蚂蚁集团 ANT GROUP

CAICT 中国信息通信研究院

大模型
行业可信应用框架
研究报告

Research Report on Industry Trustworthy Application Framework of Foundation Models

蚂蚁科技集团股份有限公司
中国信息通信研究院

2024年9月

序一

当前，新一代人工智能技术变革方兴未艾，大模型技术作为其代表，正在成为推动产业智能化升级和经济高质量发展的重要引擎。国家高度重视人工智能发展，已经出台一系列政策措施，为人工智能与实体经济深度融合指明了方向。然而，正如人类历史上的每一次科技革命，新技术的应用在带来巨大发展机遇的同时，也伴随着潜在的风险与挑战。如何确保人工智能技术安全、可靠、可控地发展和应用，构建人类信任的人工智能，已成为全球共同关注的时代命题。

这份研究报告聚焦大模型可信应用这一前沿课题，深入探讨了大模型在金融、医疗、政务等专业领域应用所面临的挑战，并提出了系统化的解决方案。报告立足于技术架构和体系建设，从数据质量提升、模型能力增强、推理过程可控、系统安全保障、评测体系健全等多个维度，构建了面向专业领域的大模型可信应用框架，为推动大模型在产业中的规模化落地应用提供了有益参考。

这份报告的发布恰逢其时，我相信它将对推动人工智能技术与产业的深度融合，促进大模型技术的可信落地应用产生积极的影响。期待未来有更多科研机构、企业和个人参与到这一领域的研究和实践中，共同努力构建出安全、可靠、可控的大模型应用生态，为新时代的智能化产业升级和经济高质量发展贡献更多智慧和力量。

蒋昌俊

同济大学讲席教授、中国工程院院士

序二

以大模型为代表的新一轮人工智能发展浪潮席卷全球，已成为各个行业全面迈向智能化的新引擎。当前，在缩放定律驱动下，模型能力还在持续提升，基础大模型的语言、视觉和多模态能力加速迭代。与此同时，支撑大模型应用的工程架构也在不断完善，智能体、检索增强生成、模型即服务等新技术新模式的出现，拉近了大模型与用户的距离。

然而，我们应该认识到，虽然大模型在智能客服、知识管理、软件开发等场景中的落地应用越来越多，但要在复杂度高、容错率低的场景中实现规模化应用，仍然面临不少挑战。比如，大模型嵌入知识的学习和更新成本较高，专业性可能存在不足且难以及时更新；大模型的推理过程如同黑盒，难以解释，这使得其难以直接应用于复杂的推理任务；大模型存在“幻觉”现象，可能导致生成的结果缺乏事实依据或数据支撑。因此，如何确保大模型应用的专业性、可控性、真实性和安全性，是跨越大模型落地最后一公里的关键一环。

这份报告结合了中国信通院研究和相关企业的实践经验，系统梳理了大模型可信应用面临的问题和实践路径。在概念上，报告分析了金融、医疗、政务等多个领域的大模型的应用实践经验，提出了“面向专业领域的大模型可信应用框架”。在实践上，报告将“专业、可控、真实、安全”等四个要素作为大模型可信应用的核心目标，并针对数据质量提升、模型知识增强、内容生成可控等七个维度提出落地实施建议，力求形成一套推动大模型可信应用的方法论。

我相信，这份报告一定能对各个行业深入探索大模型应用带来新的启发。我也期待产学研各方进一步紧密携手，不断丰富大模型可信应用实践，充分释放人工智能的无限潜能。

余晓晖
中国信息通信研究院院长

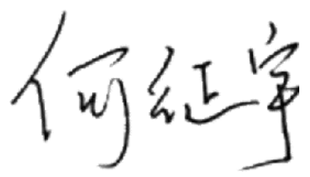
序三

人工智能大模型技术如同一颗"小火花"，正在点燃新世界的无限可能，为"新质生产力"的发展注入强劲动力。作为以科技创新为根本驱动力的科技企业，蚂蚁集团也在持续投入资源，积极探索大模型在专业、严谨领域的可信应用。我们的目标是让AI像扫码支付一样便利每个人的生活，同时确保其在专业领域的应用安全可靠。为此，我们推出了智能生活管家、金融管家和健康管家三大应用，旨在将大模型技术落地到实际场景中。

然而，我们也清醒地认识到，大模型在行业中规模化可信应用仍面临诸多困难与挑战。要让大模型在行业应用中不断提升其专业性，达到专业人士的水准，需要不断提高数据质量，特别是专业数据的质量、增强模型的专业能力、确保生成内容的可信、提升推理决策过程的可控可解释以及构建完善的专业领域深度评测体系，都是我们需要深入研究的课题。

本报告提出了面向专业领域的大模型可信应用框架，旨在为解决这些问题提供一些思路和方案。我们相信，只有通过产业上下游的通力协作，才能真正推动大模型在产业中的规模化可信应用，助力产业智能化升级。

让我们携手共进，以开放、创新的态度，共同探索人工智能大模型的无限可能，为构建智能化、数字化的美好未来贡献我们的力量。



何征宇

蚂蚁集团副总裁&首席技术官

前言

以大模型为代表的新一代人工智能技术变革仍在加速迭代，为“新质生产力”的发展注入强劲动力，助力产业智能化升级和经济发展。大模型在产业端已经开始在任务简单、容错率高的场景得到广泛地使用，例如智能客服、文案生成等场景，但是在医疗疾病诊断、金融投资理财等任务复杂度高、容错率低的场景，大模型的规模化落地仍然遇到不少阻碍。为进一步释放以大模型为核心的人工智能新技术范式生产力，需要解决大模型的“可信应用”问题，支撑大模型在金融、医疗、政务等专业领域的深化应用。为此本报告提出了面向专业领域的大模型可信应用框架，即大模型为核心的智能系统在面向金融、医疗、政务等专业领域的应用中，为确保应用的专业性、可控性、真实性和安全性，应当如何构建系统的技术架构和体系。

从技术实现来看，企业建设一体化的大模型可信应用框架目标是在大模型开发和应用的不同环节，施加相应的技术保障手段，以提升落地应用的可信程度。目前主要手段包括提升数据供给质量、增强模型在应用领域的专业能力、提高模型生成内容的可控可信、提升智能体推理的可信、使用围栏工具和安全保障措施、建立全面充分的评测体系、构建“反馈-迭代”的良性循环体系等手段，综合实现大模型在专业领域中的稳妥应用和持续深化拓展。**从产业应用实践来看，**目前在金融、医疗、政务等行业中已有大模型应用框架的落地实践，提升了推理结果的准确专业真实性，并增强了推理过程的透明安全可控，提高了客户对于大模型应用的信赖程度。

展望未来，本报告中提出的大模型应用框架是推动大模型在产业中规模化落地应用释放价值的初步探索，未来的产业化突破还需要从前沿技术创新探索、可信应用框架落地实施、行业治理体系搭建、产业生态合作完善等多个维度统筹推进。我们期待产业上下游各方通力协作，共同推动大模型在产业中的规模化可信应用落地，助力产业智能化升级。

目 录

图 目 录

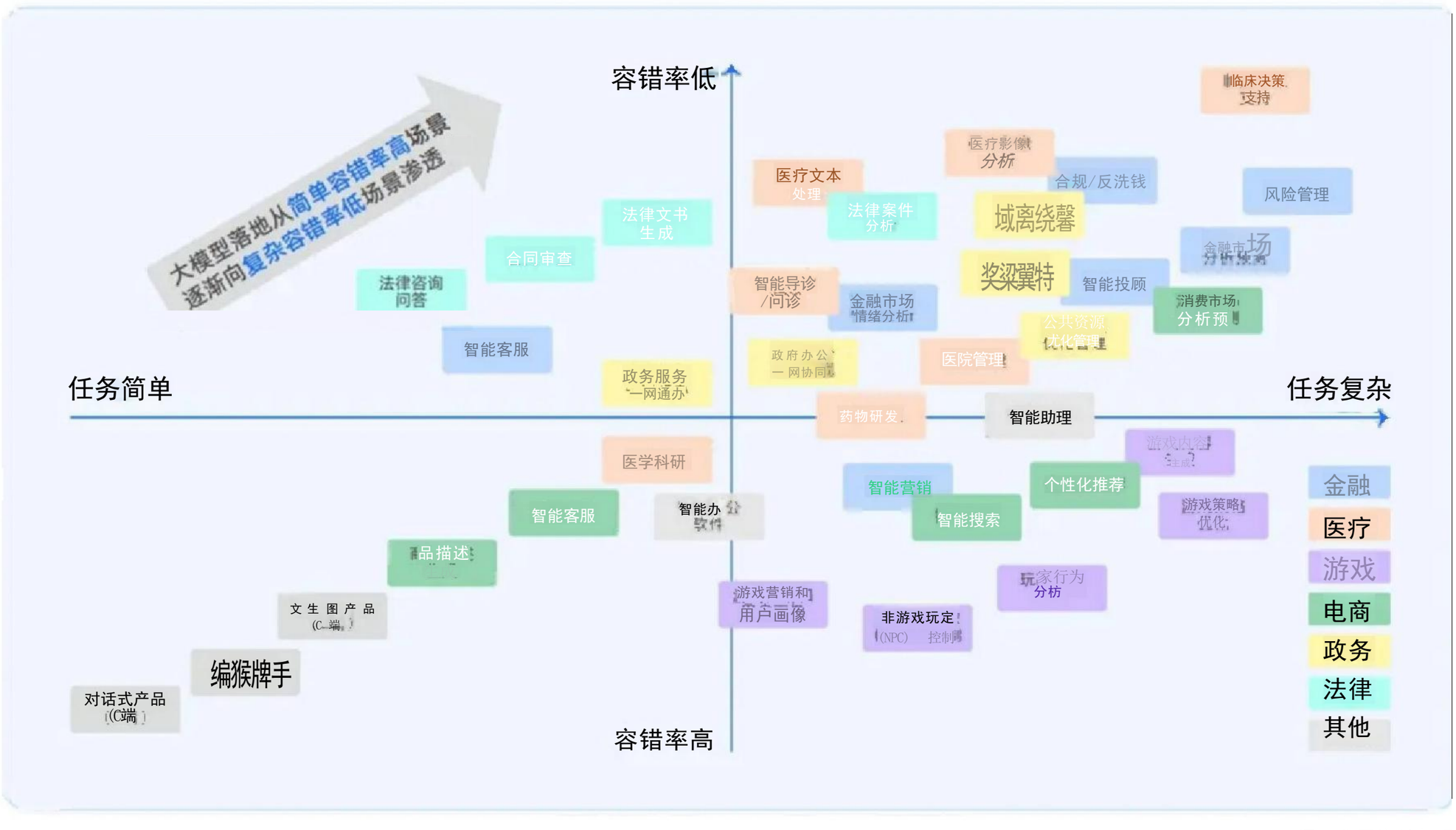
01

大模型可信应用背景和需求

本章梳理了当前大模型在垂直行业的落地场景路径，通过分析当前产业界人工智能可信的内涵和定义，进一步细化了面向专业领域的大模型落地可信核心要求，并提出了大模型可信应用框架的内涵。

（一）大模型加速落地激发应用可信需求

大模型落地应用已进入产业变现阶段，助力全球产业升级和经济发展。一方面，以大模型为代表的新一代人工智能技术变革仍在加速，为各产业带来了全面智能升级的可能性。无论是制造业、金融业、医疗行业还是教育行业，都可以通过集成大模型实现产品和服务功能的重构与提升，提高操作效率，优化用户体验，以适应更高层次的用户需求和市场期待。另一方面，大模型的通用性、泛化性进一步赋能传统机器学习模型无法支持的场景。受限于数据和实施成本等因素，传统机器学习模型通常只能面向特定场景和任务进行定制开发，而大模型具备更强的泛化能力和学习能力，能够基于较少的数据和较低的开发成本赋能海量长尾应用场景。IDC 预测①，全球生成式人工智能市场年复合增长率或达85.7%，预计2027年45%的企业将掌握并使用生成式人工智能共同开发数字产品和服务，全球生成式人工智能市场规模将接近1500亿美元。



来源：公开资料整理

图1大模型落地场景发展路径

① https://www.idc.com/getdoc.jsp?containerId=prCHC51997124

当前，大模型规模化落地应用主要集中容错率高、任务简单场景。例如，在电商领域，基于大模型的智能客服能够通过生成商品描述和智能搜索功能，快速响应用户查询，提供个性化推荐；在游戏行业，大模型通过生成游戏内容和优化游戏策略，增强玩家的沉浸感和游戏体验。主要原因**一方面**，这些场景更具备通用可复制性，通过大模型赋能实现自动化和流程化处理，显著提升效率、降低应用成本。**另一方面**，此类场景对模型输出的准确性要求相对低，即使偶尔出现偏差，也不会对用户造成显著影响，进而助力大模型能够在这类场景实现快速部署和迭代优化。

大模型落地应用将持续向复杂容错率高的场景渗透。例如，在医疗领域，大模型可被用于辅助进行更为复杂的影像分析、临床决策，以及在药物研发中进行分子结构的预测和筛选。在金融领域，大模型的应用可扩展到智能投顾、风险管理和预测。在法律领域，大模型可被用于复杂的案件分析和判决结论生成等。这些场景**一方面**涉及高度专业化的知识和复杂决策过程，要求对专业领域的知识和流程深入理解，大模型需要结合高质量领域数据集、外挂专家知识库及复杂的调优过程才能具备此类能力；**另一方面**对输出结果的准确性和可靠性有着极高要求，任何错误都可能带来较大影响和损失，对大模型的输出结果和推理决策能力提出挑战。

进一步推动大模型规模化落地需要解决大模型“可信应用”问题。从应用现状可以看出，大模型在更广泛场景深入应用的关键核心在于提升大模型应用的可信程度，即大模型是否能够在应用场景中真实、准确且安全地提供服务，行业亟需一套有参考性的系统化可信应用框架，以推进大模型规模化可信应用。

（二）人工智能可信内涵及外延不断丰富

在人工智能的发展过程中，源于对人工智能系统安全性、透明度、公平性等方面的关注，可信**(Trustworthiness)** 已成为各国政府、国际组织、企业和学术界都较为关注的理念，行业各方针对“可信人工智能”**(Trustworthy AI)** 也已制定了一系列原则、标准和实践指南。发展至今，人工智能的可信内涵已非常丰富，不仅涵盖了技术的准确性、安全性、鲁棒性和可解释性，还需要考虑伦理问题如公平性和隐私保护，以及社会影响如公众接受度和法律合规性等多个方面。

