阿里云

GenAI技术落地白皮书

目录

Contents

核心观点

生成式人工智能（Generative Artificial Intelligence，GenAI）即将迎来全面爆发，各行各业必须为此做好准备。本报告从企业视角出发，聚焦技术，阐述GenAI在企业落地时的关键考量点，提出了**“选-育-用”**方法论，覆盖了从模型和技术路线的选择，到如何培育适合企业的大模型，并将其广泛应用在企业流程实现全面创新的全生命周期，为企业规模化GenAI落地提供指导。核心观点如下：

1. 企业应充分了解不同产品服务、技术解决方案背后的技术难度、成本及其能达到的效果，结合自身的技术实力、资金储备以及业务目标，作出合适的选择；特别是面向不同应用场景时，可以采取不同的产品服务模式而不必限于单一选择。

2. **选：**企业需要结合自身情况选择构建GenAI能力的技术路线：深度研发大模型，或者基于现有大模型进行工程化适配，或者直接使用大模型服务。后两条路线适合大多数企业，此时要做好大模型的选择，形成自己的大模型池。面对具体的应用场景，选择大模型的关键是在成本、效果和性能的“不可能三角”间进行权衡和取舍。

3. **育：**定制适应企业的大模型需要基于基础大模型进行工程化适配，按照技术难度从小到大和投入成本从少到多，主要包括提示词工程、检索增强生成和微调三种方式。其中，微调会改变部分大模型参数，微调后还可以通过知识蒸馏、剪枝、量化等手段“压缩”大模型达到灵活的适应性，需要较高的技术门槛。

4. **用：**广泛应用GenAI需要解决基础设施问题。相比传统的自建或租用数据中心方式，使用云基础设施或者采用云托管大模型的方式能够节约时间成本、降低现金流压力。企业可以通过Agent将大模型的能力与企业应用紧密集成，基于GenAIOps做好跨团队紧密协作、消除流程断点，从而加速GenAI应用上线，并根据效果及时更新。此外，需要始终关注GenAI应用的信任、风险和安全管理，构筑可信任的基石。

1. GenAI构建企业竞争新优势

GenAI是一种先进的人工智能技术，它能够基于已有的数据和知识生成全新的内容。这种技术的发展得益于深度学习、大数据和计算能力的发展，特别是大型语言模型（Large Language Models，LLMs）等基础模型的进步。GenAI将逐渐改变人们与机器交互的方式，为各行各业带来前所未有的创新机遇。

当前，GenAI正处于爆炸性增长阶段，ChatGPT的火爆更是印证了这一点，它展现了GenAI在交互性、实用性和创造性上的巨大潜力。工业界和学术界都在积极投入资源，探索如何利用GenAI实现经营提效、体验提升以及业务创新。市场上涌现出各种基于GenAI的应用，比如自助式数据分析、定制化内容创作、个性化推荐、自动化客户服务以及辅助设计与研发等。与此同时，GenAI的伦理、安全和合规等潜在问题也日益凸显，如何保障GenAI的可持续和负责任发展成为各界广泛关注的问题。

打造GenAI能力，已经成为企业全面迈向智能化、构建市场竞争优势的必然选择。GenAI可以推动产品创新，通过快速生成设计和创意，加速产品开发流程；提升成本效益，利用自动化内容生成，将人力从重复性工作中解放出来，更专注于发挥创造力；降低数据分析的门槛，人人都成为数据分析师，从而实现科学决策，为企业提供精准的决策支持；改善用户体验，根据用户行为和偏好，实现高度个性化的产品和服务；基于GenAI能力打造AI原生应用，带来颠覆性的体验和价值。

企业构建GenAI能力，是一个涉及战略、组织、文化和技术等多个维度的综合问题。本研究将聚焦技术层面，分析GenAI在企业业务场景中全面落地的关键考量因素，提出**“选-育-用”**的GenAI落地方法论，从选择技术路线和基础模型入手，培育好适合企业的定制化大模型，并将其高效、安全地应用在企业的方方面面，从而助力企业充分发挥GenAI能力，构建独一无二的竞争优势，带来可观的商业价值。

2. 大模型的选择

2022年11月30日ChatGPT的面世，拉开了GenAI发展的新篇章。短时间内，GenAI取得了日新月异的发展，目前市面上已经出现众多各具特色的产品服务：产品门类繁多——有适合多种通用任务的基础大模型，还有各类适应特定行业或场景的行业大模型和场景大模型；服务模式多样——既可以像私有云一样本地化部署，还可以如公共云那般按用量付费，甚至能够类似混合云那样博采众长、多措并举。

