宜仅做一般性质量描述。各行业可结合业务重点对指标权重进行差异化设置，以提高评价结果的指导价值，如表 10 所示。

表 10 面向智能体能力构建的数据质量评价重点

一级维度	评价重点	典型关注点
通用基础质量	说明文档、数据本体、模型应用，与国家标准维度一致	文档完整、结构规范、内容真实、一致干净、安全合规、模型效果
任务描述适配	数据对任务理解与边界识别的支撑	任务清晰度、类型多样性、歧义鲁棒性
认知理解支撑	数据对意图识别、语义理解和知识组织的支撑	意图覆盖、语义充分性、知识结构化、模态对齐
规划决策支撑	数据对任务拆解和策略生成的支撑	任务分解、决策依据、场景泛化、多目标权衡
工具调用与执行支撑	数据对执行闭环和环境交互的支撑	工具链适配、参数正确性、执行逻辑、结果反馈
交互与鲁棒性支撑	数据对多轮交互与异常处置的支撑	交互流畅性、闭环完整性、异常适应性、故障恢复
持续进化支撑	数据对反馈学习和能力演进的支撑	修正样本、经验沉淀、可解释、可审计

来源：中国信息通信研究院

（2）数据质量自动评估方法

在大规模、多类型数据集建设背景下，应建立覆盖数据接入、内容校验、场景评估和结果输出的自动化质量评估方法体系，以提升评估效率和结果一致性。数据接入阶段，应开展格式校验、字段校验、重复检测、噪声清洗和敏感信息筛查。内容评估阶段，应同时覆盖规

范性、完整性、唯一性、准确性和合规性等基础指标，以及任务对齐度、逻辑一致性、事实准确性、交互有效性、泛化适配性和安全鲁棒性等智能体专属指标。对于长链路推理、多轮对话和工具调用数据，可结合规则引擎、统计方法和模型评估开展深度校验。评估结果宜形成整体评分、分项评分、问题定位和整改建议，支持后续分级治理。

从方法路径看，自动化评估宜遵循“前置筛查、深度评估、闭环验证”的递进逻辑。前置筛查负责快速剔除明显不合格数据，重点解决格式错误、重复样本、低信息量样本和敏感内容问题；深度评估负责判断数据是否真正满足任务和能力要求，重点解决逻辑一致性、事实准确性和场景适配性问题；闭环验证则通过实际任务测试反向评估数据质量对智能体表现的支撑效果。

表 11 数据质量自动评估能力分层

能力层级	重点内容	建设要求
基础能力	格式校验、噪声清洗、重复去重、敏感信息筛查、基础评分	应满足
提升能力	任务对齐、逻辑一致性、事实准确性、交互有效性、场景适配	宜满足
高阶能力	多模态统一评估、对抗评估、端到端闭环验证、智能化诊断建议	宜逐步建设

来源：中国信息通信研究院

（3）基于智能体任务表现的数据质量反馈机制

数据质量最终应以任务表现为检验标准。应建立基于智能体任务表现的数据质量反馈机制，将任务执行结果、过程日志和用户反馈反向映射到数据质量问题，形成“任务表现观测—质量问题定位—数据优化迭代—反馈验证”的闭环。

任务表现观测涵盖结果数据与过程数据两大维度。结果数据应采集智能体任务执行相关核心指标，直观反映任务最终输出效果，过程数据则记录任务规划步骤、工具调用逻辑、推理日志及异常触发点等，挖掘隐藏在执行过程中的隐性问题。采集方式采用自动化采集与人工补充相结合，既依托监控系统实现任务日志、输出结果的自动采集，减少人工干预，也通过人工抽样、用户反馈补充主观指标，确保采集数据的全面性与准确性，为后续环节提供可靠支撑。

质量问题定位关键是建立任务表现偏差与数据质量缺陷的精准映射关系。首先对采集到的任务表现数据进行分类梳理，识别各类偏差类型。随后反向追溯根源，将各类偏差与数据集质量缺陷对应。结合智能体核心任务目标，对各类数据质量问题进行优先级排序，优先处理影响任务表现最大、发生频次最高的缺陷。

数据优化迭代针对定位到的不同质量缺陷，制定针对性优化策略。优化过程需注重可追溯、可复现，借助专业工具提升效率，确保每一项优化操作都能精准解决对应质量问题，推动数据集质量迭代升级。

反馈验证闭环采用对照测试与批量验证相结合的方式，将优化后的数据集应用于智能体同类任务。重点对比相关核心指标的变化，若未达到预期目标，则重新分析偏差根源、调整优化方案，直至符合要求。结合验证结果复盘总结，持续完善反馈机制，确保数据质量优化始终与智能体任务需求同频同步，最终实现数据价值最大化。

四、面向智能体能力构建的行业数据集建设路径

（一）数据集建设阶段框架

1.需求摸底与规划设计阶段

需求摸底与规划设计是数据集建设的顶层设计环节，本阶段锚定智能体的闭环能力需求，建立业务场景与数据要素的精准映射关系，确保数据集建设目标与智能体“感知—推理—决策—执行”的核心能力诉求高度对齐，避免数据集与智能体应用脱节。

需求清单落地需紧扣智能体的核心能力诉求。区别于传统 AI 模型，智能体的优势在于自主闭环运行能力，通过输出标准化、精细化需求清单，确保数据精准驱动智能体能力的构建与进化。场景拆解需立足智能体实际应用场景，结合行业特点明确核心业务场景，确保场景拆解贴合实际业务需求，不脱离智能体应用痛点。将各核心业务场景拆解为可执行、可落地的具体任务节点，明确每个任务节点的核心目标与执行流程。数据映射需针对每个任务节点，逐一明确所需数据的核心要求：在数据类型上，区分文本、图像、语音、视频等单模态与多模态数据；在数据规模上，结合任务复杂度与智能体训练需求，明确数据采集的最低量级与覆盖范围，确保数据量满足模型训练要求；在数据质量上，明确数据准确性、完整性、时效性等核心指标。

数据资源摸底旨在厘清现有数据资产状况。应对组织内部业务系统数据（如结构化数据库、非结构化数据存储系统数据、业务系统日志、知识文档等）、外部可获取数据（如行业公共数据、开源数据集等）进行全面盘点，评估其覆盖度、完整性、时效性与可用性。梳理

数据缺口，明确需新增采集或标注的数据范围，并注意规避数据重复建设。摸底过程中应建立数据资产目录，为后续建设规划提供决策依据。

建设规划制定需统筹考虑资源约束与建设目标。基于需求紧迫性、资源可行性及技术成熟度，确定数据集建设的优先级、阶段性目标、资源投入及计划表，形成兼具建设战略前瞻性与技术路线可行性的顶层设计。

2. 基础数据集建设阶段

基础数据集建设是数据集落地的核心环节，核心目标是构建覆盖智能体基础应用需求、具备行业适配性的数据集底座。此阶段遵循“先通用后行业、先单模态后多模态”的分步建设逻辑，兼顾建设效率与资源优化配置，为智能体模型训练、能力提升提供基础数据支撑。

在建设顺序上，坚持“先通用后行业”的建设次序。通用数据集建设是行业数据集的能力底座。优先构建通用基础数据集，涵盖语言理解、逻辑推理、常识知识等基础能力所需的数据资源，满足各类智能体的基础能力训练需求，确保智能体具备基本的交互素养。在此基础上，结合具体行业场景特点，补充行业专属数据，实现数据集与行业智能体应用场景的深度适配，培育智能体的垂直领域专长。

在模态演进上，遵循“先单模态后多模态”的技术路径。先构建文本、图像、语音、视频等单模态数据集，夯实智能体单维度感知与处理能力；再逐步融合多模态数据，构建“文本+图像+语音+视频”的多模态融合数据集，适配智能体多维度感知、跨模态推理的核心需