在国际标准**ISO/IEC 22989中**，可信的定义是“能够以可验证的方式满足相关方的预期的能力”，并认为构成可信的关键要素包括：责任性(Accountability)、准确性(Accuracy)、真实性(Authenticity)、可用性(Availability)、可控性(Controllability)、完整性(Integrity)、隐私(Privacy)质量(Quality)、可靠性(Reliability)、弹性(Resilience)、鲁棒性(Robustness)、安全性(Safety)、信息安全(Security)、透明性(Transparency)和易用性(Usability)等，如图2所示。此外，还有观点认为应包括公平性(Fairness)、多样性(Diversity)和包容性(Inclusiveness)等。

由此可见，人工智能的“可信”是一个内涵丰富的概念，从不同视角和拟解决的问题域出发，会有不同的关注侧重点。例如，**从功能视角看，**会更关注准确性、可靠性、完整性、可用性、信息安全等要素；**从伦理道德视角看，**会更关注公平性、安全性、多样性、包容性、责任性等要素；**从应用交互视角看，**会更关注可控性、真实性、透明性、易用性等，这些要素可以帮助用户更好理解、控制和使用AI系统。

来源：公开资料整理

图2可信人工智能的内涵丰富

（三）专业领域大模型应用关注四大要素

综上所述，如果抛开具体的问题领域讨论人工智能的“可信”，将容易陷入概念繁多重叠、范围界定不清晰的情况，业界难以形成共识并推动问题解决。为了解决我们所关注的人工智能产业应用问题，尤其是大模型在专业领域落地推广问题时，需要结合场景范围进一步厘清“可信”的关键要素，明确“可信”目标。总体来看，专业领域对大模型可信应用的要求主要体现在：

专业领域知识及业务流程复杂，大模型需具备专业性。专业领域涉及复杂的专业知识和业务流程，大模型需要对该领域的术语内涵、表达方法、推理逻辑和操作流程等有足够深入的认知和遵循。例如，**医疗领域医**学术语众多且专业性强，大模型需要对这些术语的确切含义和使用场景有准确的理解，同时医疗诊断和治疗涉及复杂的生理机理和临床流程，大模型需能够准确推理和遵循这些流程。**金融领域**交易和风险管理有严格的法规和操作规范，同时金融数据分析需要复杂的统计和建模方法，大模型也需能够准确理解和遵循这些规范或方法。

专业领域强调决策逻辑严谨透明，大模型需确保可控性。专业领域的决策需要有科学的方法、严谨的过程，明确的依据。大模型在提供决策支持时需要符合专业指引，推理过程需要可解释和可追溯，以确保其输出结果的专业性、科学性与合规性，同时在发现不合理输出时可以进行纠正。以医疗领域为例，大模型在辅助诊断和治疗时需能够解释其推理过程，说明为何做出特定诊断或治疗建议。医生需要了解模型的判断依据，以确保其与医学常识和经验相符，如果模型做出违背医学规范的建议，医生能够通过交互引导模型进行纠正。

专业领域要求结果准确实时，大模型需保障真实性。专业领域对推理结果的真实性和准确性有极高要求，也就是说大模型在生成内容时，需基于真实且没有事实性错误的数据，部分应用场景对于结果的实时性也有极高的要求。以金融领域为例，大模型在辅助进行风险评估和投资决策时，必须基于真实的市场数据和财务数据进行分析和预测，对虚假信息或事实性错误容忍度较低。同时，金融市场变化迅速，大模型应能够及时获取并根据最新的市场信息和数据来生成内容或进行推理决策，以应对该领域信息的动态变化。

专业领域涉及高价值数据和场景，大模型需确保安全性。专业领域一般涉及到高价值数据，安全合规要求较高，大模型需能够确保自身安全，并符合相应业务的合规性、安全性要求。例如，专业领域大模型通常需要处理大量敏感数据，如个人隐私信息、商业机密等，这些数据一旦泄露或被恶意利用，可能会造成严重的经济损失和声誉损害等。同时，在专业领域部署的大模型也更具攻击价值，在训练和部署过程中会面临着更大安全风险，大模型应用系统需具备抗攻击能力，能够识别和抵御各种攻击手段，确保模型的安全性和可靠性。此外，专业领域通常有严格的法规和标准要求，如金融行业的反洗钱法规、医疗行业的隐私保护要求等，大模型应符合这些合规性要求，避免在应用过程中违反相关法律法规。

（四）大模型可信应用技术框架及体系

结合大模型在专业领域的可信应用要求，以及当前在金融、医疗、政务等领域的落地实践，本报告提出面向专业领域的大模型可信应用框架，并优先选择“**专业、可控、真实、安全**”四类要素作为可信应用框架的关注重点。通过这一框架，期望能够帮助行业中的开发者和应用者更好地推动大模型在各行各业中的应用深化和场景拓展。

具体而言，本报告中大模型可信应用框架的内涵为：在面向如金融、医疗、政务等专业领域应用中，构建一个以大模型为核心的专业智能服务体系，该体系应确保应用的专业性、可控性、真实性和安全性，以满足专业领域高标准要求，如图3所示。

专业性

·大模型不仅需要掌握所应用领域的专业知识，生成与预测结果需准确符合专业领域要求，还需要在提供决策支持时遵循专业领域的方法、过程和指引。这意味着，模型需要能准确理解专业领域的术语内涵和用户问题意图，并能够在领域专业语境下进行问题分析，通过符合专业规范的流程实践，给出符合专业逻辑和规范的输出。

·例如，在医疗领域，模型需要根据标准诊疗流程准确诊断疾病并提供符合临床指南的有效治疗建议；在金融投顾场景中，模型需要准确理解客户的投资需求，并给出符合金融专业标准和流程的分析和决策建议。

可控性

·大模型在应用中需要实现两个层面的可控， 一是模型本身推理过程的可控，即推理过程需要足够透明可解释；二是模型在专业领域应用中的推理结果和决策可控，即能够确保从推理结果到最终决策的过程可被有效控制和调整。也就是说，大模型在推理过程中，可以通过任务分拆执行、输出结果核验、执行反馈纠正等中间过程，引导系统的输出结果收敛到合理范围内，同时，模型的输出结果能够被专业人员有效审核和必要时进行人工干预，以保证最终决策的准确与合规。

·例如，在金融投顾场景中，模型在输出投资建议时需给出推理过程，并可结合用户输入和反馈持续改进其建议，同时专业投顾人员可以审核和必要时调整模型的建议，最终帮助用户做出可靠的投资决策。

真实性

·大模型应能生成与现实世界事实相一致的内容，并在推理过程中忠实于输入的上下文数据，尽量避免虚假信息和幻觉。这要求模型具备足够可靠的通用常识和逻辑能力，在生成时尽量不出现“幻觉”内容，确保输出是基于真实上下文数据和可靠专业知识的合理推断。

·例如，医疗领域中大模型在给出治疗建议时，需要充分结合患者的临床表现和诊疗数据进行分析，并基于循证医学证据给出可靠的结果输出。

安全性

·大模型应通过体系化的安全措施，防止外部的恶意攻击，确保模型本身、系统、数据和个人信息的安全，这是大模型在行业应用的基础保障。这需要结合系统安全、算法安全、数据安全，针对大模型在应用过程中潜在的安全风险与挑战，采用更有针对性的安全措施进行安全保障与实时对抗，如系统安全攻防、数据安全和隐私保护、身份权限管理等。

·例如，在医疗领域中，大模型应用过程需要严格保护患者的隐私数据，防止未经授权的访问和使用、或者在模型输出结果中不当泄露患者隐私，模型的管理也应该设立严格的权限体系。同时，还需要防范可能针对医疗大模型系统的网络攻击与渗透，确保系统和模型本身的安全性和可靠性。

来源：中国信息通信研究院，蚂蚁集团

图3大模型可信应用框架的内涵

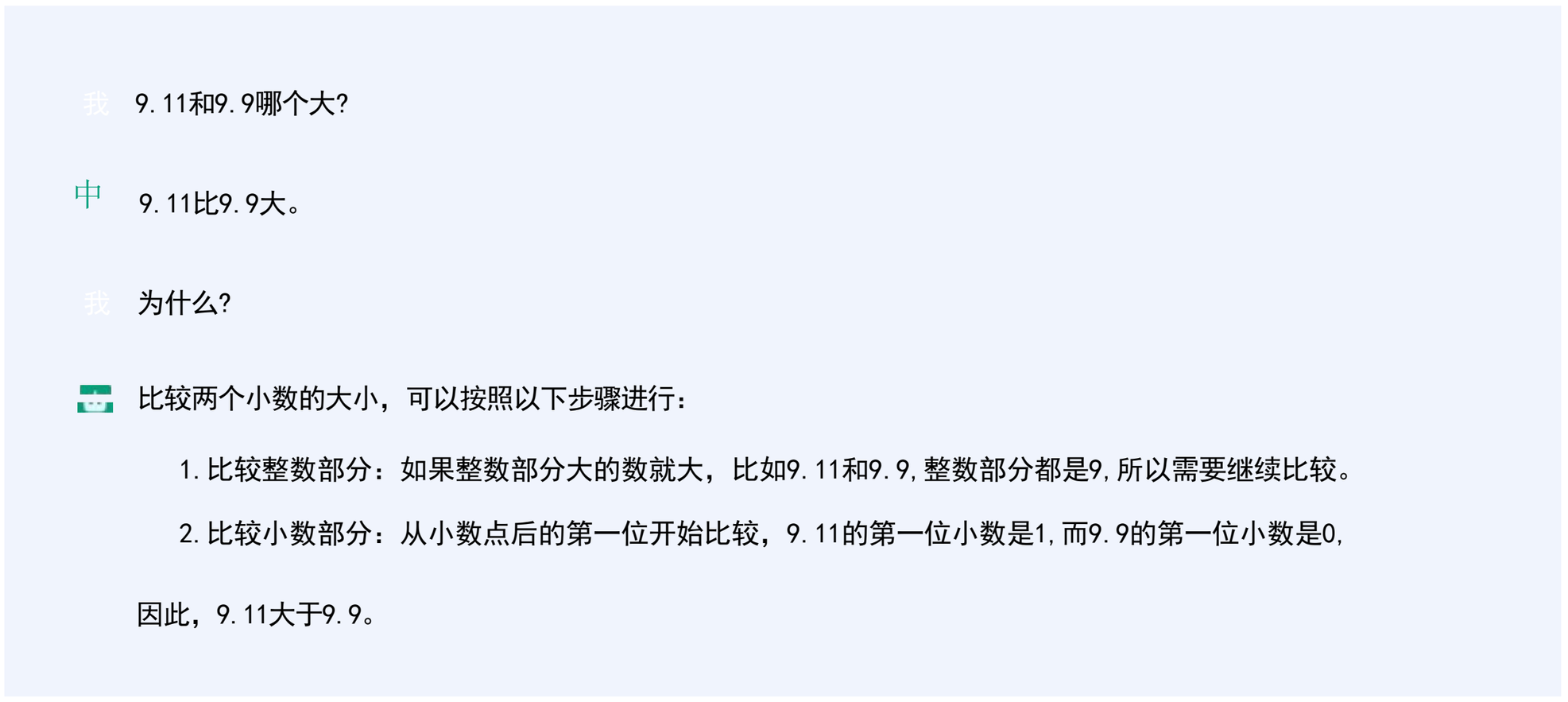
02

面向专业领域的大模型可信应用挑战

在专业领域应用时，大模型在应用领域落地所表现出的专业性、可控性、真实性和安全性均有比较高的要求，实现难度也较大。目前来看，仅依赖提升大模型自身能力一般难以直接满足在专业领域应用的要求，需要结合其他技术和工具共同构建体系化的应用解决方案，并需要依赖相应的行业治理体系和产业生态协作支撑，才能保障大模型在专业领域的可信应用落地。当前，面向通用场景生成内容的可控、真实等需求，已经有了单点的技术解决方案，但是面向专业领域更深层次的应用，仍然缺少更系统化的方法论和体系化的解决方案。

（一）模型技术能力不足，可信应用亟需高标准严要求

大模**型内置专业知识的学习和更新成本较高，专业性可能存在不足且难以及时跟上领域最新进展。具**体表现在一是专业知识存在时滞性，大模型学习到的知识一般只覆盖到所收集到的训练数据截止日期前，难以及时吸收和输出领域的最新进展数据。**二 是**知识更新成本高，对大模型内置知识的更新需要重新进行训练或微调优化，每次更新都需要耗费大量时间、资金和算力。**三 是**专业场景的逻辑推理和数值运算能力不足，大模型当前逻辑思维较弱可能导致模型输出较为低级的错误，如**图 4**所示例。