面对如此众多的市场选择，企业应当如何确定最适合自己的GenAI服务呢？我们建议，企业首先根据自身的业务需求和成本预算来选择技术路线，然后权衡模型的效果、性能等因素选择合适的大模型。特别是当企业在面向多个业务场景需求时，可以不局限于单一大模型产品服务甚至技术路线，而是根据不同场景的特殊需求和市场上相应产品服务的成熟性和契合度，分别选择最合适的产品服务。

2.1 大模型技术路线

企业使用大模型服务的技术路线，主要包括深度研发大模型、基于现有基础大模型进行工程化适配、直接使用大模型服务三种。

表1 GenAI主要技术路线的优劣势比较

大模型技术路线	总成本	技术难度	上线周期	可定制
深度研发	非常高	非常高	长	高
工程化适配*	较低	较低	较短	较高
直接使用	低	低	短	低

*注：不同的工程化适配方法在成本、技术难度、上线周期和定制化能力方面存在差异，此处为与另两条技术路线相比的平均水平。

1 深度研发大模型

深度研发大模型，是指企业从0到1完全自主研发或者基于开源模型做深度定制得到大模型。这一过程涵盖模型设计、数据准备、环境准备、模型训练、模型评估和优化等多个阶段。

深度研发大模型可以针对企业的具体场景需求进行优化设计，理论上可以更为聚焦地解决特定问题，从而拥有更好的表现。企业在研发过程中掌握充分的模型技术细节，拥有较高的自主性，从而不受外部供应商的限制。

但是深度研发往往需要投入巨大的研发成本，包括计算资源、稀缺技术人员的薪资等。从启动自研到上线应用的时间跨度长达数月甚至以年计，并且需要持续投入，以确保在快速的技术迭代中不掉队。由于技术体系复杂、研发难度大，企业可能面临模型性能不理想、项目延期或失败等风险。

总体而言，深度研发大模型是成本最高、难度最大、周期最长的一条技术路径，除非是拥有高密度AI人才、资金充足的企业，否则并不推荐。

2 基于现有基础大模型进行工程化适配

基于现有基础大模型进行工程化适配，是指企业在已有的大模型基础上，针对具体应用场景进行的技术调整和优化工作，以更好地适应企业场景。这一过程不仅涉及技术上的适配，还需要综合考量成本、性能、安全、可维护性等因素。对于用户来说，常用的工程化适配方式包括提示词工程（Prompt Engineering）、检索增强生成（Retrieval-Augmented Generation，RAG）和模型微调（Fine-tuning）。企业还可以通过知识蒸馏、剪枝、量化等手段减少大模型的参数规模，降低推理的计算量，提高大模型的响应速度。

选取这一技术路线无需为基础大模型的训练付费，从而显著减少开发成本；同时可以优化大模型在特定任务领域的输出，在特定任务上得到更好效果的预期较高。该路线尽管有一定的技术门槛，但不算太高，经过一定培训的技术人员即可掌握，因此适合于几乎所有的企业用户。特别是当市面上现有的大模型产品和服务无法直接满足企业的特定需求时，基于现有基础大模型进行工程化适配几乎成为企业的必然选择。

3 直接使用大模型服务

企业还可以直接采购已经训练好的大模型来解决业务问题。一些模型服务商提供将自家模型部署在客户环境的能力，更多模型服务商和云平台合作，采用云托管的方式，这种方式随用随取，按需使用，进一步降低了使用大模型的门槛。

直接使用大模型服务无需投入大量资源，有效降低使用成本。企业不需要深入了解技术细节，业务团队可以快速上手，直接将大模型集成到现有系统中，迅速享受到大模型的红利。部分第三方服务提供商针对市场规模较大的行业或通用性较强的业务场景推出了特定领域的专用大模型产品，例如在智能客服、信息检索、代码生成等领域，这进一步提升了大模型的使用效果和用户体验。直接使用大模型服务的方式适合于大多数企业，特别是成本预算有限、技术能力欠缺的中小微企业。另外，随着基础模型能力的不断提升，以及该方式可以与提示词工程、RAG等工程化适配方法相结合，使得云端API调用的方式被越来越多的企业重视。

2.2 基础大模型的选择

在企业构建GenAI能力的三条技术路线中，除了不适用于多数企业的深度研发，无论是对基础大模型进行工程化适配，还是大模型的直接使用，其中最关键的环节就是基础大模型的选择。在这一过程中，需要综合考量各种因素，包括企业的业务场景需求、成本预算、员工技术水平，模型的生成质量、泛化能力、响应速度等，但本质上，选择大模型服务的关键是在成本、效果和性能构成的“不可能三角”间进行权衡和取舍。