求，如具身智能体所需的“视觉—力觉—动作”全链路数据。

数据集建设过程中应注重数据质量的全流程管控，构建涵盖准确性、多样性及平衡性等多维质量保障体系，数据质量直接决定智能体的决策可靠性与泛化能力。同时，应重视数据分布的均衡性与偏差治理，避免因训练数据偏差导致智能体产生歧视性决策或场景适应性问题，防止模型幻觉与错误行为强化。鉴于智能体应用环境的动态变化特征，需建立数据的动态更新与维护机制，确保数据集的时效性，避免因知识滞后导致智能体能力退化。

3.标准规范与评测基准建设阶段

标准规范与评测基准是保障数据集质量、推动行业协同的重要基础设施。本阶段旨在建立统一的质量标尺与评估体系，确保数据集建设的规范性与结果的可衡量性。

（1）标注规范落地

标注规范落地是保障数据集质量、推动数据标准化、实现智能体模型精准训练的核心保障。标注规范应明确标注对象定义、标注类型体系、标注操作流程、质量判定标准等核心要素。针对智能体应用特点，标注规范还需涵盖意图识别、槽位填充、对话状态、动作序列等特殊标注类型的详细说明。

（2）测试数据集构建

测试数据集构建需兼顾覆盖度与代表性。测试数据集应覆盖智能体应用的典型场景、常见任务与边界情况，能够有效评估智能体在不同条件下的表现。测试数据应与训练数据保持独立，避免数据泄露导

致的评估偏差。

（3）质量基准确立

质量基准确立为数据集质量评估提供参照标准。质量基准应涵盖数据完整性、准确性、一致性、时效性等多个维度，并设定量化的评价指标与阈值标准。参考国家标准、行业标准或国际通行实践，结合行业特点制定适用的质量基准体系。

（4）评测体系搭建

评测体系搭建需建立科学、公正、可复现的评估机制。评测体系应包含评测指标设计、评测方法选择、评测流程规范、评测结果呈现等组成部分。围绕智能体在特定场景下的任务完成度、决策合理性、交互流畅性等核心能力，构建包含自动化客观指标与人工主观评价相结合的评测框架，实现从数据质量到模型能力的全链路评估。

4. 开放运营与持续迭代阶段

开放运营与持续迭代是维持数据集生命力与价值增长的长期机制，旨在构建数据应用的飞轮效应，驱动数据集质量的螺旋式上升。

（1）数据集发布

数据集发布应选择适宜的开放方式与渠道。根据数据敏感程度与知识产权要求，可采取完全开放、有条件开放、会员共享等不同模式。发布时应配套完整的数据说明文档，包括数据集概述、数据字段说明、采集与处理方法、使用许可协议等信息。建立数据集版本管理机制，清晰记录各版本的变更内容与更新时间。

（2）应用反馈闭环

应用反馈闭环是数据集持续优化的重要驱动力。建立畅通的反馈渠道，收集数据使用者在应用过程中发现的问题与建议，包括数据错误、标注偏差、覆盖不足、格式兼容等问题。反馈信息经分析处理后，纳入数据集迭代优化计划，形成“应用—反馈—优化”的良性循环。

（3）飞轮机制运转

飞轮机制运转推动数据集生态自我演进。飞轮机制运转的核心是实现业务价值与技术能力的双向赋能，推动数据集生态与智能体应用生态的协同演进，打破技术与业务脱节的困境。

从业务层面看，迭代后的高质量数据集反哺至智能体训练流程，持续提升智能体的推理精度、场景适配能力与执行效率，让智能体能够更好地完成行业业务任务，如金融智能体降低风控误判率、工业具身智能体提升设备操作效率，进而提升用户体验与业务价值，吸引更多行业主体接入智能体应用。

从技术层面看，更多行业主体的接入会产生大量高价值的真实业务场景数据，这些数据经清洗、标注后补充至数据集，进一步丰富数据集的场景覆盖度与数据质量，形成“业务场景拓展—高价值数据积累—数据集优化—智能体能力提升—业务价值升级”的飞轮效应，推动数据集与智能体的持续迭代。还可探索数据共享与交易机制，在保障安全的前提下，促进数据要素在产业链内的有序流动，最大化数据集的社会与经济价值。

（二）数据集建设的主体模式与协同机制

面向智能体能力构建的行业数据集建设涉及政府、企业、行业组织、科研机构等多方主体。不同主体在资源禀赋与核心诉求上存在差异，需结合行业特点选择合适的建设模式，建立有效协同机制，破解数据孤岛、资源分散等问题，推动数据集共建共享。

1. 企业主导模式

企业主导模式适用于业务场景明确、数据资源集中、商业化目标清晰的场景。该模式下，企业作为建设主体自主完成数据集的规划设计、采集标注、质量管控与运营维护全流程，拥有数据集完整的控制权与收益权，可根据业务需求灵活调整建设重点，实现数据资产的敏捷迭代。核心优势在于建设目标与业务目标高度契合，企业对业务场景理解深刻，能够精准把握数据需求，确保数据集的实用性与针对性。该模式要求企业具备较强的数据技术能力与持续投入能力，适用于涉及敏感数据或商业机密的数据集建设场景。

2. 行业联盟模式

行业联盟模式适用于数据分散、需求共性强、协同意愿高的行业领域。该模式下，多个主体以联盟形式共同参与数据集建设，按约定贡献数据资源或标注能力，共享建设成果。联盟组织负责制定建设规划、协调成员关系、管理数据资产，确保合作有序推进。核心优势在于能够突破单一主体的数据瓶颈，汇聚行业各方资源与专业知识，构建规模更大、覆盖更广的数据集，同时有助于形成行业统一的数据标准与评测规范。该模式的关键在于建立有效的利益分配机制，明确各

方权利义务，配套数据贡献评价规则与争议解决机制，建议优先在行业组织成熟、企业合作基础良好的领域开展探索。

3.公共平台模式：政府驱动与开放共享

公共平台模式适用于基础性强、公益属性明显的领域。该模式下，政府或公共机构作为主导力量，投入公共资源建设数据集并面向社会开放共享，旨在降低行业整体数据获取门槛，促进技术创新与应用普及。核心优势在于能够解决市场失灵问题——部分基础性数据集建设投入大、周期长、回报不确定，企业缺乏投资动力，政府主导可弥补这一不足，提供行业共用的数据基础设施。有效实施需政府持续投入并建立科学的平台运营机制，平衡开放共享与安全管控，建议优先在医疗健康、教育教学、城市治理等公共服务领域推进。

4.多主体协同的数据生态构建路径

单一模式难以满足智能体各类能力对数据多样性、规模性与专业性的复合需求，需推动多主体协同发展，构建健康可持续的数据生态。在主体分工上，企业聚焦专用数据集建设，支撑自身智能体研发；行业联盟整合行业资源，建设共性数据集；公共平台提供基础数据集，支撑通用能力构建；科研机构聚焦核心技术攻关，推动数据集质量提升；开源社区发挥聚合优势，提供低成本可复用的开源数据集，推动多源数据融合适配。在流通机制上，需在合规前提下推动公共数据向企业开放、企业数据向联盟共享、联盟数据向行业扩散，通过建立数据交易平台或数据共享空间，综合运用隐私计算、联邦学习等技术手段，实现数据价值的跨主体安全流转，同步建立数据版权保护机制，

保障各方合法权益。

（三）数据合规、安全与伦理治理

面向智能体能力构建的行业数据集建设，需坚守“安全底线、合规红线、伦理高线”的基本原则。依据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规要求，以及《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》《人工智能科技伦理审查与服务办法（试行）》等政策文件和相关规定，构建“合规采集、安全存储、规范使用、全程管控”的安全合规治理体系，防范数据安全风险，保障数据主体合法权益，推动数据集建设与应用的健康有序发展。