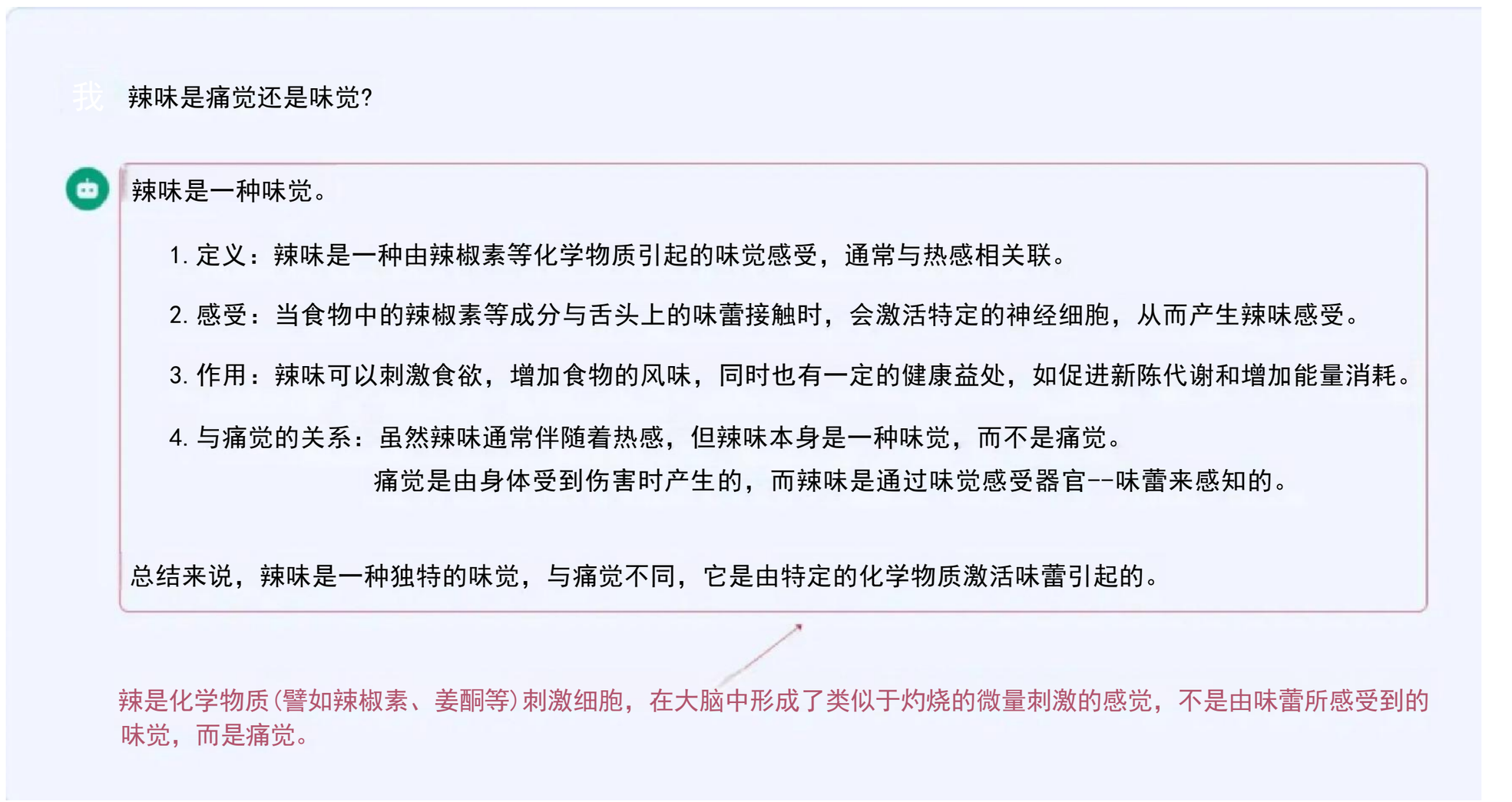


来源：某大模型对话窗口

图4示例-大模型在数值运算时可能出现错误

大模型内部的推理过程不透明，难以直接应用于专业领域中常见的复杂推理任务中。推理过程可信性不足且难以验证，大模型的内置知识是以隐式的方式存储在模型的海量参数中，内部推理过程是个“黑箱”，难以解释它们的决策依据以及得出结论的过程。专业领域的场景任务往往需要复杂的、多步骤的逻辑推理和计算，直接应用大模型进行复杂推理任务，过程无法解释、结果可能不可靠，而且即使被专业人士识别出了错误也难以进行反馈纠正，即可控性不足。

大模型的“幻觉”问题，这种缺乏事实依据或数据支撑的虚假生成结果可能会给专业领域的应用带来严重后果。大模型有时会生成看似合理但实际上是虚假的内容，这种"一本正经地胡说八道"的现象被称为“幻觉”。此类虚假内容往往看起来很真实，即使是专业人士也很难一眼就精准发现其中的谬误，而且大模型自身缺乏核验机制，难以依靠自己来纠正潜在错误或虚假认知。在金融、医疗、政务等专业领域的一些应用场景中，如果采信了“幻觉”信息，将可能会导致严重的错误，如图 5所示例。



来源：某大模型对话窗口

图5示例-大模型的输出可能出现“幻觉”

大模型在专业领域应用尚缺乏系统性的安全分析与安全措施做保障。随着大模型技术迅猛发展，其全生命周期中各个环节都面临着新的风险与挑战，需要进行系统分析并给出相应安全措施。以2023 年爆出的CVE-2023-48022 漏洞为例，它揭示了一个被广泛使用的AI开源计算框架Ray 在设计和使用上存在的安全风险，包括鉴权缺失(攻击者可以借此绕过正常的身份验证)和恶意远程代码执行(攻击者可以借此控制被攻击系统执行任意命令或加载恶意软件)等。该漏洞的影响范围广泛，涉及OpenAI、Hugging-face、Stripe 等多家企业，导致用户身份安全凭证被窃取，并由此可能导致模型被滥用或者敏感数据泄露。值得进一步注意的是，伴随大模型智能体的广泛应用，它们往往会被赋予更高的权限来访问敏感的API 或数据， 一旦攻击者通过单点突破实现了对智能体的控制，并以此为跳板访问更多敏感应用和数据，可能会导致较为严重的安全后果。

（二）保障体系构建不全，产业界体系化方案仍然缺乏

大模型在专业领域的可信应用尚缺乏体系化的解决方案。通过前面的分析可见，目前仅靠大模型自身难以直接满足在专业领域应用的高标准要求，需要结合其他技术或专业工具协助大模型“扬长避短”。例如可把大模型与知识工程相结合，通过知识结构化整合、逻辑规则构建等过程，为大模型的推理过程提供专业可信的知识基础和清晰的推理逻辑。目前行业中已存在着一些大模型与周边技术结合的落地尝试，但总体来看，目前行业中仍缺乏比较全面系统的整体解决方案指南。因此，有必要站在面向专业领域的可信应用全局视角，对大模型与多种技术的结合进行系统梳理，并提出相对全面且可落地的可信应用框架以及技术实现指南，这也是本报告核心研究的问题。

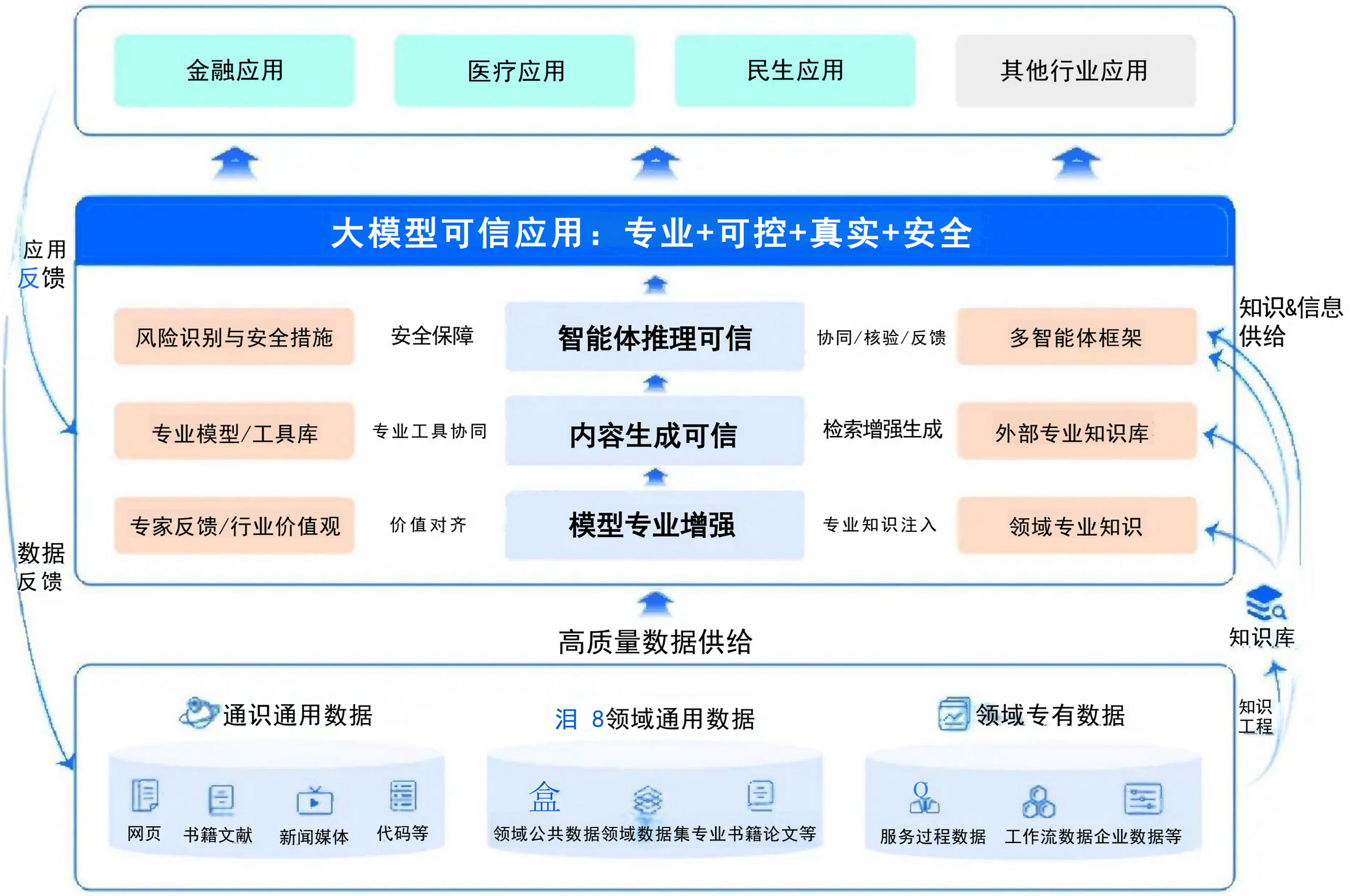
大模型在专业领域的可信应用所需的行业保障尚不完善。大模型在专业领域的规模化可信应用，不仅要依赖于技术解决方案的完善和升级，还需要健全的产业合作方式、统一的标准体系和完善的评估评测机制。一是由于大模型技术的持续快速迭代，亟需构建新型的可信应用行业合作框架，此框架应当涵盖监管机构、大模型开发企业、大模型应用企业以及独立的第三方专业评测机构的紧密协作，且能够灵活根据大模型的技术迭代实现快速升级。二是针对于金融、医疗、政务等专业领域的大模型可信应用，目前尚存在较多的技术标准和应用规范空白，缺乏统一且细化的标准指导和规范约束，企业在应用大模型时缺少有效的参照和依据。三是大模型在专业领域可信应用的评测验证机制目前也尚不完善。对于大模型评测来说，如何设计评估指标、如何构建评测数据集尤其是专业领域应用评测数据集，以及如何能够快速迭代跟上大模型能力上的发展，都亟待从行业层面共同构建以完善。

03

面向专业领域的大模型可信应用框架

（一）可信应用框架总体视图

为能够推动大模型在专业领域中实现可信应用，需在大模型开发和应用的不同环节采取相应的技术保障手段，以提升落地应用的可信程度，本报告中提出的大模型可信应用框架如**图 6**所示。



来源：中国信息通信研究院，蚂蚁集团

图6面向专业领域的大模型可信应用框架

具体而言，**一是要提供高质量的数据供给，**尤其是应用到专业领域时，需要拓展更多具有专业性和多样性的可信数据来源，并对领域相关数据进行高质量处理，以尽量确保数据集所包含数据的真实、准确、合规、一致且有较强时效性，这是大模型可信应用的基础。

二是通过丰富、完整的专业知识增强大模型的专业能力，一般需要通过知识工程抽取的领域专业知识、领域专家的经验反馈和行业价值观等，为大模型注入专业知识并实现价值对齐。

三是提升大模型生成内容的可信，一般是通过检索增强生成技术和工具学习，来实现大模型在内容生成过程中与外部专业知识库和外部专业工具/模型的协作。

四是基于智能体应用范式提升复杂任务处理可信，通过智能体对复杂任务进行规划分解、子任务推理与核验、多智能体协同优化等，可显著提升大模型处理复杂任务的能力，以及推理过程中的可控性和透明性。

五是结合应用领域的风险分析与识别，实施相应的安全措施保障，搭建出可在专业领域落地的大模型可信应用系统。

此外，在系统上线应用前，还需要通过**全面、充分的评测**来验证大模型系统已经达到预期目标。最后，考虑到大模型技术仍在不断迭代，实际应用中发现的偏差也应快速反馈到大模型应用体系中，还需构建出**“反馈-迭代”的良性循环体系，**实现大模型应用能力在专业领域中的稳妥拓展和持续强化。

（二）大模型在专业领域可信应用的技术实现

1 数据供给质量提升

在大模型的应用过程中，高质量数据对于大模型在专业领域的可信应用至关重要。大体上可以把数据分为三类：

通识通用数据 **领域通用数据** **领域专有数据**

关于每类数据的描述和质量提升方法具体介绍如下：

（1）数据分类介绍

通识通用数据：一般是通过公开渠道可获取的无标注数据，数据规模大（一般包含数千亿至数万亿个词元②），内容范围广泛，覆盖多种主题和领域。数据来源一般包括网页数据、书籍文献、新闻数据、百科全书数据、代码数据等，主要用于大模型的预训练阶段，帮助大模型学习到通用的世界知识，增强模型通识能力。

②词元，即Token，是指大模型训练和生成内容时的处理单位，可以是字符、子词、单词或短语等。

领域通用数据：一般是特定领域中通过公开渠道可获取的数据集，可用于大模型的增量预训练、专业调优对齐或推理过程，可以帮助大模型学习到特定行业或领域的专业知识，增强模型专业能力，包括：

- 政府或行业监管公布的数据，特点是权威性高、更新频率稳定，例如国家统计局网站③等；
- 研究机构或社区发布的特定领域数据集，一般经过标注并开放出来供研究和开发使用，比如医疗领域的**Huatuo-26M**④等；
- 领域的专业书籍、文献、论文等，例如BookCorpus 免费书籍、已发表学术期刊论文等；
- 从互联网上可获取的相关领域网页、资料等，例如金融领域的上市公司财报、公司公告、投资研究报告和金融新闻等。

领域专有数据：一般是企业或组织生产运营过程中积累的数据，这类数据同样可用于大模型的增量预训练、专业调优对齐或推理过程，可以进一步提升模型的专业能力以及在企业中的落地应用效果。但是这类数据往往涉及机构和企业生产运营核心环节，相对难以获得。具体包括：

- 企业知识库数据，例如企业内部积累下来的产品文档、技术文档、研究实验数据、市场动态、运营日志数据等；
- 工作流数据，例如企业内部的工作流程和操作步骤、行业或领域专家解决问题的分析框架和解决方案步骤等；
- 服务过程数据，如机构或企业提供产品和服务过程中记录下的数据，如服务过程中客服与客户的对话数据、医疗诊疗过程中医生与患者对话数据、客户对产品服务的评价和反馈数据、服务过程中的客户行为偏好数据等；
- 内部构建的专业指令数据集， 一般是模型开发者为了更好地训练大模型在特定领域的专业指令遵循能力，而专门构建的经过标注的高质量指令数据集等，实践中宜同时包含正负样例的指令数据，可以使模型获得更好的表现。

③中国国家数据，https://data.stats.gov.cn/

①医疗领域开源数据集汇总链接，https://github.com/onejune2018/Awesome-Medical-Healthcare-Dataset-For-LLM

(2) 数据质量提升方法

已有较多研究⑤表明，提高数据集质量可以有效提升大模型的性能表现，并有助于减少幻觉现象，尤其是应用到专业领域时，更需要对领域通用数据和领域专有数据进行高质量处理，以尽量确保数据集所包含数据的真实、准确、合规、 一致且有较强时效性。如图7所示，提升数据质量的主要措施包括：