图1 大模型的不可能三角

- **成本**指的是企业大模型落地的整体费用，包括大模型的训练成本、推理成本以及部署、运维和升级成本等。企业有时仅关注有形成本：例如GPU购置费用、消耗的电费，或从第三方服务商购买模型服务的费用；而会忽略无形成本：包括为实现大模型服务而配置的人力成本，以及大模型在部署、训练或调试阶段消耗的时间成本等。企业在核算成本时，需要考量总持有成本，特别是不要忽略无形成本。按成本从高到低，一般为深度研发大模型、微调、RAG、提示词工程、直接调用。

- **效果**指的是大模型生成内容的质量，包括内容的准确性，是否存在幻觉问题，或是否会生成不合适的内容。大模型效果可以基于“3H”原则进行评价：1）Helpful：内容可用有帮助，不要废话连篇、泛泛而谈；2）Harmless：内容合规无害处，符合伦理规范和监管要求；3）Honest：内容正确无幻觉，不要一本正经地胡说八道，甚至给出错误信息。通常来说，大模型的参数规模越大，生成效果越好。因此，当业务需求对生成内容质量要求严苛时，应尽量选择参数规模更大的模型。此外，目前市场上主流商业化模型的效果，大多优于同期同参数规模的开源模型。

- **性能**指的是大模型服务的速度，包括大模型的训练速度，推理时的响应速度、生成速度等。一般而言，大模型的参数规模越大，则需要的训练时间越长，即训练速度越慢，而其进行推理服务时的需求响应速度和内容生成速度也越慢。因此，大模型的效果和性能不可兼得，当成本固定时，大模型的选择主要是在效果和性能之间进行平衡和取舍。对于性能要求较高而对效果有一定容忍度的场景，可以选择参数规模相对较小的大模型。

基础大模型的选择是个综合性任务，除了做好成本、效果、性能“不可能三角”的权衡，还需要同时考虑一系列其他因素：例如集成难度，即模型服务与现有系统的集成复杂度及其所需的技术投入；技术友好性，即技术人员的学习和使用难度；模型扩展性，即模型的更新、升级频率和向下兼容性；模型生态，包括模型系列的参数尺寸全面性及其背后的工具生态系统和合作伙伴网络等；服务商可靠性，包括服务商的口碑声誉、技术实力和服务能力以及客户成功案例等。这其中，企业需要格外注意大模型服务的合规性与安全性，以免影响业务的正常开展甚至造成企业数据的泄露。在国内，提供基础大模型服务的供应商除了需要遵守数据安全相关法规，还需要完成生成式人工智能的算法备案和服务备案。

3. 大模型的培育

在大模型的三条主要技术路线中，基于现有大模型进行工程化适配是最受企业关注的一条路线：它在成本方面与直接使用大模型相持平，有一定的技术门槛但总体上难度不大，同时能够解决基础大模型或行业大模型不能实现的一些业务特殊需求。工程化适配按技术难度从小到大和成本从低到高，主要包括提示词工程、检索增强生成和微调三种方式。

3.1 提示词工程

提示词工程，是指通过精心设置提示词（Prompt），引导模型生成更准确、更有用的输出。

提示词工程的关键是清晰、明确地表达用户的意图，需要确保提示词直接、具体，减少歧义，让模型能够准确捕捉到问题的核心。因此，通常采用包括指令、上下文和期望输出格式的提示词结构，特别是可以根据模型擅长处理的格式来设计提示词模板（Prompt Template），并通过试验找到最优的提示词组合。

提示词工程能够在不修改或重新训练大模型的情况下，引导模型更加精准地完成任务，从而有效控制成本。良好的提示词设计能够显著提升模型的输出质量，使得模型效果更贴近用户期待。但同时，提示词工程高度依赖用户经验，优秀的提示词需要对领域知识和模型特性均有深入了解，这需要大量的人力投入和试错。不当的提示词还可能引入或强化模型偏见，导致模型生成不恰当甚至有害的内容。

大模型本身能力决定了提示词工程效果的上限。如果基础大模型训练时纳入了充足的行业数据，提示词工程可以有效引导模型进行高质量输出，但如果基础大模型内含的行业数据匮乏，提示词工程的作用就十分有限，此时可以采用RAG或微调的方式对基础大模型进行数据补齐。

图2 提示词工程流程示意图

3.2 检索增强生成

RAG是指从外部知识库中检索相关信息，作为上下文输入给大模型，从而提升生成内容的质量。

RAG是一种结合了信息检索和文本生成的技术方案，它需要企业构建知识库，特别是要在知识库中纳入企业希望重点服务的业务场景数据。

RAG通过引入外部权威信息，显著提升大模型内容生成的准确性和丰富度，减少错误和臆测；生成的内容可以追溯到具体的信息源，提高透明度和可解释性。同时，基础大模型只会调用相关数据，而不会吸收数据成为内含知识。RAG能够在不改变大模型本身的基础上，快速、显著地提升大模型在特定领域的表现，因此成为企业部署大模型应用的主流选择。