1. 治理框架

构建“管理层引领、技术层支撑、运营层落地”的三维治理框架。管理层明确数据治理的决策主体与职责边界。建立数据安全与合规审查机制，负责数据集建设的风险评估、分级审批与争议处置。将合规要求纳入数据集立项、采购、验收等关键节点的决策流程，确保法律合规审查前置。技术层部署覆盖数据全生命周期的安全与合规技术能力。包括数据加密、访问控制、隐私计算等安全防护技术，以及数据溯源、内容安全检测、偏见分析等合规保障工具。通过技术手段将安全策略与合规要求固化为可自动执行的规则。运营层建立覆盖数据采集、清洗、标注、存储、使用、销毁全流程的标准化操作规范。明确各环节的合规操作要点与安全控制措施。制定数据安全事件与合规风险的应急预案，建立监测预警与响应处置机制。



来源：中国信息通信研究院

图 8 综合治理框架

2.数据安全体系建设

（1）全流程安全防护

在**数据采集与接入阶段**，建立严格的数据源准入审查机制，对第三方数据、开源数据及合成数据统一开展来源验证、权属核验与安全检测，构建前置风险筛查流程，严防恶意数据、污染数据流入模型训练环节。**数据传输与存储阶段**，综合运用加密技术保障数据全链路机密性与完整性，对高敏感类别数据集实施隔离存储机制，并配套严格身份认证、最小权限访问控制与操作全程留痕管理，筑牢非授权访问与数据泄露防护屏障。**数据使用与计算阶段**，依托可信计算、联邦学习等隐私增强技术，降低数据在训练与推理计算过程中的明文暴露风险，同时建立模型推理接口对外输出监测机制，强化异常行为告警与管控，防范敏感信息通过模型输出结果被逆向还原窃取。

（2）隐私保护机制

隐私保护是数据集建设的基本要求，需综合运用管理措施与技术手段实现有效保护。

数据脱敏是隐私保护的基础技术手段。针对直接标识符（如姓名、身份证号、手机号等），需采用掩码、替换、泛化等方法进行脱敏处理，使数据主体不可识别。针对准标识符（如年龄、地址、职业等组合信息），需评估重识别风险，采用 k-匿名、l-多样性、t-接近性等模型降低识别概率。同步建立脱敏规则库与脱敏效果评估机制，确保脱敏处理的有效性与一致性。

隐私计算技术可在数据利用与隐私保护之间取得平衡。联邦学习技术支持在不共享原始数据的前提下进行模型训练，适用于多方协同建设场景；差分隐私技术通过添加可控噪声保护个体隐私，适用于统计分析类数据集；可信执行环境技术提供隔离的安全计算空间，适用于敏感数据处理场景。

数据分级分类是隐私保护的精细化管理基础。按照数据敏感程度将数据划分为核心数据、重要数据、一般数据，针对不同等级制定差异化的保护措施。同时按照数据类别进行分类管理，明确各类数据的采集标准、存储要求、使用规则与销毁流程，实现精准管控。

3.数据合规体系建设

以国家法律法规、监管政策及行业规范为根本遵循，构建覆盖数据集全生命周期、过程可追溯、责任可认定、结果可审计的数据合规治理体系，实现数据集来源合法、使用规范、流转可控，系统性防范合规风险，为智能体应用安全落地提供坚实法治保障。

（1）合规制度与标准体系

立足智能体数据集建设业务场景，对标《中华人民共和国数据安全

全法》《中华人民共和国个人信息保护法》等国家法律规制，紧跟行业监管要求及技术标准规范，建立健全内部合规制度与标准体系。明确数据采集授权、知识产权归属、使用边限定、跨境数据流动等关键环节的合规准则，厘清各主体权责边界，编制统一规范的数据合规操作指引与审查核验清单，推动合规要求全面嵌入数据采集、处理、标注、存储、使用、流转全业务流程，实现合规管理标准化、体系化、常态化，确保各项合规要求落地生根、可执行、可监督。

（2）全流程合规审查机制

建立事前准入审查、事中过程管控、事后监督审计的全流程闭环合规管控机制，实现合规风险的前置防范、事中管控与事后追溯。

在数据采集接入阶段，严格开展数据源资质审核、授权文件核验与采集行为合法性校验，重点核查数据来源的合规性、权属的明确性，坚决阻断非法及不合规数据进入数据集建设流程。**在数据处理、标注、聚合及应用阶段**，强化全过程合规校验与行为管控，建立动态监测机制，严防超范围、超授权使用数据，及时纠正不合规操作行为。每个阶段实施全链路操作留痕与存证管理，确保各环节行为可追溯、责任可界定、问题可倒查，为合规审计与责任追究提供坚实支撑。针对重点行业领域的差异化监管要求与业务场景特征，立足各行业数据管理规范与监管底线，建立场景化、精细化合规适配规则。将各行业特殊监管要求、数据安全标准与数据集建设全流程深度融合，明确不同行业数据集的合规重点、管控措施与审查标准。

4. 伦理治理体系建设

（1）基本原则

确立智能体数据集建设伦理要求：**增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险、保持公开透明、保护隐私安全、确保可控可信。**

（2）偏见识别与纠偏机制

聚焦智能体训练数据可能存在的样本偏差、分布失衡、标签歧视等伦理隐患，建立系统化、常态化的数据偏见识别、评估与纠偏机制。构建多维度数据偏见检测指标体系，依托技术工具对训练数据的样本覆盖范围、分布比例、标签设置等进行全面排查，精准识别性别、地域、职业、群体等各类潜在偏见问题，确保数据集在公平、包容、可信的轨道上持续演进。针对检测发现的偏见隐患，制定差异化纠偏方案，通过补充样本、优化数据分布、规范标注规则等方式，提升数据集的代表性与均衡性，避免因数据偏差导致智能体产生歧视性决策或输出。同时，建立标注规则伦理校验机制，组织专业团队对标注标准、标签内容进行伦理审核，坚决杜绝有害、违规标签植入数据集，从源头遏制偏见传递。

（3）伦理审查与责任机制

建立健全规范化、刚性化的伦理审查与责任追究机制，确保伦理治理要求贯穿相关人工智能科研活动全过程。建立常态化伦理审查机制，明确审查主体、审查流程、审查标准，依法依规对属于科技伦理审查适用范围的人工智能科研活动开展伦理审查，并将数据集采集、处理、标注和应用等情况作为审查的重要内容，重点核查数据伦理合

规性与潜在风险。坚持建设单位履行人工智能科技伦理审查主体责任，健全内部伦理审查机制，对不具备审查能力的建设单位，可依托具备相应条件的人工智能科技伦理审查服务机构开展审查，提升伦理治理的专业化、规范化水平。同时建立伦理风险问责机制，明确数据集建设、使用、管理各主体的伦理责任，对违背伦理原则、造成不良社会影响或损害公共利益的行为，实行责任追溯与问责。

五、发展展望和建议

（一）行业数据集发展趋势展望

围绕智能体在行业场景中的深度应用，高质量数据集建设呈现三大发展趋势：**多模态融合成为主流、数据飞轮推动持续进化、市场化流通加速。**

1. 多模态与跨模态数据集成为主流

随着人工智能技术持续向科研场景深入，高质量数据集建设正由以文本、表格和结构化元数据为主的单一形态，逐步转向图像、曲线、表格、公式、模型、文献和实验过程信息等多模态数据的综合组织。单一模态数据已难以支撑复杂场景交互、深度推理与动态规划任务。未来面向智能体的数据集将以多模态对齐、跨模态关联、全场景覆盖为核心特征，具备更强的时空一致性、语义完整性与场景适配性。

一是跨模态语义关联将成为数据集建设重点。未来科研数据集不再只是不同类型数据的简单汇集，而将更加注重图像、文本、物理量、实验参数和结论表述之间的语义映射关系，使智能体能够在不同表达形式之间建立关联，提升其对复杂科研问题的综合分析能力。**二是实**