来源：公开资料整理

图7数据质量提升处理流程

一是增加数据的多样性和丰富度。 一方面，在模型能力支持的前提下，引入不同模态的数据如图像、音频、语音、三维数据等，通过多模态数据的统一表征、对齐和有效压缩，可以让大模型从多种感知通道中获取信息，形成更全面的认知，增强模型的泛化能力。同时，通过结合多种模态信息，模型可以更好地验证和校正自身的输出，减少错误和不准确的输出内容。除了模态多样性外，还可以从语言多样性、领域多样性、格式多样性等方面增加数据的多样性和丰富度。

另一方面，对于领域通用或专有数据集，在数据收集成本较高的情况下，也可以结合一些数据增强技术来丰富数据集，例如，在医疗图像数据集中，可以通过旋转、裁剪、缩放、亮度调整等几何方式生成更多的数据样本⑥。此外，也可通过开发专门的大模型①来合成高质量数据，但应用过程中需要注意解决合成数据中可能存在的偏差和质量问题。

⑤https://cacm.acm.org/news/data-quality-may-be-all-you-need/
◎《深度学习辅助决策医疗器械软件审评要点》，国家药品监督管理局医疗器械技术审评中心，2019年7月
①https://blogs.nvidia.com/blog/nemotron-4-synthetic-data-generation-llm-training/

二是实现更为全面完善的数据处理方式。一般包括数据清洗(即去除无效、错误和噪声数据，进行数据规整，处理缺失值及异常值)、数据去重(剔除重复或冗余数据，包括段落去重、文档去重、文档相似度去重等)、数据去毒(识别并移除有害或不适当的内容)、数据脱敏(识别并脱敏个人隐私信息等)数据整合(将来自不同来源的数据进行整合形成一个统一数据集)、数据标注(在面对准确率提升困难、接近知识密度天花板、或涉及人类偏好对齐的情况时，部分数据还需要人工进行更高质量反馈评分和精细标注)、数据平衡(通过采样、欠采样等方法平衡数据集中的类别分布，避免模型偏向某些类别)等。

三是建立持续的数据质量评估机制。首先，定义数据质量评估维度，可以从数据错误、数据遗漏、数据重复、数据时效、数据安全等多个方面，从规范性、准确性、 一致性、可靠性、完整性、全面性、多样性、重复性、及时性、安全性等多个维度定义数据质量评估指标；其次，在数据处理流程中建立数据质量监控点(比如数据引入阶段、数据加工后、数据消费前、模型上线后)，结合数据质量评估指标对数据处理质量进行有效评估，并根据评估结果持续优化，例如可定义验证规则并通过单元测试(Unit Testing)方式来监控和检查数据质量；再次，可引入自动化工具提升数据质量评估效率并降低成本，比如建立专用的质量评估算子库，训练专用的数据质量评估模型，并构建自动化评估链路等。

2 模型知识专业增强

大模型在面向专业领域进行应用时，还需要基于高质量的领域专业数据集，进一步通过知识工程有效提取出高质量数据和知识，并通过模型调优微调出面向该领域的行业或专业大模型。由于实施难度成本较高、需要领域高质量数据较多， 一般该过程由大模型提供方实现，如**图 8**所示。

图8大模型专业增强处理

来源：公开资料整理

(1) 知识工程

通过知识工程进行领域专业知识抽取和知识图谱构建。领域知识抽取主要是从非结构化的数据语料中挖掘出领域相关的实体、概念和关系等知识元素，常用的方法包括本体构建、命名实体识别、关系抽取、术语抽取、事件抽取、主题建模等。目前，也可结合大模型来加速领域知识的抽取效率，例如开放知识抽取引擎**OneKE**O，基于大模型实现了中英文双语、多领域多任务的泛化知识抽取能力，并提供了完善的工具链支持。

知识图谱构建是指将抽取的知识元素组织成结构化的知识图谱，主要包括：实体链接，即将抽取的实体链接到知识库中的实体；知识融合，即将从多源异构数据中抽取的知识进行去重、去噪、补全等处理，融合形成统一的知识图谱；图谱存储，即将构建的知识图谱持久化存储，支持高效查询和推理等。目前，也可结合大模型来增强知识图谱，包括帮助图谱进行命名实体识别与链接、知识库补全、丰富图谱知识表达⑨等。

领域专家经验挖掘是专业知识的重要补充。从领域专家的经验中挖掘隐含的知识是当前高质量数据集和语料库获得的另一渠道，这对于在一些任务复杂度高、容错率低的领域如金融、医疗等尤为重要。常用的方法包括：从专业书籍如教科书中挖掘专家分析框架和标准处理流程 (Standard Operation Procedure, SOP) 模式等；专家访谈，即通过直接与领域专家进行交流获取其专业知识和见解，并生成专家规则；从服务过程数据中挖掘，例如在医疗诊疗过程中医生与患者的对话数据中来提炼有价值的知识和经验；从工作流挖掘，即从工作流数据中提取出业务流程的模型和规律，形成相应规则等。

(2) 专业知识注入

一般是指在通用大模型基础上，结合着领域数据集和领域知识图谱等，通过增强预训练、有监督微调等方法把专业数据和知识注入到大模型中，是提升其专业能力的关键步骤。当前主要技术包括以下几方面：

大模型增量预训练 (Continuous Pre-training,CPT)，把领域数据集与通用数据集混合，使用这些数据继续训练原有的通用大模型，从而提升模型的领域专业能力。例如，金融领域彭博社混合了其专有的金融数据**集FinPile** 和公开通用数据集（如**C4、The Pile**、维基百科等），训练了其金融领域专业大模型**BloombergGPT**。

@http://oneke.openkg.cn/

③S.Pan,L.Luo,Y.Wang,C.Chen,J.Wang and X.Wu,"Unifying Large Language Models and Knowledge Graphs:A Roadmap,"in IEEE Transactions on Knowledge and Data Engineering.

大模型有监督微调（Supervised Fine-Tuning,SFT），在已有的大模型上，利用高质量、标准化构建和标注的专业领域指令数据集（每条指令数据一般包括指令、输入和预期输出三个要素）来进行大模型的训练调优，这种方法可以有效增强模型的专业能力和指令遵循能力。

大模型领域知识图谱增强，在大模型预训练或专业微调优化步骤中，可以把领域知识图谱中的专业结构化知识整合到预训练或监督微调数据集中，使模型能够直接从图谱中学习专业知识。例如，医疗知识图谱中的三元组（非小细胞肺癌，治疗药物，吉西他滨）可以转换为"非小细胞肺癌可以使用吉西他滨这种药物进行治疗"，通过这种方式把图谱中的结构化专业知识转化为训练语料，并与其他语料数据集相结合。

(3) 价值对齐

价值对齐可以建立大模型与人类价值观的约束对齐，使得模型输出更符合人类的伦理和社会规范，避免产生有害或不适当的输出。主要措施包括：

专家反馈强化学习，把大模型应用到专业领域时，领域专家的高质量经验和知识反馈尤为重要，可以利用专家提供的反馈数据来构建奖励模型，并通过强化学习优化大模型的行为。此外，由于收集专家的反馈成本较高，也可以利用Al生成的反馈数据来训练大模型，即AI**反**馈强化学习（Reinforcement **Learning from AI Feedback,RLAIF）。RLAIF可**以减少对高质量专家反馈数据的依赖，降低数据收集和标注的成本，且AI生成的反馈数据具有更高的一致性和可重复性， 一定程度也可减少人类主观偏见带来的不利影响。

直接偏好优化 (Direct Preference Optimization,DPO)①，直接从人类反馈中得出的偏好数据来计算损失函数，并指导优化过程。相比近端策略优化等算法，**DPO**可以简化训练过程，可以直接与人类偏好做对齐，在保留较好效果的同时减少训练所需计算资源和时间。不过，人类偏好数据集的质量对DPO 的性能有显著影响，不准确或不全面的偏好数据会影响模型训练结果。

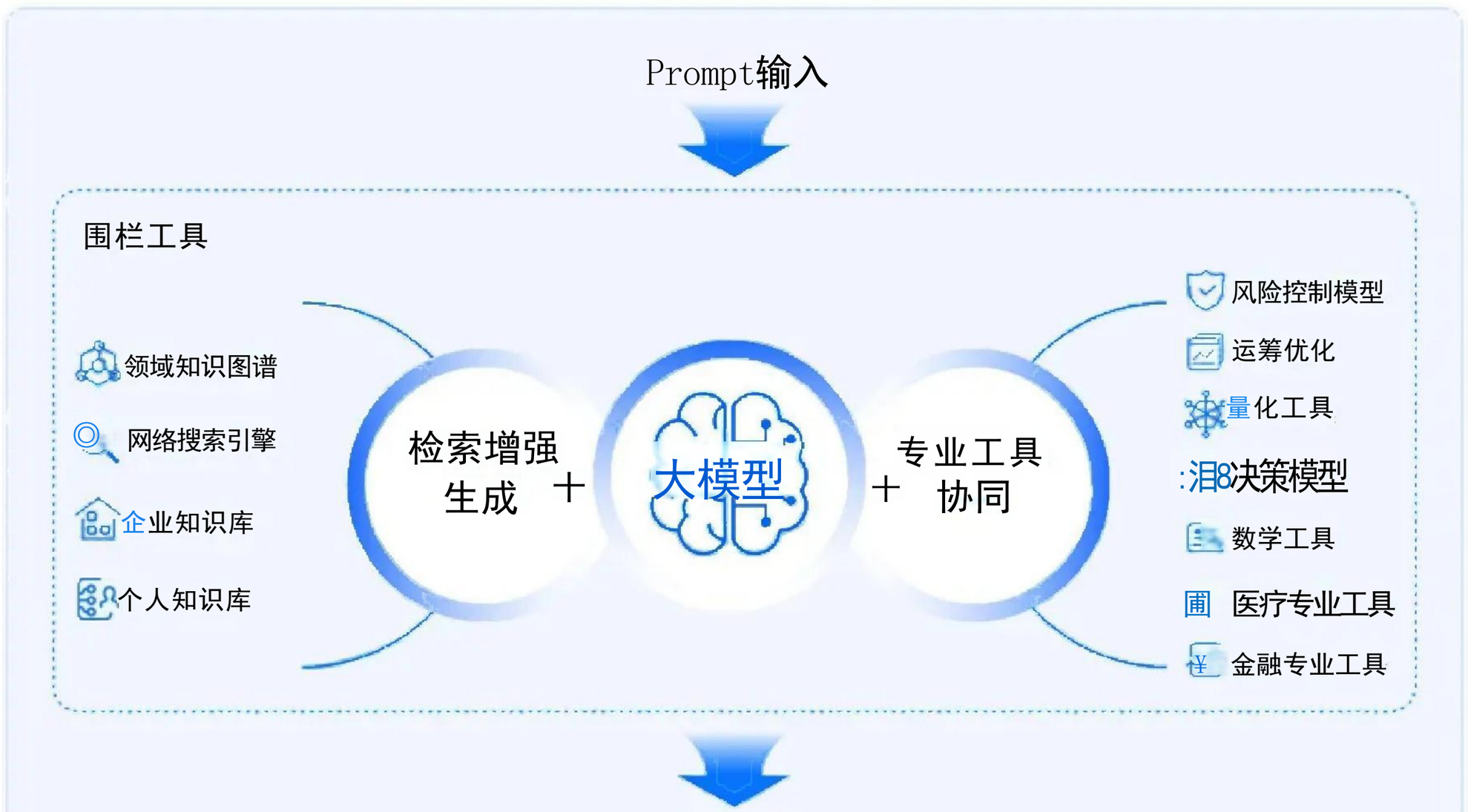
在应用到专业领域时，还需要考虑结合应用领域的特有科技伦理和价值观，对大模型在专业领域内的行为和决策进一步进行对齐约束，即**专业价值对齐**。例如，金融领域价值观一般包括“诚信、专业、稳健、创新、担当和普惠”等，医疗领域价值观一般包括“人道主义、尊重生命、公平、公正和保护隐私”等。在实施过程中，可先把专业价值观转化为机器可理解的形式、规则和约束条件等，再利用前述强化学习方法进行专业价值观对齐。

OpenAI,Weak-to-strong generalization,https://openai.com/index/weak-to-strong-generalization/
Rafailov,Rafael et al.,Direct Preference Optimization:Your Language Model is Secretly a Reward Model, https://arxiv.org/abs/2305.18290.

此外，最好可以为大模型提供较为详细的模型描述文档，当前主要方案是模型卡(Model Card)，主要内容可包括模型参数和架构、训练数据集构成、模型的性能评测情况、模型的预期应用场景及其局限性等信息，从而可提升模型应用的透明度，减少模型的滥用或者误用等。

3 内容生成可控可信

经过微调的行业大模型虽然已具备一定的领域专业知识能力，但是在更加细分和专业的场景落地使用时，他们内置知识和能力总会存在边界。主要原因**一是**最新的领域动态信息或知识难以及时更新到行业大模型中；**二 是**出于数据安全等问题，企业的内部专有知识或敏感信息也难以通过训练内置到模型中；**三是**一些专业领域的计算或任务处理上，大模型的任务完成效率和准确度上也不如已经过业务充分验证的专门工具库或者专业模型。因此，有必要在大模型对输入指令进行内容生成响应过程中，通过工程化的解决方案，结合外部专业知识库和专业工具库来提升生成质量。从技术实现上来看， 一般会结合检索增强生成、工具学习和围栏工具来实现，如**图 9**所示，具体介绍如下。



来源：公开资料整理

图9结合外部专业知识和工具实现内容生成可信

(1)检索增强生成

检索增强生成(Retrieval-augmented Generation,RAG)是把外部知识检索与大模型内容生成相结合的技术，其核心思想是通过检索与当前输入请求相关的外部知识，包括领域知识图谱、网络搜索引擎、企业和个人知识库等，并将其作为额外的上下文信息引入到模型生成过程中，从而可以提升其内容生成的可控性、专业性、准确性、实时性以及可解释性等。

目前**RAG**正逐渐成为大模型在产业应用落地过程中被广泛采用的技术，并已实际应用到多个领域和场景中，如问答系统和对话生成、搜索引擎内容生成等。以与领域知识图谱的结合为例，其价值在于**一是**图谱中的结构化知识可以为大模型提供更为清晰的上下文信息，**二是**图谱中的关系链接可以帮助大模型在一些复杂推理内容生成(如多跳推理)中更有效地综合多源信息；**三是**通过知识图谱，大模型生成的内容可以更容易地追溯到具体的知识源和推理路径，提升内容生成的可解释性和可靠性。对于大模型来说，在调用外部知识源时需能够将自然语言转换为结构化查询语言，例如N12SQL、N12Cypher、N12GraphQL等，再对相应的知识源进行查询。