但是，RAG引入检索步骤增加了系统的整体复杂度，包括建立和维护知识库、优化检索效率等，特别是生成内容的质量高度依赖于检索系统的性能和检索信息的质量，这使得RAG相比于提示词工程增加了成本，并提高了技术门槛。此外，检索过程可能导致大模型的响应速度变慢，对性能优化提出了更高的要求。

图3 RAG流程示意图

3.3 微调

微调，是指在预训练大模型基础上，针对特定任务或领域进行再训练，以提升大模型在该特定任务上的表现。

微调利用特定任务的数据集，调整大模型的部分或全部参数，进而将行业知识内化到大模型中，因此，数据质量直接影响微调后的大模型效果。同时，微调策略也直接影响大模型效果，常用的微调方法包括有监督微调（Supervised Fine-tuning，SFT），即在标注数据上调整模型参数；低秩调整（Low-Rank Adaptation，LoRA），即通过低秩矩阵减少更新参数量等。微调策略可以根据任务需求、数据量和计算资源等综合考虑。

微调能够提升大模型在特定任务上的准确性和泛化能力，特别是法律、医疗等专业性较强的领域，可以显著提升内容的专业度。微调具备较强的灵活性，可以对基础大模型进行多次微调，以适应不断变化的任务需求。

微调过程需要消耗一定的计算资源，且参数调整存在难度，找到最优的参数复杂且耗时，这使得微调相比提示词工程和RAG具有更高的技术门槛。同时，微调存在过拟合风险，如果数据量太少或过度调整，则会导致大模型的泛化能力下降。

微调是目前较为常用的行业大模型和场景大模型的构建方法，但由于其存在一定的资源和技术门槛，因此并不适合所有企业，特别是不适合小微企业。

图4 微调流程示意图

4. 大模型的使用

企业在确定GenAI的技术路线、选取合适的基础大模型并完成工程化适配后，就需要规模化进行GenAI应用开发和部署。在这个环节，企业需要构建基础设施，为大模型应用提供必要的基础资源和部署环境；可以通过Agent方式将GenAI嵌入现有业务流程进而提升效能；实施GenAIOps（生成式人工智能运维），充分发挥GenAI潜力，驱动业务创新和可持续发展；应用AITRiSM（人工智能信任、风险和安全管理），对GenAI的可信度、安全性和相关风险进行有效管理，确保GenAI合法、合规、可靠运行。

4.1 基础设施建设

当企业选择深度研发或基于现有基础大模型进行微调时，面临的首要问题就是基础设施的建设。企业构建大模型基础设施的方式主要包括自建或租用数据中心、使用公共云服务两种方式。而当企业选择以RAG、提示词工程进行工程化适配或者直接使用大模型时，模型服务商会提供成熟的产品或API接口服务，可以面向用户屏蔽大模型的基础资源消耗，用户只需关注大模型产品服务的实际应用效果，而无需分心底层的基础设施建设。

图5 GenAI基础设施服务方式

1 自建或租用数据中心

自建数据中心是指企业自行规划、设计、建设并运营管理数据中心设施。自建数据中心通常包括选址与规划设计、机房电力等基础设施建设、软硬件采购与安装配置、系统集成与测试、运维管理等多个环节。租用数据中心相比自建数据中心，减少了数据中心的选址与规划设计、机房电力等基础设施建设环节，并可以根据实际需求选择自己采购或租用硬件设备。

自建数据中心允许企业根据自己的业务需求和技术要求，定制数据中心的规模、配置和功能，从而实现对基础资源的全面掌控和管理。而租用数据中心尽管对数据中心规模的选择明显受限，但可以显著降低企业前期的资金和时间成本投入，相比自建数据中心灵活性更高。总体上，自建或租用数据中心均具备自主可控、定制化程度高等优势。

但同时，自建或租用数据中心是一个复杂工程，需要考虑诸多因素，包括技术、硬件、网络、安全、管理等。这使得自建或租用数据中心的技术门槛高、运维压力大，特别是自建数据中心建设周期长，且前期资金投入巨大，成本高昂。

可见，自建或租用数据中心能够为大模型提供高度定制和易于控制的环境，但伴随而来的是巨大的管理成本以及财务负担，因此仅适合于极少数企业或组织采用。

2 使用公共云服务

使用公共云服务构建GenAI基础设施，是指企业利用云服务商的资源和平台来搭建和运行GenAI应用的计算环境。通过公共云服务构建GenAI基础设施，主要包括设置云环境、选择计算资源、数据管理、开发环境搭建、模型训练与部署、监控与优化等环节。