验过程类数据的重要性将进一步提升。除论文、报告、图表和结构化结果外，实验记录、设备状态、异常处置和操作经验等过程性信息也将纳入数据集建设范围，有助于智能体更好地理解科研活动中的动态过程，为实验支持、过程优化和风险预警等应用提供基础。

2. 数据飞轮机制推动智能体持续进化

未来高质量数据集建设将不再局限于静态归档和一次性整理，而是逐步转向“应用牵引、反馈驱动、持续优化”的动态演进模式。通过应用侧采集的用户交互、任务执行、异常反馈、环境感知等数据，经合规治理、清洗加工与高质量标注后回流至数据集体系，形成“场景采集—治理标注—训练优化—部署应用—反馈迭代”的完整闭环。

一是长尾科研问题将成为高价值增量数据的重要来源。科研活动中的关键问题往往并非标准化、重复性高的常规问题，而更多表现为边界条件复杂、学科交叉明显、依赖经验判断的长尾问题。未来应通过用户交互记录、失败案例回收、人工复核与场景化标注等方式，持续补充此类高价值样本，提高智能体对复杂科研任务的适应能力。**二是记忆与上下文管理将推动数据集由“信息集合”向“经验集合”演进。**随着智能体上下文组织、任务连续性管理和多轮交互能力的增强，科研数据集的价值将不再局限于知识存储，还将体现在对历史经验、决策依据和问题处理路径的持续沉淀。

3. 数据要素市场化流通提速

随着科研数字化基础设施不断完善以及数据要素市场化配置改革持续深化，高质量数据集建设的重点将逐步从“资源积累”延伸到

“标准组织、可信流通和持续服务”。数据集作为数据价值化的核心产品形态，其确权、估值、交易、交付、监管体系加速完善。公共数据分类分级有序开放、行业数据合规共享、企业数据市场化交易协同推进，推动行业数据集供给从分散自建、单点自用，转向标准化生产、规模化供给、产业化复用。

一是标准化建设将成为提升数据复用能力的关键环节。未来科研数据集建设需要在元数据描述、知识组织方式、接口规范、质量评价和安全管理等方面形成更统一的规则，以提高不同机构、不同平台、不同场景之间的数据兼容性和可复用性。**二是高价值科研数据将逐步走向服务化应用。**经过处理的数据资源，有望从原始资料转化为可调用、可评估、可持续维护的知识服务能力。数据集的价值也将逐步从“存起来”转向“用起来”，更直接服务于科研支持、知识发现和技术转化，未来行业高质量数据集发展方向如表 12 所示。

表 12 未来行业高质量数据集发展方向

维度	传统数据集特征	未来高质量数据集方向	对科研场景的意义
数据形态	以文本、表格、基础元数据为主	融合图像、曲线、公式、模型、文献等多模态信息	提升复杂科研对象的表达能力
更新机制	离线整理、阶段性更新	应用反馈驱动、持续迭代优化	增强对真实科研场景的适应性
应用方式	偏存储、归档、检索	面向问答、推荐、分析辅助与过程支持	提升智能体实际可用性
组织方式	分散建设、标准不一	结构化治理、统一描述、可控共享	提高跨平台复用能力

来源：中国信息通信研究院

（二）推动行业高质量数据集发展的建议

围绕智能体能力构建的实际需求，从政策、标准、行业、企业四

个层面提出以下建议：

1.政策层面：完善数据要素政策，加大行业数据集建设支持

政府部门应进一步发挥顶层设计和政策引导作用，围绕数据要素的确权、流通、分配、应用和安全治理，持续完善制度体系，为科研领域高质量数据集建设提供稳定的政策环境和资源保障。

一是加大专项资金支持和试点示范布局力度。建议结合国家有关数据科技创新部署，持续加大对高质量数据集相关技术研发和应用示范的支持力度，重点支持面向科研活动、重大科研平台和行业关键场景的数据集构建、治理、评测及应用验证工作，推动形成一批可复制、可推广的实践案例。**二是推动“数据要素×”行动在科研领域落地实施。**建议进一步细化面向科研机构、高校和企业的实施路径，鼓励在安全合规前提下开放共享非涉密、高价值科研数据资源，探索产学研用协同参与的数据开发利用和成果转化机制，并将高质量数据集建设纳入相关科技计划和创新体系布局，提升其对科研创新和产业发展的基础支撑作用。**三是强化政策协同与多元化支持工具。**立足智能体能力构建与数据要素发展全局，将行业高质量数据集建设纳入人工智能与数据产业重点布局方向，推动国家与地方政策协同，面向电信、交通、能源、医疗、工业等重点领域，采用财政奖补、数据券、专项课题等多元化工具，降低机构数据集建设与使用成本；支持数据要素先行区开展试点示范，探索智能体数据集定向供给、安全流通、价值评估的创新路径，形成可复制、可推广的实施模式。

2.标准层面：加快建立面向智能体的数据集标准规范体系

标准先行是提升高质量数据集可复用性、可扩展性和跨平台兼容性的重要基础，应加快推动覆盖数据采集、组织、治理、接入、评测和安全管理等环节的标准规范建设。当前，人工智能相关标准化工作已形成较为系统的总体框架，也为后续面向数据集和智能体能力构建的细化标准研制提供了基础。

一是推进关键标准和技术规范研制。依托相关标准化组织和专业机构，围绕数据格式、元数据描述、知识库接入、接口调用、跨模态关联和质量控制等关键环节，加快形成面向智能体能力构建的数据集标准和技术规范，增强不同平台、不同机构之间的数据互操作能力。重点推进多模态采集、细粒度标注、质量评估、分类分级、安全合规、接口格式、流通交易等关键标准研制。**二是完善质量评价体系。**在传统的数据完整性、准确性、一致性指标之外，进一步纳入内容时效性、标注质量、场景适配性、答案可信度和检索有效性等维度，并结合 RAG 场景中常用的 Context Relevance、Faithfulness、Answer Relevance 等评估思路，提升数据集“可用、可信、可落地”的综合水平。建议建立国家级数据集质量评估与认证机制，推动标准在研发、生产、交易、应用全流程落地，提升跨行业、跨平台数据复用与互操作能力，减少重复建设与资源浪费。

3.行业层面：推动数据联盟与开放共享机制

跨机构、跨领域协同的数据生态，是缓解科研数据分散、标准不

一和重复建设等问题的重要途径。应围绕高质量数据集建设需求，逐步形成多方参与、分工协作、共建共享的行业推进机制。

一是推动建立行业协作机制或数据联盟。建议由行业协会、科研平台或相关牵头单位组织高校、科研院所、重大科研设施及产业链相关主体，共同推进基础语料、标注规范、质量要求和评测任务等方面的协同建设，通过资源共享和规则共建，降低长尾数据采集与整理成本，提升数据集建设的整体效率。以行业龙头企业、科研机构、新型研发平台为核心，组建智能体数据集产业协同组织，搭建共性数据资源池与开放共享平台。**二是探索开放科学框架下的共享利用模式。**在符合数据安全、知识产权和科研伦理要求的前提下，逐步推动科研数据、方法经验和案例资源的分级开放与规范共享，探索更加透明、可追溯、可复核的数据利用机制，并结合隐私计算、联邦学习等技术手段，提升高价值科研数据在跨机构场景中的合规利用能力。推动公共数据面向智能体研发定向开放与授权使用，提升公共数据资源供给效率；建立行业数据共享治理规则，明确数据使用边界、利益分配与责任分担机制，鼓励企业开放脱敏后的非敏感数据、标注样本与基准测试集，共建行业知识库、工具链与测试体系，破解数据孤岛、供给不足、质量参差不齐等痛点，支撑全产业协同创新。