在实际落地过程中，RAG 对于大模型生成内容的改进效果，很大程度上取决于RAG 外部知识库准备、知识索引、知识检索和模型优化等主要步骤的质量，实现过程中需注意：**一是**知识库准备过程中，需要尽量提升外部专业知识的质量，减少出现知识错误或者知识冲突的情况，并能够及时把专业领域的最新知识更新到库中；**二是**在知识索引过程中，关键是对于知识分块(Chunk) 的优化，可以通过滑动窗口、索引文档摘要、增加额外元信息 (Metadata)、 使用粗粒度和细粒度相结合的多级索引方案等来改进；**三是在**知识检索过程中，可以结合查询(Query) 优化、混合检索、对检索结果重排序或者压缩等方法，提升检索出的上下文质量；**四是**在模型优化过程中，可以有针对性地对大模型RAG 生成能力进行优化，如进一步增强模型对输入上下文的忠实度、反事实鲁棒性、噪声鲁棒性、拒答能力等。此外，也需要建立比较完善的RAG 实现效果评价机制，**对** RAG 的实现效果进行评估，在识别出性能瓶颈后有针对性地进行改进。

(2) 工具学习

工具学习 **(Tool Learning)** 是指对大模型进行“工具学习”训练②，让它学会识别何时以及如何调用外部专业工具。学习过程一般包括：**首先是**用户意图理解和指令分解，即模型能够通过输入指令理解用户意图，并将其分解为若干子任务，并判断哪些任务应由模型自身完成、哪些任务需要调用外部专业工具；**其次**是训练大模型理解外部工具的功能，并能够根据任务选择合适的外部专业工具；**再次**是训练大模型学会如何正确调用和使用外部工具，包括API 调用、参数赋值等，并根据这些工具的API接口来构建请求和处理响应，从而获取所需的信息或执行特定操作。

通过训练大模型学会如何使用领域的专业工具库或者垂直专业模型，可以实现大模型+专业工具/模型的协同，从而提升大模型在专业领域的生成内容可控性和专业性。例如，在金融领域，通过与原有金融专业工具库的协同，可以提高大模型处理和分析这些复杂金融数据的能力，同时也可以利用专业工具库内置的安全和合规功能，确保数据处理过程符合金融领域的行业标准和法规，提升安全合规性。同时，经过多年业务实践，金融机构也已经沉淀了很多经过业务充分验证的专有模型如反洗钱模型、风控模型、投资组合优化模型等，这些也可以与大模型结合起来，发挥各自所长。

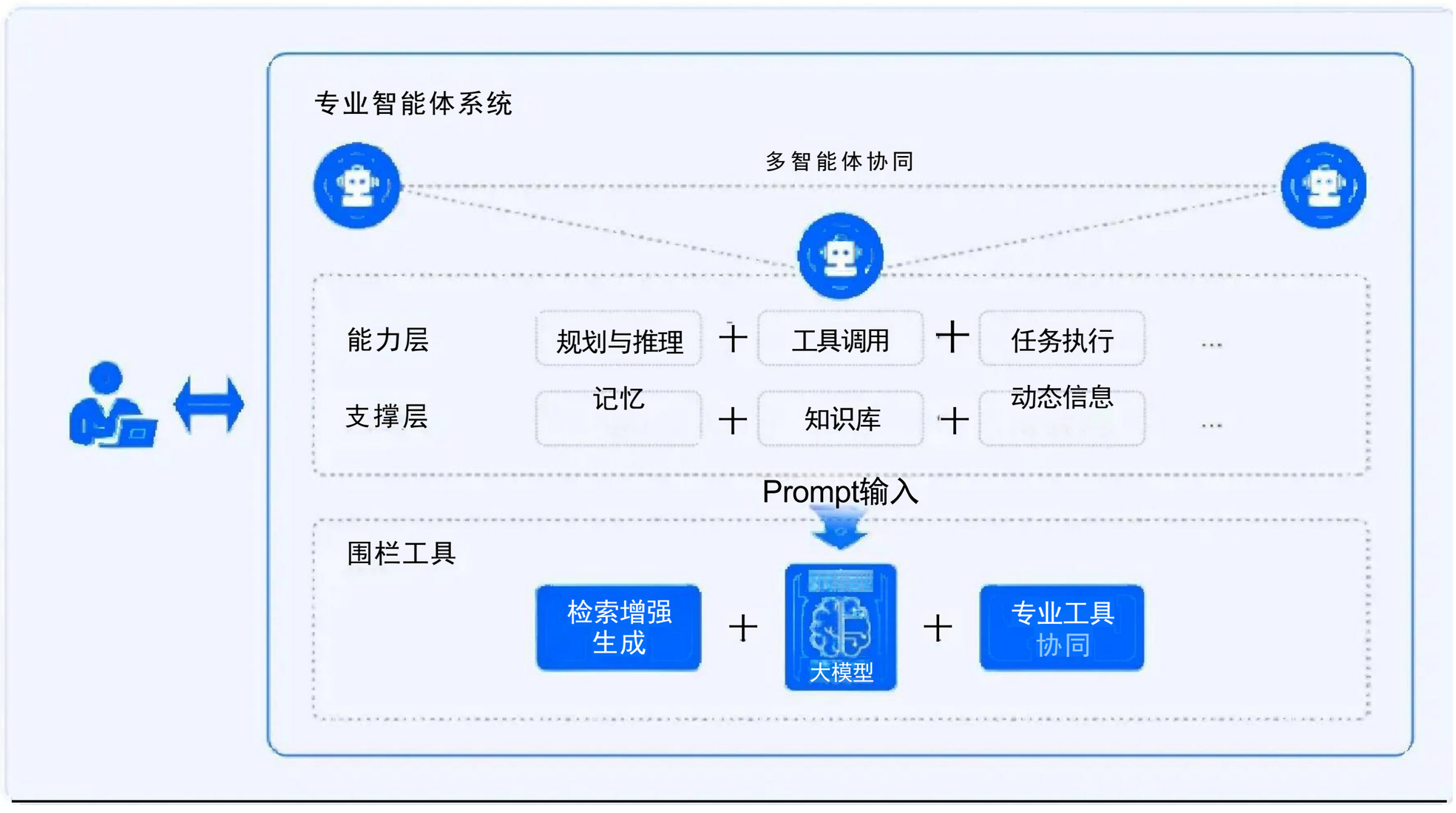
(3) 围栏工具

围栏工具相当于在大模型外围加上了一个“防护盾”，可以从三个方面提供价值： **一是**对输入请求进行精准的意图识别，可以帮助大模型挡住外界的恶意提问，或者过滤掉不符合系统可接受范围内的输入请求，确保大模型不会处理和响应不适当或有害的内容；**二是**对模型输出的生成内容进行监控和审查，并设置拦截或安全改写机制，确保输出内容不包含有害信息(如涉黄、涉暴或者包含敏感个人信息等)，符合安全合规要求；**三 是**还可提供安全攻防能力，帮助大模型抵御受到的外界恶意攻击。

⑩T Schick,JDwivedi-Yu,R Dessì,R Raileanu,M Lomeli,E Hambro,L Zettlemoyer,“Toolformer:Language models can teach themselves to use tools”,Advances in Neural Information Processing Systems,2024.

4 智能体提升复杂任务处理可信

智能体应用范式③已经是当前大模型的重要应用模式，可以在以大模型为核心“大脑”的基础上，结合着规划与推理、工具调用、任务执行、多智能体协同等能力，显著提升大模型系统在复杂任务时的能力表现，以及大模型推理过程中的可控性和透明性，为任务处理提供保障。



来源：公开资料整理

图10基于智能体实现复杂任务处理可信

（1）智能体范式拓展大模型可信应用范围

智能体扩宽了大模型可执行任务的范围，将大模型的使用方式从局部、单一的生成和问答任务转向现实世界更加复杂的任务，其关键能力模块包括：

Zhixuan Chu,Yan Wang,Feng Zhu,Lu Yu,Longfei Li,Jinjie Gu,Professional Agents --Evolving Large Language Models into Autonomous Experts with Human-Level Competencies,https://arxiv.org/abs/2402.03628.

规划与推理：

智能体在对推理任务进行感知理解后，可以将推理任务分拆成一系列可管理的子任务，分拆过程中可通过反思改进、用户交互等方式对任务分拆的合理性、颗粒度和完备性等进行评估和调整。在专业领域应用时，智能体可以采用专家经验SOP 增强的任务规划与推理，以确保推理任务分拆的专业合理。例如，在把大模型应用到金融、医疗等专业领域时，需与行业专家共创提炼出专业的决策框架及行业最佳实践，并约束智能体在规划推理时遵循。

工具调用：

智能体在与环境交互的过程中可以将复杂任务分解为子任务，通过构建大小模型协同链路，合理选择API、图像处理库、数学计算库、代码解释器、搜索引擎等工具来完成相应子任务，提高智能体在复杂业务场景中的效率。智能体工具调用能力涉及到任务识别与分解、工具选择与配置、数据准备与传输、结果获取与整合等多个方面，是拓宽智能体应用场景的关键。

任务执行：

智能体任务执行能力可以将智能体的推理结果转化为具体的操作，并执行具体的任务，比如与物理环境进行交互等。智能体在执行任务时通常需要与用户或其他智能体进行交互，以获取进一步指令、反馈信息或协同完成任务。任务执行完成后，智能体会对任务执行结果进行评估，这包括检查任务是否成功完成、评估任务执行效率和质量等方面。基于任务评估结果，智能体会对自身的知识库、算法模型和执行策略进行优化改进，进一步提高智能体在未来任务执行中的效率和准确性。

由上可见，智能体可以提高推理过程的专业性、可靠性、可控性和可解释性，能够为大模型在专业领域的推理可信提供基础保障。例如，在评测大模型代码生成能力的HumanEval 数据集上，GPT-3.5 （零样本）的正确率为48 . 0%, GPT-4 （零样本）的正确率为67. 0%, 远远高于GPT-3.5 。 如果在GPT-3.5 上搭配智能体，GPT3.5的推理表现可超过GPT4④。

(2) 智能体通过人机交互进一步确保决策质量和安全

智能体在执行任务的过程中，可以通过人机交互提高工作效率、增强用户体验，并提高智能体的决策质量和安全。**一 是**智能体通过交互式学习可以实现优势互补。人类在模糊概念理解、创造性思维、情感判断等方面具有特定优势，智能体在数据处理、数据计算、推理决策等方面更具优势。通过交互式学习，智能体可以逐步积累更多的人类经验，提高决策的可信程度。

https://www.youtube.com/watch?v=sal78ACtGTc

二是人机多模态交互提高交互的准确性和效率。多模态的交互方式，比如通过文本、图像、视频、动作等，使得智能体能够更好地理解用户的意图和需求。同时，智能体能够根据用户的行为习惯和环境变化调整其交互方式和决策策略，展现出强大的适应性。

三是通过人机交互可以更好地对齐人类价值观。智能体在与人类交互的过程中，可以逐步学会理解、尊重人类认可的一系列道德、伦理和社会原则，提高决策推理的安全可信。此外，伴随着智能体系统的自我优化和可信应用水平提升，可以减少人工干预和交互。自我优化减少了对人工的依赖，降低了维护成本，提高了智能体的自主性和智能化水平。

(3) 多智能体协同进一步提升推理可信程度

类似于人类社会，通过专业化、精细化的分工与协作，可以突破个体能力的局限，多个智能体也可以通过相互间的协作、互动和反馈，可以在专业领域应用过程中综合多个专业智能体的优势，避免单一智能体的局限性，从而进一步提升整体系统的专业严谨、可控透明。

在多智能体协同框架中，不同智能体可以被赋予不同的专家角色，例如任务规划智能体、任务执行智能体和核验智能体等。其中，任务规划智能体负责对推理任务进行规划和编排，任务执行智能体负责具体的任务执行，并在执行过程中调用外部工具或专业知识库。核验智能体则可以进一步细分为子任务核验智能体、推理核验智能体、隐私数据检查智能体和内容合规检查智能体等，负责在推理过程中对任务规划和执行的各个环节进行验证和核实，指出不足并提出改进反馈。

通过这种多智能体协同机制，可以实现： **一是**专业化分工，不同智能体专注于特定的任务或领域，可以最大化其专业知识和技能水平，提高整体系统的效率和准确性；**二是推**理过程中实现自核验，在任务执行过程中，核验智能体可以监控、评估和核验执行智能体的执行结果可信程度，并分析原因和提供即时反馈，帮助规划智能体做出及时的改进措施。

5 新安全范式保障可信

需要针对大模型应用带来的新风险引入新的安全范式。当前大模型执行的工具调用和数据访问普遍缺乏身份和凭据透传，也就是说，用户只要通过了大模型的访问权限校验，就可能通过大模型畅通无阻地对其他专业工具或数据库进行调用或访问，这可能带来较大的安全漏洞。

因此，有必要对大模型应用过程带来的新安全风险进行全面分析，并有针对性的提出相应安全措施进行缓解。可以依托安全平行切面技术⑤,通过原生安全范式引导并建设大模型专业应用的基础底座安全保障，具体而言：

运用智能体切点植入技术，提升基于身份的访问控制、工具调用，Prompt **会话交互等数据内视能力 。**要完整研判一个访问是否合法，应该基于该访问的操作者 **(Operator)** 链路和凭证 **(Voucher)** 链路。操作者链路保障了真实操作者身份的透传，而凭证链路进一步确保当操作者具备访问权限的时候，只有在具备凭证的前提下行使该权限，做到可追溯，可审计。如**图** 11所示，是为保障大模型安全应用提出的基础安全范式OVTP 可溯范式(Operator-Voucher-Traceable Paradigm)和NbSP 零越范式(Non-by-**passable Security Paradigm)。**