公共云平台提供了丰富的产品服务，从计算资源、存储服务、网络配置到一系列AI开发和部署工具甚至各种基础大模型和数据集，使得用户无需自建昂贵的数据中心，即可快速部署和扩展AI项目，从而帮助企业极大地减少初期投资、有效控制成本。同时，企业可以根据业务需求，灵活选用公共云上的产品服务层次，既能够全面基于云服务商提供的产品构建大模型服务能力，也可以只采购云服务商的硬件基础资源。此外，云服务商负责底层硬件和软件的维护，能够减少企业运维负担。而为了更好地帮助用户，云服务商还会提供专业的技术支持和持续的服务更新。

尽管公共云平台会提供各种资源、工具甚至技术支持，但对于多数企业而言，GenAI计算环境的配置和调优仍然颇具技术难度。使用公共云服务，还意味着用户需要依赖云服务商的服务和技术，因此，企业需要仔细评估公共云提供商的安全性和合规性，并采取适当的数据安全保护措施。

使用公共云构建GenAI基础设施是一种灵活、高效的方式，尤其适合那些希望快速部署、按需扩展、并专注于核心业务而非基础设施管理的企业。

4.2 通过AI Agent升级业务流程

AI Agent，即人工智能智能体，是一种能够感知环境、进行决策和执行动作的智能实体。它以大模型为核心驱动力，并通过记忆、规划和工具等组件分别实现信息存储、决策制定与反思总结、任务执行等功能，从而实现特定目标。