4.企业层面：建立数据工程核心能力，构建数据竞争壁垒

科研机构、科研平台及相关建设单位应将高质量数据集建设纳入长期能力布局，推动数据工作由阶段性任务逐步转向体系化、持续性

建设，为智能体能力构建和科研知识服务提供稳定支撑。

一是做好体系规划，明确建设目标与场景边界。应结合自身科研方向、业务流程和应用需求，系统梳理数据集建设所需的数据范围、内容边界、组织方式和质量要求，明确面向知识服务、科研辅助、实验支持等不同场景的重点任务，围绕智能体能力构建需求，统筹规划知识数据、任务数据、工具数据、交互数据和评测数据等资源体系，形成与应用场景相匹配的数据资产布局。**二是加强工程化建设，完善数据治理与组织能力。**围绕数据采集、清洗转换、知识抽取、结构化加工、质量控制、版本管理和合规审查等关键环节，逐步建立完善的数据工程体系。重点强化非结构化数据解析、多模态数据处理、知识图谱构建、RAG 知识库加工、Agent 轨迹生成与标注、数据合成增强等能力建设，通过对技术文档、实验记录、科研论文和专家经验等资源的持续整理与组织，推动分散知识向可维护、可调用、可复用的数据资产转化。加大数据集自动化生产工具、质量管控平台、安全合规系统研发投入，提升数据生产效率与质量稳定性。**三是建立持续监测与迭代机制。**面向数据建设和智能体能力构建全过程，建立覆盖需求分析、数据采集、预处理、标注校验、应用评测和反馈优化等环节的质量监测机制。结合自动化评测工具和人工复核方式，对知识检索效果、工具调用表现、任务完成质量和用户反馈情况进行动态评估，形成“数据构建—模型优化—应用验证—反馈迭代”的持续改进闭环。聚焦垂直领域深耕高壁垒、高价值、高适配的专属数据集，发展数据即服务（DaaS）、数据集即产品（Dataset as a Product）等新模式，以

持续积累的数据资产和数据工程能力构筑智能体应用长期竞争优势。



附件 行业智能体高质量数据集建设代表性实践

（一）企业运维领域：面向企业内部 IT 运维智能体的高质量数据集建设

（1）数据集建设背景

企业 IT 系统架构和业务链路日趋复杂，运维数据分布在服务管理、配置管理、监报告警、日志分析、知识平台和运维数据湖等系统中，涵盖工单、配置项、指标、告警、日志、拓扑、知识文档及工具调用记录等多种类型。传统数据治理能够保障基础质量，但业务术语、指标口径、对象关系和运行上下文仍较多依赖专家经验，导致智能体难以准确识别告警影响对象、关联配置与拓扑、追溯分析依据并调用合适工具。为支撑智能体规模化应用，**联想（北京）有限公司**围绕“数据可理解、知识可检索、关系可推理、工具可调用、过程可追溯、效果可反馈”，建设面向智能运维任务的高质量数据集与人工智能就绪数据工具链。

（2）场景与需求

一是监控智能体需要融合指标、日志、告警、工单、配置、拓扑、历史故障案例和工具接口，支撑异常识别、告警归并、根因定位、影响评估与修复。**二是故障诊断智能体**需要建立应用系统、配置项、监控指标、告警、事件单、根本原因和解决方案之间的语义关系，为跨域关联分析提供推理依据。**三是用户服务智能体**需要贯通用户咨询、智能体对话、服务请求、知识命中、转人工和满意度反馈，支撑问题识别、服务推荐及运营优化。**四是管理洞察智能体**需要统一服务水平、

工单效率、系统稳定性和用户体验等 KPI 指标口径与数据来源。**五是工程师智能体**需要将供应链订单流转、任务状态和异常告警映射到监控对象，为业务影响分析、风险识别和辅助决策提供实时证据。**六是守护智能体**需要采集 Agent 会话、踪迹、模型及工具调用、响应耗时、词元消耗、错误、质量评价、风险样本和人工反馈，对智能体运行质量、安全合规、资源消耗和执行风险进行持续观测与治理。

（3）建设路径与做法

一是建设统一数据治理底座。对监控、日志、配置、事件、知识等多源数据进行接入、清洗、去重、标准化和语义对齐，自动采集数据库、表、字段及数据处理任务等技术元数据，并通过业务域、术语、指标、标签、责任人和数据血缘记录数据含义、计算口径、责任归属及上下游链路。**二是构建运维语义层与本体关系骨架。**将底层表和字段映射为标准运维术语、业务指标和计算规则，并与实际字段、查询逻辑及血缘绑定；围绕应用系统、业务服务、配置项、接口、指标、告警、事件、责任组、根本原因和解决方案建立本体模型，定义部署、依赖、调用、监测、触发、影响和处置等关系，使原始数据转化为智能体可理解、检索和推理的语义数据。**三是建设场景化任务数据集。**围绕工程师问题处理流程，形成“业务问题、运行状态、监控证据、异常原因、解决方案、业务影响”六类关联数据，并通过标准化智能体间交互协议统一调用意图、业务上下文、执行动作和验证结果；围绕智能体守护，融合踪迹、会话和 Span 数据，沉淀异常样本、风险样本及干预案例，形成覆盖规划、执行、反馈的闭环数据。**四是打通**

智能体消费与反馈通道。通过模型上下文协议检索元数据、术语、指标和规则，通过图谱增强检索关联本体关系，并将知识库、检索增强生成、技能、工具及权限治理纳入统一能力底座。智能体执行结果、质量评价和人工反馈持续回流，用于优化知识、规则、样本和服务策略。

（4）应用成效与典型成果

一是监控智能体基于统一语义模型自动关联指标、日志、告警、拓扑和历史案例，开展异常检测与根因定位并形成证据链，分析结果准确性超过 95%，故障修复平均修复时间提升 60%。**二是基于本体的故障诊断智能体**沿“故障现象、影响对象、关联告警、可能原因、处置方案”路径开展分析，已支持工厂现场 IT 人员日常单据处理，复杂任务处理效率提升 80%。**三是用户服务智能体**数据闭环能够识别高频问题、未解决问题、低满意度会话、知识缺口和转人工原因，推动知识优化、意图调优、服务推荐及运营策略迭代。**四是管理洞察智能体**依托统一的运维指标口径和可信数据来源，提供服务水平、工单效率、系统稳定性和用户体验等多维分析，为管理决策提供可信数据支撑。**五是工程师智能体**在供应链实时追踪场景中自动整合监控信息，生成业务运行分析和影响评估结论，预测准确性超过 95%，已拓展至六大供应链核心业务群、覆盖 400 余名业务用户。**六是守护智能体**持续监测模型调用、知识检索、工具执行和输出质量，覆盖 15 个审查指标，审查准确性超过 95%，风险预警准确率超过 98%。目前，联想（北京）有限公司已初步形成“技术元数据采集、运维语义维护、本

体关系建模、场景数据构建、智能体标准调用、运行反馈优化”的工具链，为多类智能运维智能体提供可治理、可追溯、可关联和可调用的数据基础。

（二）医疗健康领域：专病、临床辅助决策、血管智能评估智能体高质量数据集建设

（1）数据集建设背景

医疗健康数据长期存在来源分散、标准不一、质量参差、难以复用等问题，制约了医疗人工智能应用落地。随着医疗 AI 深化应用，行业对数据集的匹配度、可计算性、可追溯性和持续更新能力提出了更高要求。**东软集团**围绕诊前、诊中、诊后全周期形成统一数据底座，服务科研发现、辅助诊疗、患者管理和数据要素价值释放，融合病例样本、临床指南、人工知识等多源多模态数据，支撑疾病理解、证据检索和诊断建议生成，同时建设大规模、标准化、多中心分布的医学影像标注数据及多模态关联数据，提升 AI 模型的临床可用性。

（2）场景与需求

一是专病管理场景聚焦科研、诊疗和管理三类需求。科研侧需要快速定位患者、提取变量，支持疗效评价和循证医学研究；诊疗侧需要整合院内外临床、检验、影像、随访等多模态数据，支撑风险筛查、个性化精准决策、分组精细化管理及智能化患者服务；管理侧需要将“活数据”沉淀为可持续更新、可复用、可交易的数据产品。**二是临床决策场景**聚焦患者病情理解与诊断推理。其核心需求是将非结构化病历摘要、疾病知识、指南资料和人工整理知识转化为可训练、可检