安全平行切面白皮书，http://www.itstec.org.cn/aspect_oriented_security_white_paper.pdf

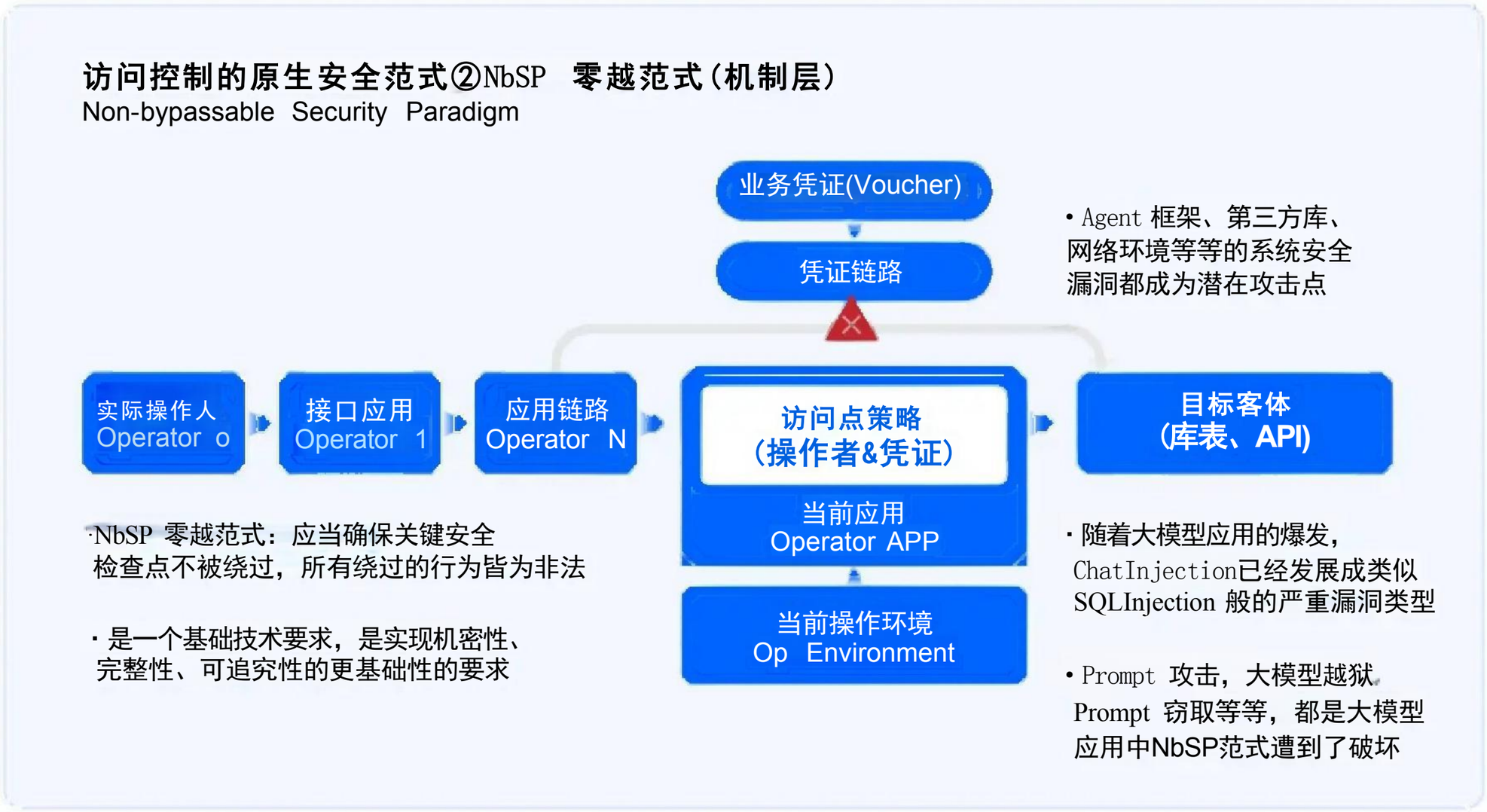


图11大模型应用原生安全范式OVTP 和NbSP

来源：蚂蚁集团

其中，在**OVTP**可溯范式指导下，可实现访问控制点应追溯，保障每一次大模型请求都可以追溯到实际操作者，避免攻击者通过智能体代理进行恶意操作后身份无法识别的窘境。随着大模型应用的功能增加，整体调用链路会越来越复杂，OVTP 范式可以确保在多智能体协同过程中的整体调用链路是可信的，以及权限不被滥用。同时，为了保证关键访问控制点不能被攻击者绕过，基于NbSP 零越范式，可以在大模型专业应用关键链路上设计和布置访问控制点，比如在大模型输入、或者智能体交互的时候设置检测点，识别类似**Prompt** 注入攻击的安全风险⑥，同时也可以在大模型输出的时候设置检测点，保障敏感数据不被泄漏。

6 反馈迭代提升体系

为实现大模型在专业领域中的稳妥应用和持续强化，构建“反馈-迭代”的良性循环体系非常重要，使得大模型的研发运营体系实现持续的改进和优化： 一是建立高效的反馈机制，确保从模型生产、部署到实际应用的每一个环节都能实时收集和分析有价值的数据，如用户行为、模型性能和推理运行指标等，以随时掌握模型运营情况，及早发现问题。二是结合持续集成、持续部署、持续训练等流程，提高快速响应变化的能力，及时进行模型迭代更新，提升模型稳定性；三是通过体系的建设和管理，不断提高过程管理

LangChain 注入攻击漏洞.https://nvd.nist.gov/vuln/detail/CVE-2023-36258.

能力，减少过程风险，提升模型过程可信。如**图 1 2 所**示，大模型在专业领域的“反馈-迭代”机制可包括：

数据和案例反馈：建立应用过程中的服务过程数据和场景应用效果(包括成功案例、失败案例)的反馈机制， 一方面可以帮助模型开发者和应用者分析大模型在实际应用中的表现，识别问题和改进点；另一方面，可以在对反馈数据进行脱敏后，整理成高质量的数据丰富到领域数据集中。

模型和领域数据集优化：根据反馈的应用案例和领域数据，可以分别对大模型能力、整体应用系统、领域专业知识库、领域评测体系(包括评测数据集、评测任务和指标等)等进行优化和升级，提升大模型在专业领域应用的可信程度。

评估和验证：在优化后的大模型可信应用系统上线前，仍需要进行全面的测试和验证，确保其在不同场景下的可信应用程度符合预期目标。

场景拓展和深化： 基于可信应用程度的不断提升，稳妥推动大模型应用在专业领域的场景拓展，进一步深化应用到任务复杂度更高、容错率更低的场景中。

来源：公开资料整理

图12大模型可信应用的“反馈-迭代”机制

7 应用系统评测验证

构建全面、充分的面向智能应用系统的评测体系，对于专业领域的可信应用也尤为重要。评测可以评估当前模型的技术能力、助力产品研发、支撑行业应用，是促进企业内部智能水平和可信能力提升的重要手段。企业自身应该构建覆盖大模型“建用管”全生命周期的评测体系，建立契合业务特性要求的高质量评测数据集，采用主客观复合的测试方法，对大模型进行全面、客观的评测。除此之外，还可采用第三方中立的评测体系来验证大模型在专业领域中的表现。例如中国信通院从指标体系、测试方法、测试数据集和测试工具四个维度出发，建立“方升”大模型基准测试体系，重点面向产业应用效果进行评估。

大模型技术的快速发展和应用，也给大模型的评测带来较大挑战，具体包括：**首 先 ，**大模型技术的快速发展和对多任务处理能力的提升，使得如何划定评测的任务边界成为挑战。**其次，**评测体系需要设计合适的评估指标和构建高质量的评测数据集，要全面评估大模型的性能，需要花费大量的时间和资源。**再次 ，**大模型技术更新速度快，评测体系需要不断更新和优化，以适应新的技术和应用场景。**最后，**面 向 专业领域应用过程中，还需要在通用测评的基础上，构建针对专业领域应用场景的垂直评测基准，不同专业领域有其独特的需求和挑战，需要大量的行业领域知识和经验来指导评测基准的设计与构建。

来源：中国信息通信研究院

图13大模型可信应用评测验证体系示意图

针对大模型在专业领域的评测体系，可采用“模块化评测”+“端到端评测”的思路，如**图 1 3**所示。其中：

模块化评测：针对大模型可信应用系统中的各个主要模块进行评测，模块化的评测更为适用于识别系统瓶颈，并有针对性的进行开发改进。在大模型可信应用系统中，主要的评测模块任务分为安全保障、大模型、智能体三类，具体评测任务详见**图13所**示。

端到端评测：是从真实的专业领域应用出发，针对典型应用场景设计出完整的测试任务(一般是需要进行深度推理)并对系统进行评测，端到端的评测更为适合专业领域应用方对大模型可信应用系统进行整体评估、对比和验收。例如，在医疗领域中目前已出现了结合临床诊断场景设计的端到端评测基准@,可以用来评测医疗大模型在诊断和治疗任务上的深度推理表现，测试任务不再局限在医学问答或医学信息提取等相对较为简单的任务上。

@Lei Liu,XiaoyanYang,Fangzhou Li,Chenfei Chi,Yue Shen,Shiwei Lyu Ming Zhang,Xiaowei Ma,Xiangguo Lyu, Liya Ma,Zhiqiang Zhang,Wei Xue,Yiran Huang,Jinjie Gu,Towards Automatic Evaluation for LLMs'Clinical Capabilities:Metric,Data,and Algorithm,https://arxiv.org/abs/2403.16446.

04

大模型可信应用框架助力千行百业智能化转型

（一）大模型助力金融场景智能化转型

大模型在金融领域的主要应用场景包括智能客服、智能营销、智能投资顾问、金融市场情绪分析、金融市场分析预测、风险管理、合规和反洗钱等。此处选择了智能投顾场景作为典型场景来分析大模型可信应用框架在金融领域的落地实践，具体如下。

1. 应用挑战

一是金融投资市场信息量大、更新快，应用过程中大模型需要能够快速理解和处理这些实时数据，同时又因存在较多的冲突和噪声，需要大模型并在解决信息冲突和过滤噪声后，提取出有价值的信息并提供给用户进行决策支持。

二是用户投资决策是在信息不对称情况下做出，大模型在给出投资建议时，需能够根据客户需要详细解释推导过程以及所依赖的信息来源依据等，并确保符合金融领域的合规性和适当性等要求，帮助用户在金融市场信息不对称情况下更加理性地做出合理决策，增强用户的信任感。

三是数据安全与隐私保护，智能投顾场景中可能会需要收集和处理大量敏感的个人和财务数据（例如在对客户风险偏好进行分析时），大模型在处理相关敏感数据时必须确保数据安全，防止数据泄露和不当使用。

2. 实践方案

如 图 1 4所示，以某单位面向金融专业人员的智能投顾助理为例，目前已经可以用到金融知识挖掘、资产市场行业分析、新闻政策事件解读、公告研报财报解读、量化代码生成、财经稿分析报告撰写等场景中，帮助提升金融服务链条各职能专家专业水平和生产效率。具体可信应用框架的实现措施介绍如下。

金融知识挖掘	资产/市场/行业分析	新闻/政策/事件解读	公告/研报/财报解读	量化代码生成	财经稿/分析报告撰写

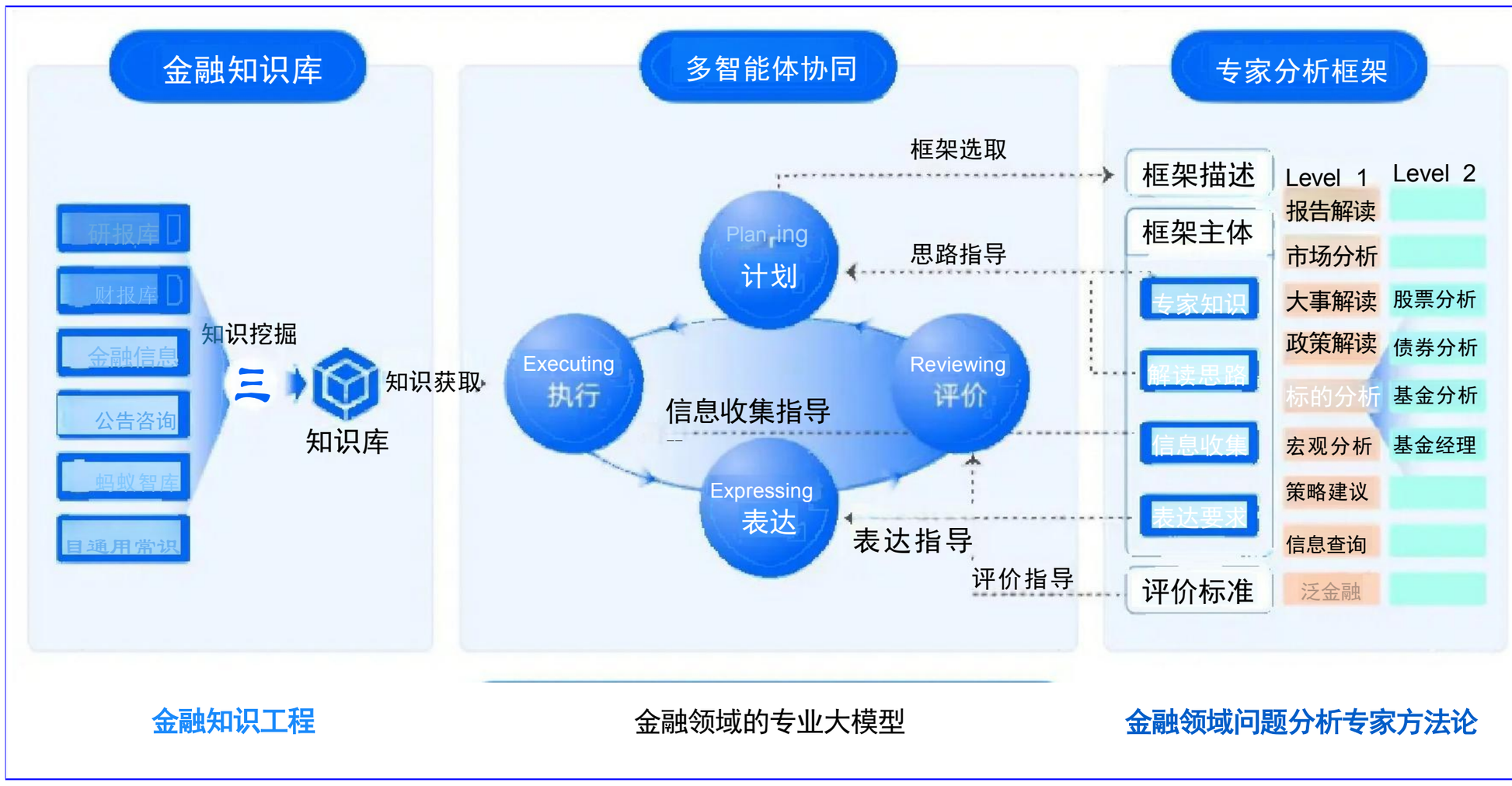


图14智能投顾助理场景中的应用

构建高质量的动态金融专业知识库。首先该系统整合了各种来源的投研报告、公司公告、新闻资讯、行情数据等，目前已形成了包括研报库、投报库、金融信息、公告资讯等多个知识库，同时通过多种知识挖掘链路来过滤冲突信息和数据噪声，提取到实时金融市场知识并动态更新到金融知识库中。**其 次 ，**还邀请金融专家对金融领域中典型任务整理出问题分析专家方法论，沉淀多类典型定性分析场景和多个细分专家框架，用于后续问题拆解、知识运用、推理核验和合规表达过程中。