图6 AI Agent工作原理示意图

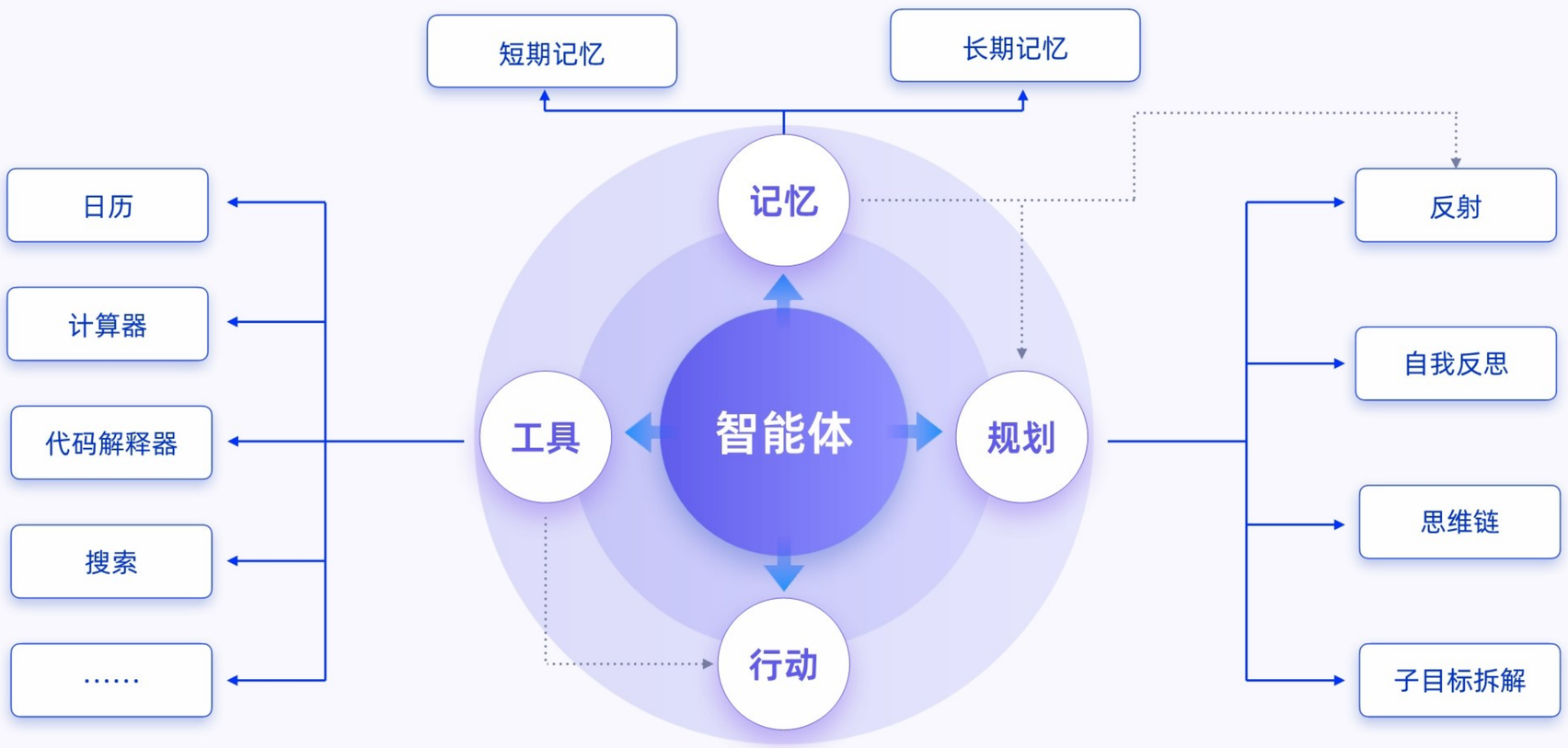


图7 AI Agent对工作范式的改变

	传统工作范式	AI Agent 工作范式
架构方式	面向过程	面向目标
思维方式	软件工程思维	Agent 工程思维
协作方式	人为中心，AI为辅助	AI为中心，人为辅助
特征	• 固定的交互界面 • 预定义有限域的任务 • 信息分发平台 • 海量数据 • 规模化的生产方式	• 动态的交互界面 • 无预制开放域的任务 • 知识和世界模型的分发平台 • 海量任务 • 兼具规模化和个性化的生产方式

AI Agent的核心在于其自主性、智能性、反应性和交互性，能够根据预设的目标或学习到的知识自动完成任务。因此，AI Agent将改变工作范式，重塑业务流程。在AI Agent的推动下，人机协作方式将从目前的“以人为中心，AI为辅助”转向未来的“以AI为中心，人为辅助”，极大地扩展工作的范畴和方式。作为一种先进的技术解决方案，AI Agent可以显著提高生产力，将广泛应用于各行业，例如，在个人助理方面，能够安排日程、回答问题、控制家居设备；在客户服务方面，能够提供24小时在线服务，解答疑问，处理投诉；在工业制造方面，能够监控设备状态，优化生产流程，预测维护需求；在医疗健康方面，能够辅助诊断疾病，监测患者健康，提供治疗建议；在金融服务方面，能够进行风险管理、欺诈检测、智能投顾，并提供个性化财务建议；在教育方面，能够智能答疑，并提供个性化学习路径规划和学习效果评估等。

AI Agent是具有广泛应用前景和巨大发展潜力的智能实体。随着技术的进步和场景的拓展，AI Agent将充分挖掘每个业务流程的效能和潜力，推动企业步入“人机协同”智能时代。

4.3 高效跨团队协作与持续效果提升，实现GenAIOps

现代软件开发中强调开发团队与运维团队之间的紧密协作与整合（DevOps），以实现快速、持续、高效的软件交付。这一实践在机器学习（Machine Learning）领域的落地即为MLOps（Machine Learning Operations），强调数据工程师、数据科学家、算法工程师以及业务开发人员的高效协作，保障线上、线下数据的一致性以及实现模型的训练、部署、监控、重新训练等一系列流程持续运转，确保模型的效果始终满足业务预期。在GenAI爆发的今天，大模型作为内核，将会与众多企业应用进行更加紧密地交互，迫切需要企业借鉴MLOps的理念，实现GenAIOps，以达到团队高效协作、缩短产品迭代周期、加速产品上市、监控与评测效果、及时升级产品以及确保产品安全合规等目的。GenAIOps的范畴包括：

1 增进团队协作

GenAI原生应用开发，或者基于GenAI能力的现有应用升级都涉及多个团队的紧密协作。在模型调优或工程化适配阶段，需要数据团队提供精准业务应用范围的微调数据或外挂知识库数据，需要算法科学家与AI工程师团队进行微调、RAG或提示词的工程化，特别是需要业务团队对于提示词的控制、对AI Agent与应用协同的控制以及对应用上线后如何衡量应用效果以及是否达到业务目标给出建议和改进意见。在整个过程中，安全合规团队要确保数据没有被滥用以及从负责任（responsible）的角度确保数据安全以及大模型的表现不出现偏见，符合法律、伦理与道德。