索、可引用的数据体系，解决原始材料结构差异大、知识可用性不足、专业性和可追溯性要求高等问题。**三是血管智能评估场景**聚焦全身动脉 CTA 影像分析，覆盖血管分割、斑块识别、血管分段和狭窄检测四类任务，核心需求是构建符合医学共识的标准化标注体系，覆盖不同部位、病变类型、狭窄程度及阳性与阴性样本。

（3）建设路径与做法

一是围绕专科专病的深入需求，依托洞察平台汇聚 HIS、EMR、LIS、PACS 等系统数据，完成数据清洗、治理、标准化和结构化建模，构建专病数据库和专病场景化高质量数据集。同时，结合大模型和知识图谱融合的记忆体技术，建设专病知识库、特征提取模型和智能体服务，将专病数据集、知识库和模型能力编排并转化为医生侧、患者侧和管理侧应用。**二是围绕智慧医院和智慧服务的共性需求**，构建病例样本数据、临床知识数据和人工整理知识三类数据体系。病例样本以公开病例摘要为基础，经字段级语义对齐、病历化重构和约束式合成增强，提升模型对主诉、症状、检查结果和疾病概念的理解能力；临床知识以 ICD-10 疾病术语为锚点，通过检索、抓取、抽取、去噪、归一化和知识块切分，形成可检索、可引用的 RAG 增强数据；人工知识围绕高风险场景补充疾病定义、诊断关注点、鉴别诊断和危险信号识别等内容，强化关键场景的边界控制。**三是围绕心脑血管和全身血管影像分析需求**，汇集 15000 例阳性和 5000 例阴性 CTA 数据，制定统一标注规范和标签体系，依托平台集成的 150 余种预标注算法，采用“AI 预标注+人工精修+三级质控”模式，完成全量标注，标注

效率较纯手工提升 150%，并通过专家审核、第三方盲审和一致性评估保障数据质量。

（4）应用成效与典型成果

一是专病智能体方面，通过数据集支撑开发覆盖儿科、精神心理和脑重大疾病“筛防诊治康”全过程的智能体；支撑建设胃癌、肺癌、肠癌等专病数据库，服务临床研究并孵化智能应用，建设 34 个专病全周期管理平台，形成数据驱动的全病程智能管理创新服务模式。**二是临床辅助决策智能体方面**，通过病例与知识协同建设，系统在临床理解、知识补充、证据追溯和专业稳定性方面显著提升。目前覆盖 10549 种疾病，Top5 诊断准确率达到 91.54%。**三是血管影像智能评估智能体方面**，打造符合医学共识、适配算法训练的高质量全身动脉 CTA 标注数据集；显著提升 Vasis 智能体在血管分割、斑块识别、分段及狭窄检测四大场景的模型性能，大幅提升影像诊断效率与精准度，支撑多学科精准诊疗落地，助力血管精准诊疗智能化、规范化发展。

（三）通信领域：网络运维智能体高质量数据集建设

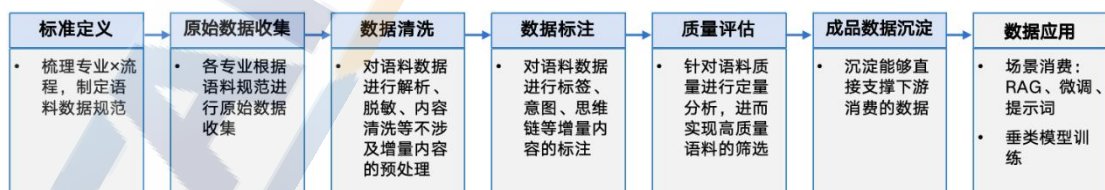
（1）数据集建设背景

当前，通信行业的语料建设存在四大核心困境。**一是语料资源孤岛化**，不同地区、不同系统、不同厂商的语料数据各自为战，无法实现共享复用；**二是数据建设成本高**，构建高质量语料数据集需要投入大量人力物力进行标注和梳理，高昂的成本限制了语料库的扩展和更新；**三是语料标准碎片化**，行业内缺乏统一的语料建设标准，数据格

式、标注方法五花八门，无法高效整合协同；四是语料质量参差不齐，专业术语不统一、模糊表述频现，导致 AI 模型训练效率低、识别准确率不足。基于网络运维各阶段的痛点问题，浙江移动部署实施自智网络“点金行动”，以“流程、数据、AI”融合为驱动，推动运维模式升级，打造 30+ 专业智能体，提升网络质量与业务韧性。

（2）建设路径与做法

标准定义方面，明确数据标准，包括数据收集的业务范围（通信运营商网络运维领域各生产管理流程）和类型范围，制定通信行业网络运维领域语料规范。**原始数据收集方面**，制定收集机制与收集模板，从各专业领域全面收集原始语料。数据清洗方面，获取原始语料后，进行解析、脱敏、内容清洗等预处理操作。**数据标注方面**，采用人工+自动结合的方式，对数据进行标签、意图、思维链等增量内容标注，为大模型应用提供关键支撑信息。**质量评估方面**，对处理后的数据进行定量分析评估，筛选出高质量数据语料予以保留，低质量数据则进行搁置或进一步迭代优化。**成品数据沉淀方面**，形成可直接支撑下游应用（如需求场景智能体建设）的成品数据集，满足 RAG、微调、提示词等不同应用场景对数据的需求。



（3）应用成效与典型成果

问答智能体结合行业“黑话”增强改写、关键词与向量混合检索、召回内容重排清洗等手段，实现用户的查询需求的精确检索，并利用

大模型对召回文本进行总结提炼输出。**家宽业务装维支撑智能体**支撑全省 7000+ 名装维人员，装机自助处理率 35%+，每月自智化解决工单数 5w+，综调支撑人员减少 12 名，咨询单处理时长压降 80%。**网络安全分析智能体**已接入安全告警中台，支撑告警降噪以及事件分析流程。单条告警研判时间由 1 分钟缩短至 0.1s 内，分析准确率达 90% 以上。**政企业务支撑智能体**已接入政企业务网络支撑系统、掌中集客 APP、政企产品交付工作台等生产系统，为 1800+ 一线人员提供智能化支撑，月交互量超 3w 次，协助一线人员工作提效 300%，支撑降本 300 万/年。

（四）知识产权领域：专利智能体高质量数据集建设

（1）数据集建设背景

高质量专利数据集作为专利智能体能力建设的核心底座，直接决定了智能体在专利场景的专业度、合规性与任务执行能力。八月瓜科技有限公司围绕知识产权领域专利智能体的全场景应用需求，系统阐述专利智能体高质量数据集建设的核心思路、实施路径与应用成效，为垂类领域智能体数据集建设提供实践参考。

（2）场景与需求

一是多智能体协同专利撰写场景的结构化数据需求。要求数据集拆解专利撰写核心任务节点，构建匹配各智能体角色的专项数据集，明确各环节输入输出标准、技术逻辑规则与合规要求，支撑多智能体分工协作与全链路任务执行。**二是垂类大模型微调的场景化指令数据需求。**要求数据集覆盖专利撰写全场景的指令任务，包含完整的技术

逻辑关联数据与专业解释数据，确保模型通过微调能够精准掌握专利文本的撰写规范、逻辑结构与专业术语使用规则。**三是 RAG 检索增强的实时性知识数据需求。**需要通过 RAG 检索增强能力，确保智能体生成内容的合规性与准确性。该场景要求构建全量向量化的专利知识库数据集，覆盖全技术领域的专利文献、法律法规、审查规范，具备实时更新能力与高效精准的检索能力，为智能体提供实时、可靠的知识支撑。**四是跨技术领域的场景适配数据需求。**要求数据集构建完善的技术领域分类体系，覆盖各细分技术领域的专利样本与撰写规则，支撑智能体自动识别技术领域，精准匹配对应领域的专业要求。