应用多种方法增强模型专业性，并利用金融知识库和专业金融工具库，提升模型生成内容的真实可控。结合高质量千亿级**Token** 公开金融知识语料库和百亿级**Token** 企业专有金融语料库并采用直接偏好优化算法等，优化大模型的金融专业能力。在此基础上，通过检索增强生成（RAG） 技术与动态金融专业知识库打通，并通过多种改进策略提升RAG 优化效果能够达到金融领域业务指标的要求。同时，打通与企业目前已沉淀出的现有多个专业金融接口工具，包括行情解读、产品评估、行为分析等，进一步提升模型的生成内容可控性。

创新金融专家多智能体协同框架，实现投资研究分析过程的透明可控。通过仿金融专家的工作流程，将复杂的问题抽象为计划(Planning) 、执 行(Executing) 、表 达(Express-ing) 和评价 **(Reviewing)** 四个环节，并构建出相应的金融专家多智能体协同框架。在处理特定金融问题时，“计划”智能体首先梳理问题的核心要素，并规划出解决问题的策略(可基于已积累的问题分析专家方法论)；“执行”智能体随后对这些策略进行深入的信息搜集和逻辑推理；“表达”智能体将复杂的数据和逻辑推理结果转换成易于理解和结构专业的报告；“评价”智能体最终确保输出内容的质量，验证分析结果是否达到了预定的专业标准。

推出专业金融领域任务测评集，对大模型在金融领域的可信应用程度进行充分评估。推出**FIN-EVA**金融领域中文语言专业数据评测集，覆盖财富管理、保险、投资研究等多个金融场景以及金融专业主题学科，目前已开放的总评测题数目达到**1.3万**条以上⑩。测试任务涵盖了金融、经济、会计和认证的学科知识，以及大模型在多金融场景中的金融认知、金融知识、金融逻辑、内容生成以及安全合规五大类能力，可以对大模型在金融领域的可信应用程度进行充分评测。

数据来源：FIN-EVA**公开数据，**https://github.com/alipay/financial_evaluation_dataset

3. 应用成效

目前该智能投顾助理产品已覆盖中美港市场的重要上市公司与基金产品、800 余行业板块、200 余热点事件以及百余家核心独角兽公司，实现对金融市场动态的分钟级分析响应和多角度分析，能大幅提升投研分析师、理财顾问的工作效率和效果。具体应用示例如图15 所示，可见政策解读中不仅能够给出专业、真实的结果，而且解读的推导过程透明可控。

政策解读示例

提问：央行下调LPR将对股市带来怎样的影响?

专家框架匹配
Level 1　政策解读
Level 2　货币政策

智能体主动策划

问题细化
?央行下调LPR的具体时间和政策内容
?央行下调LPR政策的主旨 智能体自主决策
关注时间上的影响变化
?央行下调LPR政策的工作目标
?央行下调LPR政策与上一期政策的具体变化
?央行下调LPR政策前后股票市场的历史表现
?类似货币政策出台前后股市的市场表现
?央行下调LPR会影响哪些行业的服票
?央行下调LPR对投资者心态的影响
?央行下调LPR对未来股市走势的预期影响

智能投顾助理

中国人民银行近期下调了贷款市场报价利率(LPR)，特别是对5年期以上的LPR进行了自2019年LPR改革以来的最大幅度调整，从4.2%降至3.95%,降幅达到25个基点。这一政策调整被预计将对股市产生积极的影响。主要原因在于LPR的下调有助于降低实体经济的融资成本，从而激发经济的有效需求。

具体来说，这次政策调整的背景涉及多个方面，包括央行领导预告的降准和下调支农支小再贷款、再贴现利率，以及《金融时报》提及的贷款利率下调空间，旨在刺激消费和购房，同时考虑到CPI和PPI持续低位的状况。

这次LPR 的调整策略不仅反映了对当前经济形势的积极应对，也是在努力增强中国货币政策的自主性和提高对外部冲击应对的韧性。特别是5年期以上LPR的下调幅度超出预期，这被认为是一种强烈的政策信号，显示出央行支持长期资产，如股权资产和房地产价格的决 心。这种非对称的降息方式，被视为历史上“最不对称”的降息，从而释放出金融向实体经济让利的明确政策导向。

从政策的主旨和工作目标来看，央行的这一调整旨在稳定宏观经济和楼市，推动经济回稳向上。这次大幅度的LPR 下调打破了LPR 与中期借贷便利（MLF）　利率之间的线性关系，意味着后续对LPR走势的判断将变得更加复杂，需结合负债成本、风险和市场利率等多重因素进行综合分析。

政策对市场的影响预期是积极的，尤其是在当前全球经济形势复杂目不稳定的环境下。首先，通过降低企业和居民的融资成本，进一步促进社会融资成本稳中有降，有利于激发实体经济有效需求。其次，展现了货币政策的自主性和灵活性，有助于提高应对外部冲击的韧性。此外，LPR 的下调也有助于稳定经济增速，为保持合理的经济增速提供支持。总体而言，央行下调LPR 的举措，尤其是对5年期以上LPR的调整，将对股市带来积极影响，支持长久期资产的价格，并有利于中国经济的稳健增长。

图15示例-智能投顾助理在政策解读的应用

（二）大模型助力医疗场景智能化转型

大模型在医疗领域具有广泛应用潜力，主要的应用场景有临床决策支持、医疗影像分析、医疗文本处理、智能就医（导诊/问诊）、药物研发、医学科研等。在当前国家高度重视“互联网+医疗健康”背景下，此处选择智能就医场景作为典型场景来分析大模型可信应用框架在医疗领域的落地实践，具体如下。

1. 应用挑战

一是就医相关的信息完备度与准确性要求极高。模型必须达到专业标准，并且有询证逻辑，做到来源可信，需要掌握大量的专业医疗知识。 **二是**医疗不同场景下需要为患者提供差异化服务。在了解用户意图后，大模型除了进行正确分诊、挂号就医外，还需按照医生专业口吻给出建议，并兼顾到病人情绪提供病人安抚服务。 **三是需**在遵守相关的医疗法规和标准同时不影响模型性能效果。例如，医疗数据的敏感性要求模型在使用过程中的用户信息和关键知识库必须确保数据不被未授权访问和泄露。

2. 实践方案

如图16所示，以某面向患者的智能就医助理为例，通过在原有就医全流程客户端中应用医疗大模型和数字人等技术，为用户提供了智能挂号、 AR 导航、诊疗引导、互动服务等，解决当前患者在就医体验方面存在的**就医全流程服务复杂分散患者体验不佳、患者健康咨询有诉求但缺乏服务途径**等问题。具体可信应用框架的实现措施介绍如下。

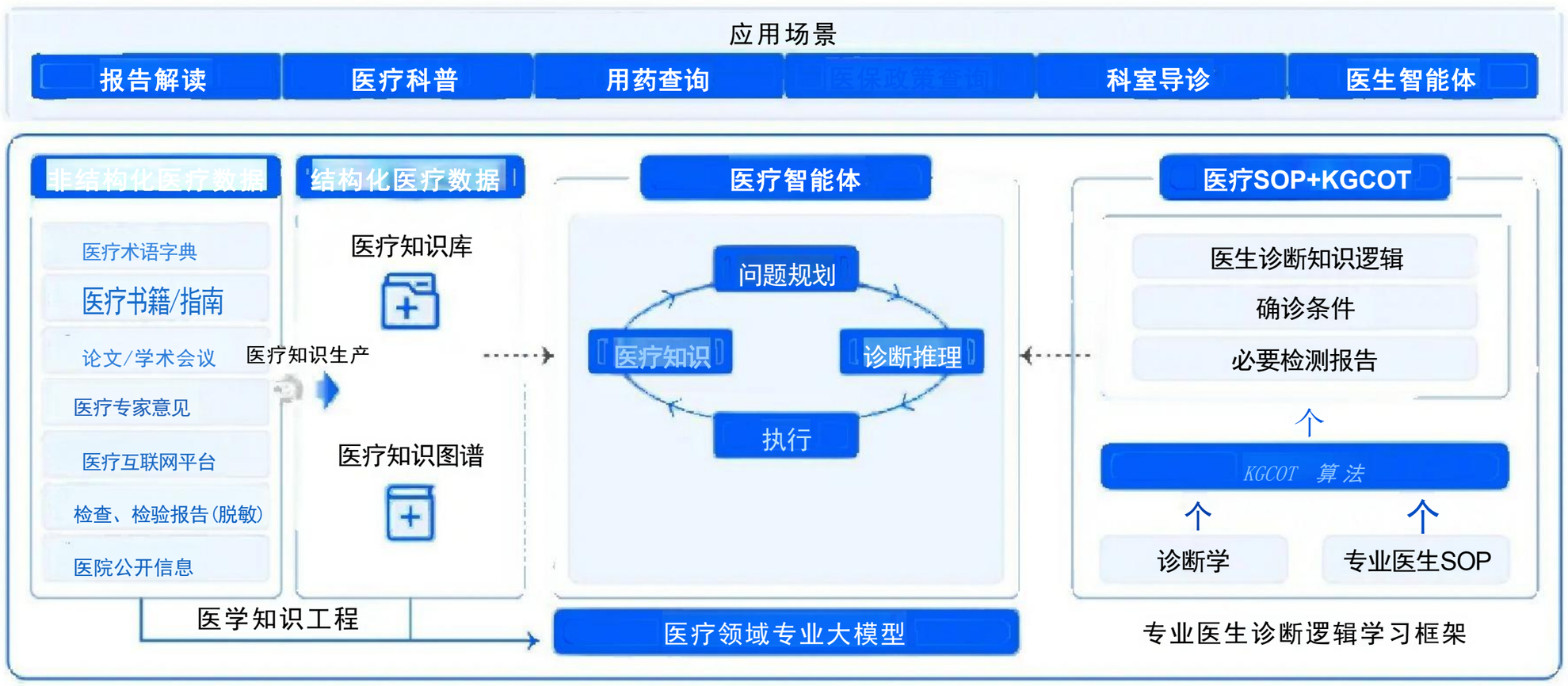


图16智能就医助理场景中的应用

构建高质量的医疗数据集，涵盖领域知识图谱、知识库以及垂搜内容。从业界权威数据源头获取数据，包括但不限于人民卫生出版社的相关数据、**pubmed** 论文与书籍、中华医学会指南、药品药械说明书等，通过高效数据抽取搭建出医疗专业知识图谱，涉及病、症、药、术、检、医生、医院等关键医疗信息，并结合内外部专家撰写、与医院打通等多种方式，构建出有专家背书的医疗知识库。同时，还实现了医疗领域专业全网搜索，相关网站来源均有认证，保障数据的实时性和准确性。

结合医疗知识库通过RAG 技术提升模型生成内容的真实可控。通过RAG 技术打通构建出的医疗知识图谱、知识库和全网垂搜能力，并对从三路检索召回的相关上下文信息进行再排序，以及对大模型针对性调优以增强模型选择正确性语料的能力，构造出真实可控的大模型回答生成结果，同时结果中对数据来源进行标识做到回答有据可循。

应用SOP 增强的推理过程实现复杂的医疗决策支持。以国内外大量医疗指南提供的标准操作流程作为基础，保障诊断推理逻辑过程的专业、可控，同时结合诊断交互过程可依据经验不断动态调整的特性，做到与具体诊断任务的灵活适配。推理过程中可将知识图谱与思维链相结合，并基于用户检查报告中的异常指标和症状(可通过多模态识别进行分析)，进行可控的疾病诊断推理。

推出专科医疗人工智能评测数据集，对大模型在医疗领域的可信应用程度进行充分评估。RJUA-AQ是模型研发方与医院合作创建了专科医疗人工智能数据集⑩，并为保障数据集的专业和权威性，做了以下设计与考量：数据集包括单轮临床问答、多轮诊断对话以及临床诊断推理三方面，其中单轮临床问答的数据来源于临床专家根据咨询经验编写的虚拟案例，使得数据集更加逼真，并确保数据隐私；多轮诊断对话的问题涵盖泌尿学的多个方面，占所有泌尿疾病的95%；临床诊断推理，数据集提供了详细的专业证据和推理过程。在此基础上，可以对大模型在医疗领域的可信应用程度进行充分评测。

RJUA-AQ项目地址：https://github.com/alipay/RJU_Ant_QA。

3. 应用成效

目前该智能就医助理已在浙江全省92 家医院使用，服务点击可达1.3万/日，具体健康问答应用示例如图17所示。

后续，就医助理将继续在病历生成、辅助诊断等方面进行应用拓展，以进一步提高诊断效率与质量，让每个患者都有贴心的智能陪诊员。

图17示例-就医助理在健康问答的应用

（三）大模型助力政务场景智能化转型

大模型在政务领域加速应用落地，目前主要的应用场景有政务服务、政府办公、城市治理、热线问答等。此处选择城市治理场景作为典型应用来分析大模型可信应用框架在政务领域的落地实践，具体如下。

1. 应用挑战

一是城市多主体参与的统筹协调挑战，城市治理需要协调政府、企业、社会组织和市民等多方利益，这要求模型在推理决策时能够兼顾到不同主体的需求。

二是治理多层次问题交织的挑战，城市治理问题往往涉及多个层面，如规划、建设、治理等，对大模型跨领域的专业知识储备、治理问题的分析框架与决策流程严谨性提出了更高要求。

三是城市治理事件的动态发展和不确定性，要求大模型能够结合着实时更新的信息进行数据处理和分析，并给出真实可信的推理结果。

四是城市治理的整体性和系统性需求，城市是一个复杂系统，不能仅仅关注孤立的问题或单一业务领域，各项治理措施要相互协调、相互支持，推理决策过程需能够考虑到不同的约束条件，对大模型的可控性提出了较高要求。

2. 实践方案

以某面向城市治理领域的数字社工产品为例，其基于大模型能力，利用城市治理领域法规条例、事项及工单等海量业务数据进行大模型的专业能力调优，可实现工单事件的智能分派、行政执法智能辅助等功能，解决管理人员事件调度分析难、办事满意度低、事件处置效率低等问题，助力城市治理水平提升，如图 1 8所示，具体可信应用框架的实现措施介绍如下。