2 消除流程断点

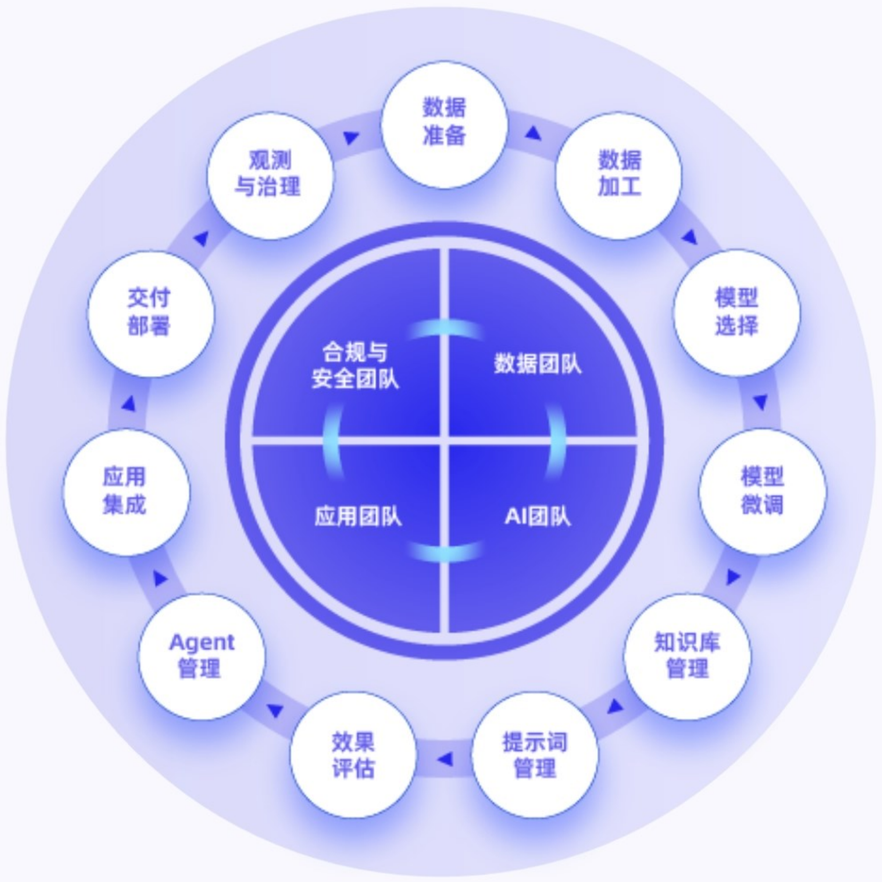
在开发GenAI应用的过程中涉及多个阶段，流程间的断点将会导致开发速度减慢、出现潜在错误或者不能及时发现已有问题。重要的流程自动化包括：

从数据准备到模型工程：从企业内大量的非结构化数据以及规章流程等条款中，提炼出可用于调优的语料，或者可输入的提示词，或者可供大模型调用的知识库等，均需要实现自动化，以免因数据问题影响模型效果。

从模型就绪到应用上线：企业可能会选择新的模型，或者对已有模型进行调优，之后需要将新模型嵌入到应用中。如果业务逻辑发生变化，或者通过AI Agent调用应用系统的插件发生改变，也需要及时更新。这一流程需要实现自动化链路以避免出现模型结果到执行动作之间的错位。

从应用上线到效果监控到持续提升，形成闭环：大模型应用上线只是开始，随着时间的推移，应用逻辑的变化、在线数据的漂移等都可能影响应用效果。这需要持续不断地监控模型表现和应用效果，及时反馈到数据准备、模型工程并重新部署上线，进而最大化应用价值、最小化业务风险。

图8 GenAIOps理念



GenAIOps不仅是技术框架，也是一种战略思维，它能够帮助企业充分利用GenAI的潜力，并管理好伴随而来的复杂性与风险，从而在快速变化的商业环境中获得竞争优势。

4.4 利用AITRiSM保障GenAI安全

AITRiSM，即人工智能信任、风险与安全管理（AI Trust, Risk, and Security - Management），是用于保证AI系统可信度、安全性并对相关风险进行有效管理的方法论和技术实践，它涵盖了AI从设计、开发、部署到运行的整个生命周期，确保AI应用不会损害用户隐私、数据安全或产生不可预测的行为，同时增强用户对AI系统的信心。

随着AI技术的广泛应用，特别是GenAI的兴起，企业面临着前所未有的挑战，包括内容偏见、数据泄露和AI应用漏洞等。传统的安全控制不足以应对这些新兴风险，因此需要专门针对AI的管理策略以保障系统的稳健性、合规性并获得用户信任。AITRiSM可以帮助企业识别并应对AI带来的风险，确保AI应用的合规可靠运行，保护企业的资产和声誉。

1 风险管理

风险管理主要涉及识别、评估和缓解与AI应用相关的风险。它包括内容异常检测，确保输入和输出内容的准确性、适宜性和合规性；数据保护，防止数据泄露和滥用，确保数据在传输和存储过程中的安全；AI应用安全，保护企业免受黑客利用GenAI进行攻击。企业应该采用内容异常检测工具来限制不当或非法的模型行为；使用AI应用安全产品来防御外部威胁；同时，建立PoC（概念验证）来测试新兴的AITRiSM产品，并逐步将其应用于生产环境。