（3）建设路径与做法

以专利智能体的全环节应用为核心，构建“基础层-能力层-应用层”三级数据集架构，**一是建立数据集与智能体能力的映射体系。**明确数据集与智能体能力的一一映射关系，确保数据集建设全程贴合智能体落地需求。**二是构建全技术领域覆盖的数据分类体系。**基于 IPC 国际专利分类号，构建多层级数据分类树，覆盖机械、电学、化学、通信、人工智能等全部核心技术领域，为跨领域智能体适配提供分类体系支撑。**三是明确全流程数据质量管控标准。**围绕专利场景的专业性与合规性要求，建立数据完整性、规范性、关联性、时效性四大维度的质量管控标准。

（4）应用成效与典型成果

专利智能体高质量数据集体系，已全面落地应用于妙算大模型专项微调、专利多智能体系统研发与知识产权全场景业务服务。

一是垂类大模型适配能力提升，生成内容专业度与合规性优化。

妙算大模型专利文本生成专业术语准确率从 82.3%提升至 96.7%，技术问题-技术方案-技术效果逻辑一致性达标率从 78.5%提升至 95.2%，专利申请文件格式规范合规率从 85.1%提升至 98.4%，权利要求书新颖性支撑表述达标率从 76.2%提升至 93.8%。二是多智能体协同执行能力全面增强，专利撰写效率与质量实现双提升。专利智能体完成完整发明专利申请文件初稿的平均耗时从 15 个工作日压缩至 2 小时，撰写效率提升超 90%；相较于单模型生成模式，多智能体协同生成的专利文件代理机构一次通过率从 68.7%提升至 92.3%，降低了人工修改与校验的工作量。三是 RAG 检索增强能力稳定落地，合规性支撑能力达到业务应用标准。专利智能体向量检索的 Top10 命中率达 97.2%，单条检索请求响应时间低于 200ms。

（五）汽车领域：汽车金融智能体高质量数据集建设

（1）数据集建设背景

汽车金融业务具有链条长、参与方多、规则复杂和风险约束强等特点，既涉及客户、经销商、主机厂和金融机构的协同，也涉及产品政策、授信规则、影像资料、合同文本和贷后行为等多类信息。随着客户服务线上化、流程管理精细化和风控要求动态化，传统依赖人工和单点系统协同的作业方式，已经难以满足高频咨询、复杂审核和闭环执行并存的业务需求。东风汽车金融有限公司形成“多源采集、场景组织、分层标注、质量评估、闭环迭代”的建设路径。

（2）场景与需求

一是客户服务场景，智能体需要处理贷款咨询、产品对比、首付试算、利率政策解释和申请条件确认等任务。要求数据集不仅具备产品知识和政策语料，还应覆盖客户画像、咨询上下文、追问逻辑、推荐依据和工具调用结果。**二是业务审核场景**，智能体需要协助完成进件材料识别、资料完整性检查、字段一致性比对、规则命中校验和异常升级流转。要求数据集同时覆盖影像样本、结构化申请信息、审核日志、规则配置、外部校验结果和人工复核结论，并能够表达“材料识别-规则判断-异常发现-人工确认”的完整过程。**三是流程协同场景**，智能体需要在经销商、客服、审核员、风控人员和贷后运营之间完成任务接力和状态同步。要求数据集具备多角色、多轮次、多节点的过程轨迹，能够支撑智能体理解任务流转关系、识别当前节点职责并生成符合流程约束的动作建议。**四是风险处置场景**，智能体需要围绕欺诈风险、逾期风险、资料造假和规则冲突等问题开展识别与辅助处置。要求数据集覆盖风险触发条件、证据链条、回退路径、人工仲裁结论和处置结果。

（3）建设路径与做法

一是在数据采集环节，建立面向业务全链路的多源归集机制。将客服咨询记录、线上对话日志、产品政策文档、进件申请表、身份证明与收入证明影像、OCR 识别结果等数据统一纳入建设链路，并建立字段规范、任务标签规范和元数据规范，对数据来源、采集时间、场景类型、脱敏等级和版本状态进行统一登记。**二是在场景化组织环节**，围绕重点任务构建场景模板，每个场景模板明确任务目标、参与

角色、输入信息、关键动作、工具依赖、约束条件和结果判定，并保留异常分支和中间状态。**三是在标注体系环节**，采用“基础层、语义层、行为层、评价层”框架。对于高风险场景，额外设置安全与风险标签域，对敏感信息、风险触发片段和处置建议进行专项标注。**四是在复杂数据构建环节**，重点加强工具调用样本和多步任务样本。围绕“智能顾问推荐”“资料核验助手”“审核协同助手”“贷后风险协同助手”等应用，构建了工具调用、规则判断和多轮协同样本。**五是在质量评估环节**，建立“基础质量+能力适配+任务反馈”的评估体系。对低质量样本实行分级门禁，不符合标准的数据进入复标、补采或剔除流程。**六是在迭代机制上**，建立“问题发现-数据补齐-效果验证-版本发布”的闭环。设置月度修订、季度升级和年度重构三层节奏，对数据结构和标签体系进行系统调整。

（4）应用成效与典型成果

一是在客户服务场景中，基于场景化数据和产品知识增强数据训练的智能顾问智能体，试点阶段客户意图识别准确率提升至 93%左右，复杂咨询场景下的首轮有效应答率较传统问答方案提升约 18%，人工客服转接率下降。**二是在进件审核场景中**，资料核验智能体在重点渠道试点中，材料缺失和字段冲突类问题的自动识别率显著提升，人工初审平均耗时缩短约 30%，审核人员能够将更多精力投入复杂疑点和高风险样本。**三是在流程协同场景中**，智能体根据节点状态生成下一步动作建议、自动汇总上下文信息，并提示补充材料和升级事项，部分协同工单的流转往返次数明显减少。**四是在风险处置场景中**，贷后

风险协同智能体可辅助识别重点风险对象、生成处置建议并提示证据材料。高风险判断仍保留人工最终决策，但智能体能够稳定承担风险线索整理、规则依据匹配和处置建议生成等辅助任务。

（六）产业园区领域：园区智能体高质量数据集建设

（1）数据集建设背景

国家级高新区在推进“一网通管”、“产业大脑”等数字化转型过程中，普遍面临数据资源分散、多模态数据融合难、智能体数据供给不足等挑战。**集商网络科技（上海）有限公司**启动“基于园区智能体的多模态高质量数据集建设”，旨在构建一套覆盖全产业、全场景的高质量多模态数据集体系，为园区治理、产业分析、企业服务 etc 智能体提供坚实的数据基础。

（2）场景与需求

围绕园区治理与产业促进的核心业务，识别出的典型场景对多模态数据提出明确需求。一是**产业链分析与强链补链场景**。园区需要实时掌握主导产业链条的运行态势，分析上下游企业的依存关系、技术短板与投资机会。智能体需要融合企业工商、专利文本、供应链新闻、产品图像等多模态数据，构建动态的产业知识图谱，以实现产业链风险预警与补链企业自动推荐。二是**企业全景画像与精准服务场景**。为提供精准的政策匹配、融资对接与空间资源推荐，智能体需要对企业进行全方位、动态的感知。三是**园区安全环保智能预警场景**。园区安全环保监管涉及大量视频监控、传感器数据与环境图像，传统人工巡检与阈值报警方式存在滞后性与误报率高的问题。智能体需要实时处

理海量视频流与物联网时序数据，通过图像识别、异常检测等技术，精准识别烟雾、人员闯入、污染物排放等事件，并自动关联周边企业信息与应急预案，实现从“事后处置”向“事前预警”转变。