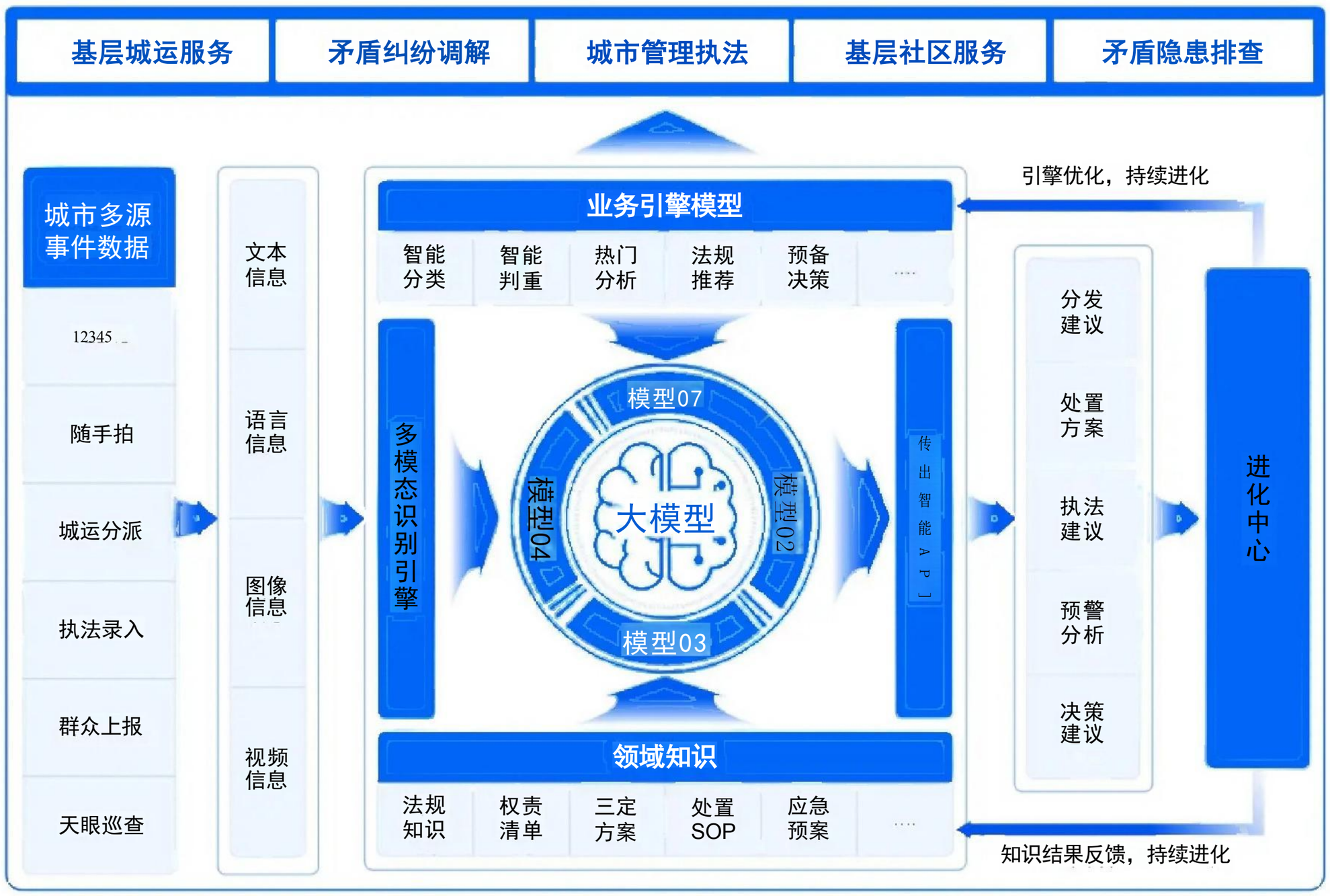


图18城市治理场景中的应用

多模态多源城市治理信息的采集与融合。城市各类事件成因复杂、涉及面广，且事件演变路径不清晰，需要全方位、深入理解事件信息。通过接入**12345** 便民热线、市民小程序随手拍、城运中心派单、基层执法人员录入、城市天眼巡查等众多信息渠道，实现城市治理事件各类信息(文本、语音、图像及视频)的数据采集和汇总。基于这些多模态多源数据， 一方面可以通过训练调优提升了大模型在复杂城市事件治理的专业能力，另一方面可以在推理过程中为大模型提供更为及时的数据，以支撑其在各类复杂城市治理事件的及时发现、准确判定及精准立案，如**图 1 9 所**示 。

图19多模态多源信息的理解和融合示意

跨领域知识库构建及业务协同处置建模。城市治理涉及的业务领域及关键主体众多，涵盖市场监管、综合执法及环保等领域，各业务领域知识专业度高、业务复杂性高。通过收集相关领域的法律法规、处置规范、优秀历史案例等构建出高质量的城市治理数据集，并进行知识抽取形成全面准确的行业知识库，搭建出城市治理各领域知识模型。应用中，依托大模型的语义理解能力，精准匹配各类事件权责知识和流转知识，实现对复杂城市治理事件的深度理解、智能匹配及分派。同时，依托大模型的事件库还能实现对重复上报事件、相似案件的精准识别，有效减少重复派单现象的发生。此外，大模型会基于事件分派的执行情况和反馈信息，持续学习相关的案件知识，不断迭代、优化事件分派原则及决策建议。如**图 2 0 所**示。

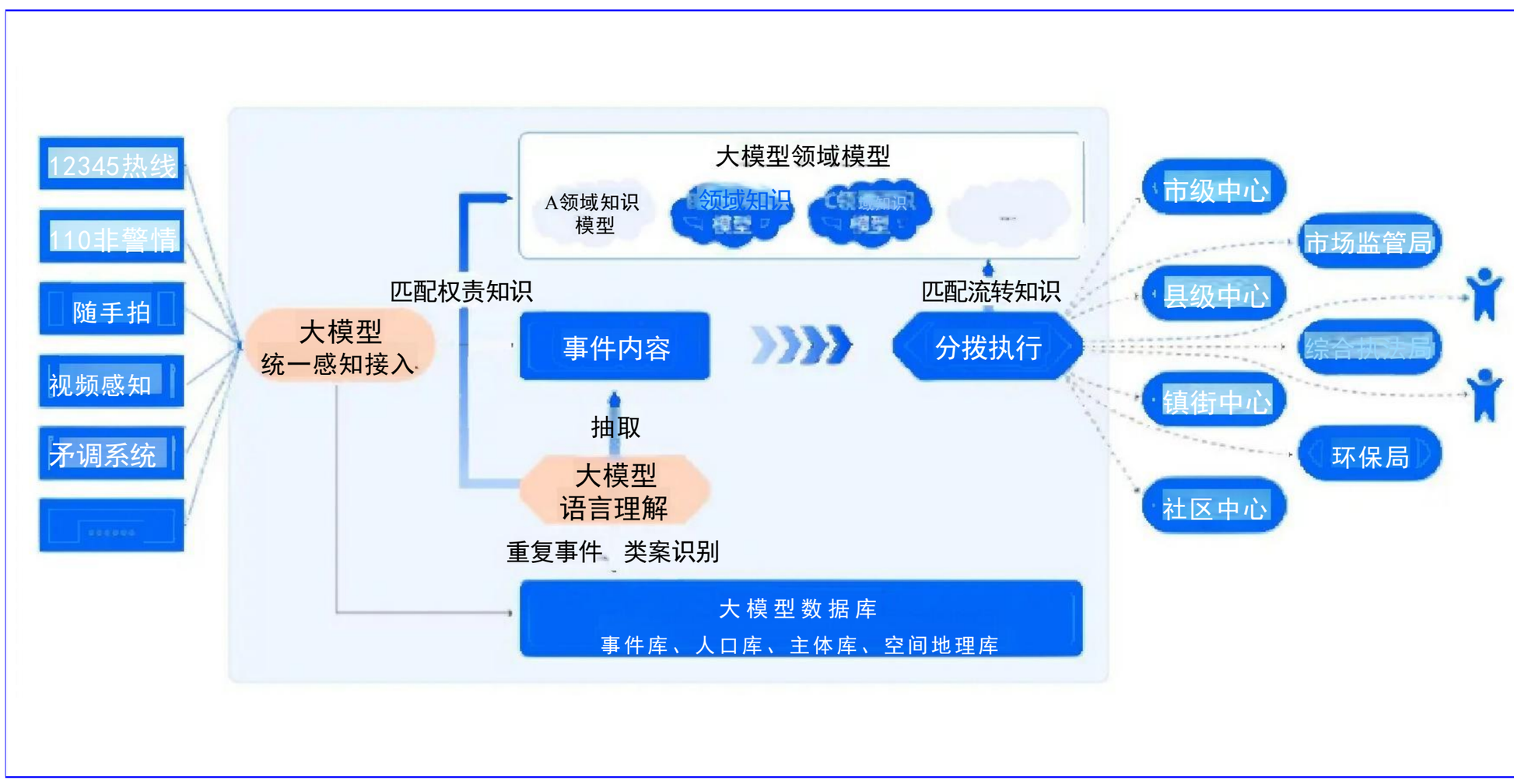


图20跨领域业务协同处置建模示意

领域知识模型优化，形成可复用可迁移的知识构建逻辑。以领域专家提供的专业知识为基础，通过构建领域知识学习（持续学习和数据积累，不断丰富领域知识库）、融合（将不同来源和类型的数据进行有效整合，形成统一的知识体系）、进化（根据新的数据和反馈自动调整和优化其知识结构和处理策略）、反馈（接收来自用户和环境的反馈，以实现持续改进和性能提升）等自闭环能力，形成社会治理行业的方法论，叠加新行业和地区，实现大模型跨领域知识自进化学习，达到触类旁通，随着系统处理的数据量和覆盖的领域不断增加，系统的学习能力和治理效率也将呈指数级提升，实现更加高效、智能和可持续的治理模式。

推出专业政务领域任务测评集CGA-Eval，**对大模型在政务领域的可信应用程度进行充分评估。**CGA-Eval 政务领域中文语言专业数据评测集覆盖政务通用、政务安全、行业知识、应用场景四个一级维度，总的评测题数目达到1万条以上。其中，政务行业知识覆盖常见政策法规、公积金、交通管理、市场监管、财政税收、规划和自然资源、社保医保、户籍制度等多个典型场景，评测题型包括单选、多选、填空、判断、问答等多种类型。评测的政务应用场景重点涉及智能客服、公文写作、智能导办、政策查询四个典型应用，通过单轮、多轮问答的方式可以对大模型在政务领域的应用程度进行全面评测。

3. 应用成效

目前该数字社工产品已在国内多个城市街道应用落地，工单智能派单、智能处置、智能催办、智能质检、智能研判等助手有效助力基层治理现代化，实现派单准确率95%以上，派单效率提升60%，处置效率提升60%、隐患事件发现率提升200%。使用界面如图21所示。

图21示例- 数字社工系统使用界面

05

未来展望

未来展望

新一代人工智能技术，特别是以大模型为代表的技术，正处于快速迭代之中，为“新质生产力”的提升注入了强大的动力，推动了产业智能化和经济的增长。只有大模型能够被“可信”地在金融、医疗、政务等专业领域进行规模化应用并带来更为深远的应用价值时，才是这轮以大模型为核心的人工智能新技术范式生产力得到充分释放的标志。

本报告提出了面向专业领域的大模型可信应用框架，即在面向如金融、医疗、政务等专业领域的应用中，构建一个以大模型为核心的专业智能服务体系，该体系应确保应用的专业性、可控性、真实性和安全性，以满足专业领域的高标准要求。**在技术实施层面，**构建大模型可信应用框架的核心目标是在其开发和部署的各个阶段引入适当的技术保障措施，以增强实际应用中的可信度。在**应用实践层面，**该 技术框架已经在金融、医疗、政务等领域落地实践，应用成效逐渐显现，提高了客户对于大模型应用的信赖程度，推动大模型在专业领域的应用拓展和深化。

展望未来，大模型技术和工程架构仍在快速演进当中，同时企业对于智能化技术的深度应用和价值释放的需求也日益强烈。本报告提出的大模型行业可信应用框架是产业中规模化落地应用释放价值的开端，未来的产业化突破还需要从前沿技术创新探索、可信应用框架落地实施、行业治理体系搭建、产业生态合作完善等多个维度持续推进，具体包括：

从算法演进来看，新的基础架构和创新技术突破值得期待。**一是**在基础模型架构方面，探索更加高效、灵活、可扩展的大模型底层模型架构**如Mamba、MOE-Mamba** 等；**二是**价值对齐方面，探索更加高效、鲁棒、可扩展的对齐方法如基于准则的对**齐**Aligner、实时偏好优化等；**三是**模型透明可控方面，目前在机制可解释性、特征重要性分析、注意力机制可视化等方面正取得进展；**四是**结合隐私计算技术提升安全性、结合元学习等进一步提升模型的泛化能力等。

从工程突破来看，**一是**算力基础设施方面，通过采用更先进的芯片技术、分布式计算架构和优化调度技术，实现更高的计算性能和更低的能耗，提升算力基础设施利用率和可扩展性；**二是开**发部署工具链和平台方面，注重工具链和平台的自动化和模块化，通过引入自动调参、智能诊断等功能降低技术门槛，提高开发效率，更进一步支持大模型的高效开发与部署；**三是**模型应用生态方面，通过共享数据集、构建开发者社区等方式促进知识共享和技术进步，推动应用生态开源开放；**四是**风险防控体系方面，通过隐私保护机制、模型透明度提升、伦理审查流程建立完善的风险防控体系，为大模型研发应用可控可信提供支撑。

从**应用落地**来看，**一是可**结合着大模型在不同行业和场景的应用风险程度进行分类分级，并相应构建出可信应用框架的不同可信等级和对应技术要求(如有需要可制定出相应技术标准)；**二是评**测验证体系是确保落地实施时能达到相应应用场景可信要求的重要保障，需进一步细化并构建出在不同行业和场景的特定评测验证基准，并在大模型应用上线前进行充分测试；**三 是**可以结合行业实践进一步总结出可信应用框架的细化实现指南或最佳实践等，降低产业各方落地实施成本。

从**治理体系**来看，由于大模型技术的持续快速迭代，亟需构建新型的大模型可信应用行业合作框架，这一框架应当涵盖监管机构、大模型开发机构、大模型应用机构以及独立的第三方专业评测机构等。通过建立常态化的沟通互动机制，及时分享相关技术和产业进展，促进各方相互理解和共识达成，并基于监管要求与行业自律相结合的方式，构建出一种鼓励创新与安全发展并重、可动态升级适应大模型技术迭代节奏的新型行业治理体系。

从**产业生态**来看，大模型产业链上下游要加强合作，如通过共同成立大模型可信应用联盟等方式，携手推动技术创新和应用落地，加强技术开源开放、共建共享高质量行业数据集、联合制定产业技术标准和最佳实践，并联合专业评测机构构建权威的大模型可信应用评测体系等，为产业发展提供助力和指引。

欢迎扫码对研究报告内容进行反馈

蚂蚁集团 ANT GROUP

CAICT 中国信息通信研究院 China Academy of Information and Communications Technology