2 信任管理

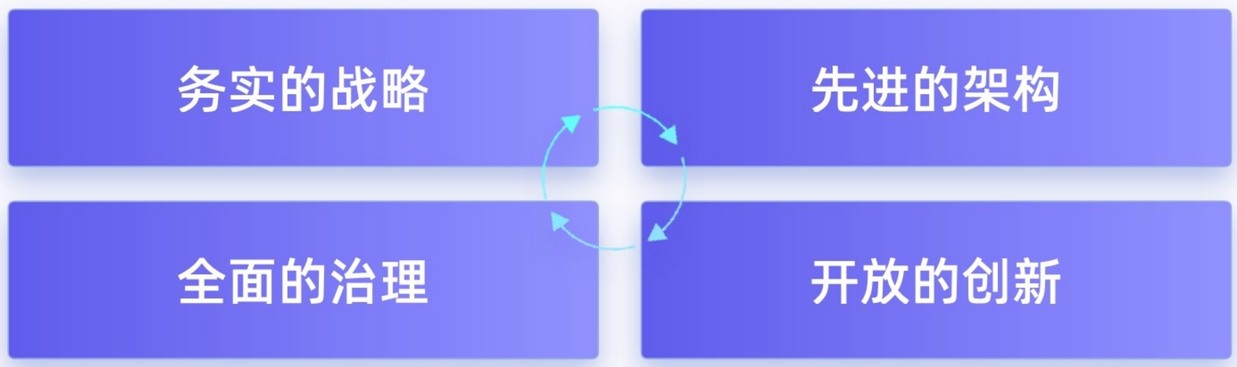
信任管理关注于建立和维护用户、合作伙伴和社会对AI系统的信任。这包括但不限于确保模型决策的透明度和可解释性，让用户理解模型如何做出决策；实施公平性与无偏见管理，避免算法歧视；提供可靠的隐私保护措施，确保用户数据的安全。实现信任管理的关键在于实施透明的治理框架，进行定期的伦理审查，并提供清晰的用户沟通，使用户了解AI系统的工作原理及其背后的数据处理方式。

3 安全管理

安全管理的目的是确保AI系统不受未经授权的访问、篡改或滥用，并防止数据泄露或损坏。安全管理的措施包括实施严格的访问控制，确保只有授权人员可以访问；进行定期的安全审计和漏洞扫描，及时发现并修复安全漏洞；建立应急响应机制，确保在发生安全事件时能够迅速响应并减少损失。特别是要关注新兴威胁，如针对AI模型的攻击，需采用专用的安全工具和策略对模型进行防护。

5. GenAI技术落地策略总结

随着GenAI技术的逐步成熟和市场潜力的不断显现，企业面临着多样化的产品服务和解决方案选择。在这一浪潮中，企业必须从自身的战略出发，综合考虑成本、效果和性能，制定合理的技术架构，以应对GenAI应用的快速发展和潜在的市场需求爆发。在GenAI落地过程中，企业特别需要关注四个要素：



务实的战略

GenAI应用的技术落地是复杂的系统性工程。企业在选择技术路径时，可以考虑深度研发、工程化适配、直接使用等方式。每种方式都有其优势和局限，企业需要根据自身的资源禀赋、技术能力和业务需求来做出选择。

在选择基础大模型时，企业需要在成本、效果和性能之间找到平衡点。企业需要考虑的成本不仅包括直接的经济成本，还涉及到时间成本、机会成本等。当成本固定时，大模型的选择就是在效果和性能之间进行平衡。

先进的架构

企业需要一个灵活、可扩展的技术架构，以支持GenAI应用的快速迭代和升级。这要求企业在技术选型时，因充分考虑到技术的兼容性、集成性和未来的发展潜力，以适应不断变化的技术和业务环境。

全面的治理

在GenAI技术落地的过程中，企业需要识别和管理各种潜在风险，包括技术风险、市场风险、法律和伦理风险等；建立全面的AI治理体系，能够帮助企业及时应对挑战，确保GenAI应用的稳健发展。

开放的更新

为了保持竞争力，企业还需要持续关注GenAI技术的创新，并寻求与其他企业、科研机构和行业组织的合作机会。通过开放的合作模式，共同推动GenAI技术的发展和应用。

总之，GenAI技术的落地并非一蹴而就，而是需要企业进行周密的规划和持续的努力。通过综合考虑各种因素，制定合理的技术架构，企业将在GenAI应用的爆发中占据有利地位，实现可持续的创新和发展。

阿里云

出品团队

策划指导：

刘湘雯　阿里云市场部总裁

穆　飞　阿里云研究院院长

研究撰写：

麻　芃　阿里云研究院高级研究专家

王巍令　阿里云研究院研究专家

创意设计：

张师华　阿里云创意设计专家