（3）建设路径与做法

一是开展顶层设计与标准规范建设，确立数据建设与应用统一框架。制定了《园区数据资源管理办法》《数据分类分级指南》等政策文件，明确了数据权属与流通规则。在此基础上，编制了《多模态数据质量评估标准》《数据标注规范》，强化对非结构化数据的质量要求。**二是构建多模态数据资源池，实现跨源异构数据的统一汇聚与关联。**建设统一的数据采集交换平台，支持 API、库表、文件、流数据等全量接入方式。数据源规划覆盖政务与产业数据、物联网与空间数据、影像与视频数据、互联网公开数据。**三是建立高质量数据治理与加工体系，**以企业实体为核心建立跨模态关联模型，将原始数据转化为面向智能体的知识资产，建立跨模态数据的关联模型。开展专业化数据标注，组建由行业专家与技术工程师组成的标注团队，对图像、视频进行高质量标注、对文本数据进行实体抽取、关系分类等知识图谱构建性标注，所有标注结果均经过三级审核，确保准确率。**四是围绕智能体应用场景，开发主题化、可服务的数据集产品。**封装开发了标准化、场景化的数据集产品，并通过统一数据服务平台提供 API 接口、数据文件等多种服务方式，直接支撑上层智能体应用。

（4）应用成效与典型成果

一是建成面向智能体的核心数据资源库。已初步形成 PB 级的园

区级数据资源库，覆盖文本、数值、图像、视频、时空地理等五种主要模态，数据资产目录清晰。**二是**“情报招商线索数据集”成功支撑产业分析智能体帮助客户识别产业链潜在断链风险点，并自动推荐补链目标企业超百家，线索准确率较传统方法提升 35%。**三是**环保智慧监管场景中基于“重点区域环境与安全影像感知数据集”训练的智能体，识别准确率达 94.2%，平均响应时间从过去的 30 分钟缩短至 5 分钟以内。能耗管理场景中智能体通过融合企业生产报表与设备红外影像数据，定位出 7 家企业的 12 处高耗能异常点，预计年节能量可达 800 吨标准煤。**四是**形成的《高质量数据集 数据标注 产业监测指标体系》及系列标准规范，为产业园区数据要素价值化提供实践范例。

（七）科研领域：模拟计算智能体高质量数据集建设

（1）数据集建设背景

模拟计算作为材料科学、化学、物理学等基础研究的核心手段，因计算流程复杂、专业门槛高、数据处理规模大等特性，成为智能体技术落地的重要场景。当前科研领域模拟计算面临三重核心挑战：**一是计算方法多元**，第一性原理计算、分子动力学模拟等不同方法的数据格式、参数体系差异显著，整合难度大。**二是专业门槛高**，研究人员需熟练掌握复杂计算软件与领域知识，学习和实操成本高。**三是流程繁琐且依赖人工**，全环节人工介入导致效率低、误差高。**成都天玑算科技有限公司**遵循需求驱动、标准先行、质量为本、持续迭代的原则，系统推进模拟计算智能体数据集建设。

（2）场景与需求

模拟计算智能体核心应用场景包括：（1）计算任务自动化场景，智能体根据研究目标自动完成计算方案设计、参数配置、任务提交及结果收集全流程。（2）计算结果智能分析场景，智能体自动解析、可视化呈现与知识提取，辅助科研人员快速获取有价值的科学洞察。

（3）领域知识问答场景，智能体基于专业知识库，为科研人员提供计算方法选型、参数优化、软件操作指导等精准知识服务。支撑三大场景应用的数据集建设需求包括：（1）指令-执行对齐数据，记录自然语言指令与模拟计算操作步骤的映射关系。（2）任务导向标注数据，围绕模拟计算流程，标注任务类型、工具调用、参数配置等核心信息。

（3）领域知识数据，涵盖计算方法原理、软件使用规范、领域专业知识等，构建智能体知识增强检索体系。（4）计算案例数据，用于智能体少样本学习与推理能力培养，提升复杂任务应对能力。（5）结果解析数据，制定计算输出文件解析规则，配套分析结论，训练智能体结果理解与分析能力。（6）动态更新与版本管理数据，建立动态更新机制，纳入最新研究成果与计算方法，实现版本可追溯。

（3）建设路径与做法

一是规划数据资源体系。软件操作数据涵盖主流计算软件的输入输出格式、参数配置规则和操作命令序列，领域知识数据整合计算方法文献、软件官方文档、领域专家经验总结等，计算案例数据收集整理典型研究课题的完整计算流程和结果分析过程，问答对话数据基于真实用户咨询场景，构建问题-答案配对数据。**二是实施多源融合采集与处理。**采用文献挖掘、软件文档解析、专家访谈等多源融合策

略，同时建立全流程清洗流程，通过“自动化脚本+人工审核”模式，统一数据格式与编码标准，校验数据准确性与完整性，去重消歧、修正矛盾信息并对敏感内容进行脱敏处理，确保数据安全可用。**三是数据标注与组织。**内容方面，对计算指令进行意图分类和槽位标注，对知识文本进行实体识别和关系抽取，对计算结果进行结构化标注和结论归纳。流程方面，采用“初标-复核-抽检”三级质量控制体系，确保标注准确率不低于 95%。数据组织方面，采用分层分类的存储架构，基础层存储原始采集数据，加工层存储清洗后的结构化数据，应用层存储面向特定场景的标注数据。**四是落实数据安全与合规。**对不同敏感级别的数据实施分级保护，采用加密存储和传输技术，防止数据泄露风险，建立数据备份和恢复机制，确保数据可靠性。、严格遵守数据保护相关法律法规，对涉及第三方知识产权的数据获取合法授权，建立数据使用审计机制。**五是建立持续迭代与质量评估机制。**自动收集智能体在真实科研任务中的交互日志，经专家复核后纳入训练集。建立多维度质量评估指标，包括完整性、一致性、有效性等。采用版本控制策略，记录数据集的更新历史与变更内容。

（4）应用成效与典型成果

基于建设数据集研发的模拟计算智能体系统在实际科研服务中得到应用验证，任务执行准确率方面，智能体在计算任务自动化场景中的意图识别准确率达到 92%，参数配置正确率达到 89%，显著降低了人工操作错误率。效率提升方面，通过智能体辅助，典型计算任务的准备时间从平均 4 小时缩短至 30 分钟，效率提升约 87%；结果分

析时间从平均 2 小时缩短至 15 分钟，效率提升约 88%。用户满意度方面，基于 200 余名科研用户的调研反馈，智能体系统的整体满意度达到 4.5 分（满分 5 分），其中“操作便捷性”和“结果准确性”两项指标得分最高。

编制说明

本研究报告自 2026 年 4 月启动编制，分为前期研究、框架设计、文稿起草、征求意见和修改完善五个阶段。本指南聚焦智能体(Agent)能力构建对数据的新需求，围绕智能体数据集建设面临的问题，系统提出面向智能体能力构建的高质量数据集建设总体框架、建设流程、关键技术和质量保障方法，并结合典型应用场景形成实践参考，为开展智能体高质量数据集建设提供方法指引和实施参考。本报告由中国信息通信研究院人工智能研究所、联想（北京）有限公司合作撰写。

参编单位：东软集团、东风汽车金融、浪潮卓数、中电信数智、科大讯飞、中科院上海微系统与信息技术研究所、合肥数据空间研究院、中科院上海高等研究院、北京邮电大学经管学院、砺英数智、海天瑞声、浩鲸科技、苏州市轨道交通集团有限公司、广州赛宝认证中心、北京长安汽车工程技术研究有限责任公司、标贝（青岛）科技、航天云路、中汽智造、东方财富、扬州市数据局、北京工业大学研究院、北京科杰科技有限公司、广域铭岛数字科技有限公司、安徽飞数信息、中国司法大数据研究院、中科斯欧（合肥）、积家传承创新（北京）科技有限公司、数据要素社。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn

